Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society

Kenneth Forbus, Dedre Gentner and Terry Regier
Editors

August 4-7, 2004
Chicago, Illinois
USA

Distributed by
Lawrence Erlbaum Associates, Inc.
10 Industrial Avenue
Mahwah, New Jersey 07430

ISBN 0-8058-5464-9

ISSN 1047-1316

Printed in the United States of America

FOREWORD

These proceedings hold the talks, posters, tutorials, and symposia presented at the 26th annual meeting of the Cognitive Science Society. The conference took place in Chicago, Illinois, at the Westin River North Hotel, August 4-7, 2004.

Each year, the CogSci conference chairs highlight a particular area of cognitive science – we chose to highlight *higher-order cognition*. This is reflected in our choice of Susan Goldin-Meadow and Doug Medin as our two plenary speakers, and Larry Barsalou and Art Markman for a debate on embodied cognition, as well as in the themes of the invited symposia: language and thought, qualitative reasoning, higher-order cognitive neuroscience, and large-scale representation systems. The rest of the program was determined by the submissions we received, and by the reviewers' enthusiasm for them.

In total, we received 370 submissions of 6-page papers, publication-based abstracts, and symposia. Of these, 115 were accepted for presentation as a talk or symposium, and 150 were accepted for poster presentation. We also received over 150 member abstract submissions, which by design were only lightly reviewed, and almost all of which were accepted.

Chairing a CogSci conference is a lot of work! We couldn't have done it without a lot of help from many people. We would especially like to thank the following:

The Governing Board, for inviting us to chair the meeting.
The Program Committee, for managing the review process.
The almost 500 reviewers, for providing professional reviews.
Wayne Gray, for helpful advice on all aspects of conference organization.
Chris Schunn and Rick Alterman, for their advice as chairs from earlier years.
Frank Ritter, for organizing and coordinating the tutorial program.
Frank Lee, for coordinating the student volunteers.
Deborah Gruber, for her help with local arrangements and the proceedings.
Jonathan Cohen and Jennifer Stedillie, for coordinating conference preparation.
James Stewart, for heroic responsiveness in operating the reviewing website.
The student volunteers at Northwestern University and the University of Chicago, especially Fey Parrill (UC) and Kate Lockwood (NU).
Dan Veksler, for producing a roommate-matching website at very short notice.
Ken Nielsen, for designing the poster.
Rebecca Asburst, for creating the logo, and Kathleen Braun, for designing the T-shirt.
Our sponsors: The Robert J. Glushko and Pamela Samuelson Foundation; Northwestern University; ONR; DARPA; AFRL; the University of Chicago; CHI Systems; and Trends in Cognitive Sciences.
The plenary speakers, Susan Goldin-Meadow and Doug Medin.
The debaters, Larry Barsalou and Art Markman.
And above all, the authors, symposium participants, and attendees.

Kenneth D. Forbus, Dedre Gentner, and Terry Regier
Conference chairs, CogSci 2004

Conference Chairs

Kenneth D. Forbus, *Northwestern University*
Dedre Gentner, *Northwestern University*
Terry Regier, *University of Chicago*

Conference Program Committee

Martha Alibali
Erik Altmann
Bruno Bara
Lawrence Barsalou
Miriam Bassok
William Bechtel
Paul Bloom
Morten Christiansen
Gary Cottrell
Gary Dell
Judy DeLoache
Kevin Dunbar
Shimon Edelman
Susan Epstein
Fernanda Ferreira
Robert French
Adele Goldberg
John Goldsmith
Rob Goldstone
Wayne Gray
Jim Greeno
James Hampton
Stephen Jose Hanson
Muhammad Ali Khalidi
Boicho Kokinov
David Leake
Vera Maljkovic
Barbara Malt
Art Markman
Laura Michaelis
Naomi Miyake
Michael Mozer
Laura Namy
Srini Narayanan
Nancy Nersessian
Partha Niyogi
Ken Paller
David Plaut
Naomi Quinn
Michael Ramscar
Paul Reber
Frank Ritter
Brian Ross
Franz Schmalhofer
Christian Schunn
Colleen Seifert
Stuart Shapiro
Steve Sloman
Steven Small
Linda Smith
Michael Spivey
Keith Stenning
Josh Tenenbaum
Barbara Tversky
David Uttal
Janet Werker
Amanda Woodward

Local arrangements: Deborah Gruber
Tutorials coordinator: Frank Ritter
Conference software coordinator: Wayne Gray
Registration website: Thomas B. Ward
Conference website: Kenneth D. Forbus
Roommate-matching website: Dan Veksler
Proceedings: Deborah Gruber

Cognitive Science Society
Governing Board

Nancy J. Nersessian (Chair) Program in Cognitive Science, College of Computing Georgia Institute of Technology 2001-2006

Keith Stenning (Past Chair) Centre for Human Communication & Informatics, Edinburgh University2000-2005

Wayne D. Gray (Chair-Elect) Department of Cognitive Science, Rensselaer Polytechnic Institute 2003-2008

Thomas B. Ward (Executive Officer) Center for Creative Media, University of Alabama, 2003-2005

Arthur B. Markman (Past Executive Officer) Department of Psychology, University of Texas, 2004

William Bechtel Department of Philosophy, University of California, San Diego 2002-2007

Peter Cheng Department of Cognitive Science, University of Sussex 2004-2009

Dedre Gentner Department of Psychology, Northwestern University 1999-2004

Robert L. Goldstone (Journal Editor) Department of Psychology, Indiana University 2001-2005

Edwin Hutchins Department of Cognitive Science, University of California, San Diego 2001-2006

James L. McClelland Center for the Neural Basis of Cognition, Carnagie Mellon University Ex-Officio member of board as chair of the Rumelhart Prize Committee

Douglas L. Medin Department of Psychology, Northwestern University 1999-2004

Naomi Miyake School of Computer and Cognitive Sciences, Chukyo University 2004-2009

Johanna Moore Human Communication Research Centre Edinburgh University 2002-2007

Michael Mozer Computer Science Department, University of Colorado 2000-2005

Linda B. Smith Department of Psychology, Indiana University 2003-2008

Barbara Tversky Department of Psychology, Stanford University 2003-2008

Richard M. Young Psychology Department, University of Hertfordshire 2002-2007

CogSci 2004 Sponsors

The Robert J. Glushko and Pamela Samuelson Foundation
Northwestern University
Office of Naval Research, Cognitive, Neural, and Social S&T Division
Defense Advanced Research Projects Agency, Information Processing Technology Office
Air Force Research Laboratory, Human Effectiveness Directorate
University of Chicago
CHI Systems, Inc.
Trends in Cognitive Sciences

About the Society

The Society is a non-profit professional organization, and its main activities are sponsoring an annual conference, publishing the journal Cognitive Science, and promoting research interactions across traditional disciplinary boundaries. The Society was incorporated as a non-profit professional organization in Massachusetts in 1979.

The Cognitive Science Society, Inc. brings together researchers from many fields that hold a common goal: understanding the nature of the human mind. The Society promotes scientific interchange among researchers in disciplines comprising the field of Cognitive Science, including Artificial Intelligence, Linguistics, Anthropology, Psychology, Neuroscience, Philosophy, and Education.

You may contact the Society at:

Cognitive Science Society, Inc.
University of Texas at Austin
Department of Psychology
1 University Station A8000
Austin, TX 78712
cogsci@psy.utexas.edu
(512) 471-2030

CogSci 2004 Reviewers

Agnar Aamodt
Herve Abdi
Mark Ackerman
Mauro Adenzato
Woo Kyoung Ahn
Adam Albright
Gary Allen
Kai Alter
Richard Alterman
Eric Amsel
Hanne Andersen
Patric Andersson
Sachiko Aoshima
Shlomo Argamon
Mark Aronoff
Kevin Ashley
Richard Aslin
Janet Astington
Harald Baayen
William Badecker
Collin Baker
Benjamin Balas
Sasha Barab
Elizabeth Baraff
Brigid Barron
Karen Bartsch
Laurie Bauer
Eric Baum
Cristina Becchio
Roman Belavkin
Aaron Benjamin
Benjamin Bergen
Robert Berwick
Susan Birch
V. Blackwell-Hardie
Mark Blair
Isabelle Blanchette
Volker Blanz
Enrico Blanzieri
Stephen Blessing
Sergey Blok
Jean-Francois Bonnefon
Lera Boroditsky
Nick Braisby
Holly Branigan
Andrew Brook
Barbara Bruschi
David Bryant
Johno Bryant
Joanna Bryson
Monica Bucciarelli
Raluca Budiu
Marc Buehner

John Bullinaria
Judee Burgoon
Denis Burnham
Bruce Burns
Jerome Busemeyer
Mike Byrne
Ellen Campana
Thomas Carr
Nicholas Cassimatis
Anne Castles
Richard Catrambone
David Chalmers
S. Chandrasekharan
Franklin Chang
Nancy Chang
Nick Chater
Xiang Chen
Zhe Chen
Patricia Cheng
Paolo Cherubini
Micki Chi
Yoonsuck Choe
Jessica Choplin
Marvin Chun
John Clapper
John Clement
Jonathan Cohen
Mark Cohen
John Coley
Eliana Colunga
Louise Connell
Anne Cook
Rick Cooper
Robin Cooper
Fintan Costello
Anna Cox
Richard Cox
Matthew Crocker
Fred Cummins
Suzanne Curtin
Florin Cutzu
Delphine Dahan
Rick Dale
Markus Damian
David Danks
Nicolas Davidenko
Janet Davidson
Jim Davies
Wim De Neys
Jean Decety
Doug DeGroot
Simon Dennis
Sharon Derry

Barry Devereux
Mike Dickey
Eric Dietrich
Kristien Dieussaert
Peter-Ford Dominey
Leonidas Doumas
Yoshimochi Ejima
Michelle Ellefson
Randi A. Engle
Blanzieri Enrico
Richard Epstein
Michael Erickson
Zachary Estes
Hany Farid
Aidan Feeney
Michele Feist
Jerome Feldman
Lisa Feldman-Barrett
David Feldon
Christopher Fennell
Ron Ferguson
Victor Ferreira
Leo Ferres
Anna Fisher
Cynthia Fisher
Joan Fisher
Alan Fiske
Nick Flor
Peter Foltz
Joao Fonseca
Kenneth Forbus
Ken Forster
Malcolm Forster
Craig Fox
Michael Frank
Nancy Franklin
Michael Freed
Eric Freedman
Daniel Freudenthal
Wai-Tat Fu
Jonathan Fugelsang
Michael G. Dyer
Liane Gabora
Christina Gagne
Susanne Gahl
Vittorio Gallese
Patricia Ganea
Linda Garro
Michael Gasser
Ronald Giere
Art Glenberg
Ashok Goel
Alvin Goldman

Louis Gomez
Charles Goodwin
Alison Gopnik
Mike Gorman
Arthur Graesser
Laura Granka
Jonathan Gratch
Jackie Griego
Zenzi Griffin
Thomas Griffiths
Maurice Grinberg
Jonathan Grudin
Sabine Gueraud
Markus Guhe
Glenn Gunzelmann
Frank Guo
Prahlad Gupta
C. Hadjichristidis
Verena Hafner
York Hagmayer
Ulrike Hahn
Rogers Hall
Christine Halverson
Joy Hanna
Andrew Hanson
David Hardman
Uri Hasson
Giyoo Hatano
Bruce Hayes
Mary Hegarty
Julie Heiser
Seth Herd
Henry Hexmoor
Graeme Hirst
Cindy Hmelo-Silver
Stephen Hockema
John Hoeks
Alex Holcombe
Keith Holyoak
Terry Horgan
Alexandra Horowitz
Autumn Hostetter
Andrew Howes
Eva Hudlicka
Bernard Huet
John Hummel
Tom Hummer
Julie Hupp
Edwin Hutchins
Mutsumi Imai
Bipin Indurkhya
Haythem Ismail
David Israel

Michael Israel
Jana Iverson
John Jacobson
Sophie Jacques
Vikram Jaswal
Heiswan Jeong
Christine Johnson
Christopher Johnson
Joe Johnson
Catholijn M. Jonker
Robert Kail
Ed Kako
Charles Kalish
Mike Kalsher
Michael Kaschak
Elham Kazemi
Miguel Kazen
Mark Keane
Fred Keijzer
Deborah Kelemen
Christopher Kello
Spencer Kelly
Charles Kemp
Trina Kershaw
Alan Kersten
Daniel Kimberg
David Kirsh
Roberta Klatzky
Celia Klin
Heidi Kloos
Pia Knoeferle
Ken Koedinger
Jean-Pierre Koenig
Boicho Kokinov
Janet Kolodner
Timothy Koschmann
Maria Kozhevnikov
Joachim Krueger
M. Krych-Appelbaum
Tevye Krynski
Sven Kuehne
Kenneth Kurtz
Howard Kurtzman
Elke Kurz-Milcke
Tamar Kushnir
Aarre Laakso
Christophe Labiouse
David Lagnado
Itziar Laka
Koen Lamberts
Pat Langley
J. Larreamendy-Joerns
Michael Lee
April Leininger
Benoit Lemaire
Ernie Lepore
Dan Levin
James Levin
Gina Levow
Rick Lewis
Linda Liu
Vincenzo Lombardo
Tania Lombrozo
Max Louwerse
Brad Love
Marsha Lovett
John Lucy
Christian Luhmann
Dermot Lynott
Lorenzo Magnani
James Magnuson
Kareen Malone
Pete Mandik
Gary Marcus
Massimo Marraffa
Massimo Marrassa
Richard Marsh
James Marshall
Bridgette Martin
Paolo Martini
Michael Masson
Rui Mata
Stefan Mateeff
Holly Mathews
Teenie Matlock
Toshihiko Matsuka
Rich Mayer
Rachel McCloy
Chris McCollum
Bob McMurray
Timothy McNamara
Nicole McNeil
Alexander Meadow
David Medler
Cristina Meini
David Melcher
Alissa Melinger
David Mendonca
Lise Menn
George Miller
Debra Mills
Toben Mintz
Daniel Mirman
Yoshio Miyake
Eva Mok
Padraic Monaghan
Chris Moore
James Morgan
Edward Munnich
Eric Murphy
Gregory Murphy
Christopher Myers
Jay Myung
Hari Narayanan
Mitchell Nathan
Daniel Navarro
Edward Necka
Hansjoerg Neth
Nora Newcombe
Elissa Newport
Wendy Newstetter
Natika Newton
Sourabh Niyogi
Timothy Nokes
Donald Norman
Catherine Norris
Robert Nosofsky
Adam November
Laura Novick
Lynne Nygaard
Tim Oates
Edward O'Brien
Stellan Ohlsson
Danny Oppenheimer
Tom Ormerod
Andrew Ortony
David Over
Praveen Paritosh
Neal Pearlmutter
David Peebles
M. Sandra Peña
Alfredo Pereira
Amy Perfors
Magnus Persson
Alexander Petrov
Rosalind Picard
David Pierce
Zygmunt Pizlo
Jodie Plumert
Kim Plunkett
Jesse Prinz
James Pustejovsky
Daniele Radicioni
David Rakison
Ashwin Ram
Michael Ranney
William Rapaport
Amnon Rapoport
David Rapp
Stephen Read
Terry Regier
Bob Rehder
Steven Reznick
Daniel Richardson
Lance Rips
Debi Roberson
Robert Roe
Ardi Roelofs
Jeremy Roschelle
Paul Rosenbloom
Karl Rosengren
Glenn Ross
Michael Rudd
Laura Sabourin
Katiuscia Sacco
Matthias Schlesewsky
Mike Schoelles
Jonathan Schooler
Lael Schooler
Wolfgang Schoppek
Walter Schroyens
Laura Schulz
Daniel Schwartz
David Schwarzkopf
Darryl Seale
Mark Seidenberg
Adrienne Seiffert
Priti Shah
Murray Shanahan
Hajime Shirouzu
Yuichi Shoda
Bradd Shore
Robert Siegler
Cynthia Sifonis
Chris Sims
Murray Singer
Jeremy Skipper
Aaron Sloman
James Slotta
Vladimir Sloutsky
Steve Smith
Jesse Snedeker
Gregg Solomon
Jessica Sommerville
Barbara Spellman
Nathan Sprague
Rohini Srihari
Craig Stark
Timo Steffens
Reed Stevens
James Stewart
Gert Storms
Claudia Strauss
Patrick Sturt
Masaki Suwa
John Sweller
Daniel Swingley
Niels Taatgen
Joanna Tai
Jean Pierre Thibaut
Maurizio Tirassa
Peter Todd
Russell Tomlin
Greg Trafton

Roy Turner
Peter Turney
Ryan Tweney
Mieko Ueno
M. Afzal Upal
Andrea Valle
Ezra Van Everbroeck
Michiel Van Lambalgen
Christof van Nimwegen
John van Opstal
Scott VanderStoep
Kurt VanLehn
Vladislav Veksler
Michael Verde
Gail Viechnicki
Gabriella Vigliocco
Carl Vogel
Eric-Jan Wagenmakers
Muhammad Walji
Clare Walsh
Hongbin Wang
Pei Wang
Thomas Ward
Duane Watson
Sandra Waxman
Dan Weiskopf
Drew Westen
David Whitney
Katja Wiemer-Hastings
Peter Wiemer-Hastings
Anna Wierzbicka
Eric Wiewiora
Jennifer Wiley
Edwin Williams
Joe Williams
William Wimsatt
Lauren Wineburgh
Edward Wisniewski
Phil Wolff
Patrick Wong
Nicholas Wymbs
Fei Xu
Takashi Yamauchi
Yingrui Yang
Daniel Yarlett
Hanako Yoshida
Richard Young
Jeff Zacks
Gregory Zelinsky
Jiajie Zhang
Tom Ziemke
Rolf Zwaan
Rami Zwick

Tutorials

August 4, 2004

Contents

Rumelhart Symposium

Symposia

Publication-based Talks

Papers

Member Abstracts

Tutorials

CHREST Tutorial: Simulations of Human Learning

Fernand Gobet (fernand.gobet@brunel.ac.uk)
Department of Human Sciences, Brunel University,
UXBRIDGE, Middlesex, UB8 3BH, U.K.

Peter C. R. Lane (peter.lane@bcs.org.uk)
School of Computer Science, University of Hertfordshire,
College Lane, HATFIELD, Hertfordshire, AL10 9AB, U.K.

Abstract

CHREST (Chunk Hierarchy and REtrieval STructures) is a comprehensive, computational model of human learning and perception. It has been used to successfully simulate data in a variety of domains, including: the acquisition of syntactic categories, expert behaviour, concept formation, implicit learning, and the acquisition of multiple representations in physics for problem solving. The aim of this tutorial is to provide participants with an introduction to CHREST, how it can be used to model various phenomena, and the knowledge to carry out their own modelling experiments.

Developing detailed process models of cognitive phenomena is important to the development of cognitive science, as only then can cognitive theories be used to generate quantitative predictions for complex phenomena. The history of computational modelling includes many diverse approaches, from models of single phenomena (such as Young and O'Shea's model of subtraction), to integrated models covering a wide range of different phenomena (such as Soar and ACT-R), to over-arching principles, which guide the development of models in disparate domains (e.g. connectionist approaches, or embodied cognition).

The EPAM/CHREST tradition, which forms the heart of this tutorial, has been providing significant models of human behaviour since 1959. Early models of EPAM provided the impetus to develop the chunking theory, which has been an important component in theories of human cognition ever since. Focusing on learning phenomena, EPAM and CHREST place a great emphasis on how the model's information is learnt through interactions with an external environment. Thus, EPAM/CHREST models are typically developed from large quantities of naturalistic input. For example, in modelling expert perception of chess players, actual chess games are used.

Historically, CHREST is derived from the EPAM (Elementary Perceiver and Memorizer) model of Feigenbaum and Simon (1984). In both models, learning happens as the creation and elaboration of a discrimination network. In addition, CHREST has mechanisms for the automatic creation of schemata and for the creation of 'lateral links', which can be used for creating elementary productions or elementary semantic links. CHREST can thus be situated between production systems such as Soar and connectionist systems. Just as EPAM was the computational embodiment of the key aspects of the chunking theory (Chase & Simon, 1973), CHREST implements the essential aspects of the template theory (Gobet & Simon, 2000). In spite of its historical and contemporary importance, and the diversity of domains in which modelling has been successfully carried out, the number of people who use or understand the principles and operation of an EPAM/CHREST model remains small.

The tutorial is structured so that participants will:

1. Acquire a comprehensive understanding of the CHREST computational model and its relation to the chunking and template theories of cognition;
2. Explore some key learning phenomena supporting the chunking theory by taking part in a verbal-learning experiment;
3. Attempt to match their own data with the performance of a CHREST model of verbal learning; and
4. Be introduced to the implementation of CHREST in sufficient detail to begin modelling their own data.

We have chosen a verbal-learning experiment (serial-anticipation method) for introducing participants to CHREST for the following reasons: the experiment is historically important; it was one of the motivations behind the development of EPAM; it can be carried out in a short period of time; striking learning phenomena are readily observable, in spite of the brevity of the experiment; the motivation and requirements for the experiment are generally clear; and, finally, it illustrates some key features of the EPAM/CHREST architecture.

References

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 55–81.

de Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess. Heuristics of the professional eye.* Assen: VanGorcum.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305–336.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C-H., Jones, G., Oliver, I. & Pine, J. M. (2001). Chunking mechanisms in human learning. *TRENDS in Cognitive Sciences*, 5, 236–243.

Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science*, 24, 651–682.

ACT-R Tutorial

Niels A. Taatgen (taatgen@cmu.edu)
Psychology, Carnegie Mellon University
5000 Forbes Av., Pittsburgh, PA 15213 USA

ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere & Qin, in press) is a cognitive theory and simulation system for developing cognitive models. It assumes cognition emerges through the interaction of a procedural memory of productions with a declarative memory of chunks and independent modules for external perception and actions. Since its release in 1997, ACT-R has supported the development of over 100 cognitive models, published in the literature by many different researchers. These models cover topics as diverse as driving behavior, implicit memory, learning backgammon, metaphor processing, and emotion. We have recently developed a new version, ACT-R 5.0 that is more interruptible, achieves greater across-task parameter consistency, has better mechanisms of production learning, and is more in correspondence with our knowledge of brain function. The tutorial has no prerequisite knowledge, and is intended to on the one hand give an overview of the theory, and on the other hand offer some direct demonstration of ACT-R models. Although a half day is not sufficient to cover all material, it can wet the appetite for the full ACT-R tutorial that is available on http://act-r.psy.cmu.edu/. This website also provides for the necessary software, and overview of researchers using ACT-R, and it has a list of ACT-R publications (many of them downloadable).

During the tutorial, following Taatgen, Lebiere and Anderson (submitted) five popular research paradigms within ACT-R will be used as a vehicle to explain the architecture:

Instance learning

Learning by retrieving old experiences from memory, similar to Logan's instance theory.

Utility learning

Learning which of several available strategies is optimal by keeping track of costs and probability of success.

Working Memory Capacity

Models in which the amount of spreading activation is varied, which can explaining individual differences in working memory capacity

Perceptual/Motor constrained processing

Models in which the main factor in explaining human performance lies in the limitations of their perceptual and motor systems.

Rule learning

Models in which new production rules are learned on the basis of combination of old rules and substitution of declarative knowledge.

Although these individual research paradigms have produced interesting models by themselves, the full potential of the architecture can only be seen when they work together in models of complex cognition.

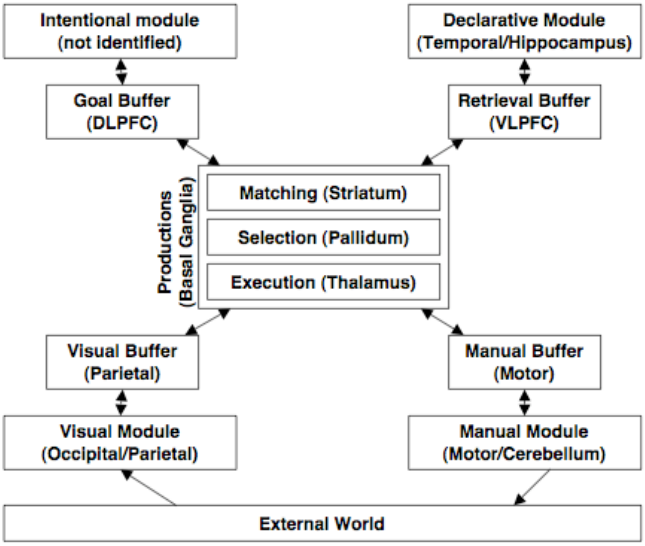


Figure 1: Overview of the architecture

References

Anderson, J. R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y. (in press) An integrated theory of Mind. *Psychological Review*. Available online: http://act-r.psy.cmu.edu/papers/403/IntegratedTheory.pdf

Taatgen, N.A., Lebiere, C. & Anderson, J.R. (submitted). Modeling paradigms in ACT-R. In R. Sun (ed.), *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press. Available online: http://www.ai.rug.nl/~niels/publications/taatgenLebiereAnderson.pdf

Development of Executable Cognitive Agents Using the COGNET Architecture and *i*GEN™ Toolset

Wayne Zachary (wzachary@chisystems.com)
CHI Systems, Inc.
1035 Virginia Drive, Fort Washington, PA 19034-3701

Michael A. Szczepkowski (mszczepkowski@chisystems.com)
CHI Systems, Inc.
1035 Virginia Drive, Fort Washington, PA 19034-3701

COGNET/*i*GEN™ is a set of software tools (i.e., a workbench) that enables human performance specialists to develop, test and deploy *cognitive agents* -- software components that exhibit a level of intelligence that mimic human thought processes. Cognitive agents represent the logical transition of research on human information processing to practical application. Cognitive agents also represent a new and growing paradigm for research in decision support, intelligent human-computer interfaces, intelligent tutoring, etc. From an application perspective, cognitive agents empower the user by combining the speed, efficiency and accuracy of the computer with the decision-making capacity, experience and expertise of human experts. From a research perspective, they allow cognitive models to be applied to problems of enhancing the interaction between people and information technology in complex work environments.

COGNET/ *i*GEN™ incorporates computational models of human cognitive processes as a basis for designing and building software agents. At the same time, COGNET/ *i*GEN™ incorporates many practical approaches from software and systems engineering to maximize its ability to meet real-world cognitive agent application needs. This makes it fundamentally different from cognitive architectures, which have been developed as vehicles to test cognitive theories (e.g., ACT-R and theories of memory; EPIC and theories of dual tasking and task performance).

This workshop introduces participants to the concepts of cognitive agents and to the COGNET/ *i*GEN™ method and tools for cognitive agent development and prepares them to undertake the development of cognitive agents applications. This tutorial provides COGSCI attendees with a view of the COGNET/ *i*GEN™ cognitive architecture that emphasizes the unique properties of COGNET/ *i*GEN™. It also provides an introduction to the concepts and methods involved in cognitive agents and their development, providing participants with an important perspective linking theory and practice.

The workshop begins with an examination of three major uses for cognitive agents:

- Intelligent training and tutoring – specifically, the use of cognitive agents to provide:
  - embedded models of the student/trainee to track student progress against the knowledge required for the skill being trained;
  - an embedded instructor/tutor that can manage presentation of information, sequence instruction, and provide feedback/remediation; and/or
  - synthetic teammates to facilitate practice and teamwork in a simulated work environment.
- decision support and electronic performance support, in the form of work-centered intelligent interfaces that assist a worker or decision maker in such functions as attention management, situation awareness, and/or contextualizing decision strategies;
- human performance simulation, in the form of simulations of system users to aid design engineering and design evaluation, and/or synthetic players for mission simulations and/or interactive games.

Examples of each type are provided.

The workshop then covers COGNET as a cognitive-agent architecture based on cognitive theory. COGNET is compared to other computational architectures that embody theories of human thought and reasoning, and the features of COGNET that support cognitive agent development are identified. Particular focus is given to the features unique to COGNET/ *i*GEN™ – including metacognition, flexible granularity, and expert-level knowledge structures – and to the constructs that specifically focus on cognitive agent requirements – temporal management, micromodels, parallel execution threads, and external application interfaces. The modeling strategies by which the system can be used to represent complex behaviors such as teamwork, coaching, and cognitive workload self-reporting are discussed, as are the development tools available to support modeling and the integration of models into larger simulations, federations, or other applications.

For more information on the COGNET methodology and *i*GEN™ toolset, go to http://www.cognitiveagent.com.

Rumelhart Symposium

ACT-R as a Unified Architecture of Cognition: A Symposium in Honor of John R. Anderson

Organizers: Kevin Gluck (kevin.gluck@mesa.afmc.af.mil) Air Force Research Laboratory
Wayne D. Gray (grayw@rpi.edu) Rensselaer Polytechnic Institute

This symposium highlights the utility of the ACT-R theory as a tool for basic and applied cognitive science research. Four colleagues of this year's Rumelhart Prize winner, John R. Anderson, will describe how his ACT-R theory has enabled and inspired their research.

An Activation-Based Model Of Sentence Processing As Skilled Memory Retrieval

Richard Lewis (rickl@umich.edu)
University of Michigan

This talk presents a theory and ACT-R-based model of human sentence comprehension that embodies the following claim: Sentence comprehension consists of a series of cue-based retrievals from short-term (and long-term) memory, subject to similarity-based interference and activation decay. ACT-R does not merely serve as an implementation language for the theory; rather it serves as the vehicle for bringing sentence processing into detailed contact with general and independently established principles of memory and skilled behavior. These principles, together with some representational assumptions from linguistic theory, provide explanatory accounts of many parsing phenomena (such as difficulty on embeddings and locality and recency effects), and generate novel predictions which can be empirically tested. Experimental work on several languages is summarized. The work includes experiments that examine the effects of interference and decay on distinct components of sentence processing, experiments that distinguish activation decay and distance-based accounts, and experiments that begin to answer one of the basic questions motivated by the theoretical framework: exactly what kinds of similarity (syntactic, semantic, phonological, positional) matters in sentence processing?

This is joint work with Shravan Vasishth, Julie Van Dyke, and JJ Nakayama. No linguistic or psycholinguistic background will be assumed.

Information Foraging Theory And The Rational Analysis Of Human-Information Interaction

Peter Pirolli (pirolli@parc.xerox.com)
Xerox PARC

Information foraging theory has been developed to provide rational analyses of how people adapt to the task of obtaining useful information to meet their ongoing goals, and how technology can be better designed to improve human-information interaction. Several ACT-R models have been developed to model human-information foraging theory. The information foraging approach will be illustrated by recent models of interaction with the Web.

ACT-R and Driving

Dario D. Salvucci (salvucci@cs.drexel.edu)
Drexel University

As a complex but ubiquitous task, driving serves as an excellent domain for both testing and applying cognitive architectures such as ACT-R. For several years we have worked to model driver behavior in ACT-R with two main branches of research. First, we have developed a computational integrated driver model that navigates realistic highway environments, accounting for various aspects of driver behavior including eye movements between regions of the visual environment and steering profiles through curves and lane changes. Second, we have worked to predict the potential for driver distraction from secondary tasks such as dialing cellular phones or even primarily cognitive tasks. Both lines of research are helping to shape a more general theoretical account of human multitasking within the ACT-R architecture, and at the same time, demonstrating how such architectures can facilitate the development of practical tools that infer driver intentions and predict driver distraction for new in-vehicle devices.

Embedded Cognition through Production Compilation

Niels A. Taatgen (taatgen@cmu.edu)
Carnegie Mellon University

Embedded cognition refers to the fact that our reasoning processes are not only driven by plans and goals in our heads, but also, or maybe even mainly, by interacting with the environment. The challenge is to develop a cognitive system that strikes the right balance between being driven by its internal goals and by opportunities in the environment. In order to explore this challenge, we have looked at a particular subdomain: following instructions. In experimental tasks participants typically are given instructions in a list-like fashion. For many tasks, strictly following the instructions will lead to brittle and rigid behavior. We will look at two examples that represent two ends of the spectrum of complexity: a basic dual-task and a complex real-time dynamic task. Both tasks have been modeled in the ACT-R cognitive architecture using the production compilation mechanism to achieve embeddedness.

Symposia

Qualitative Modeling and Cognitive Science

Gautam Biswas (Biswas@vuse.vanderbilt.edu)
EECS Department, Vanderbilt University
Box 1679 Station B, Nashville, TN 37325, USA

Bert Bredeweg (bert@swi.psy.uva.nl)
Department of Social Science Informatics, University of Amsterdam
Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

Ronald W. Ferguson (rwf@cc.gatech.edu)
College of Computing, Georgia Institute of Technology
801 Atlantic Avenue, Atlanta, GA 30332

Peter Struss (struss@in.tum.de)
Institut für Informatik, Technische Universität München
Boltzmannstr.3, 85748 Garching b. München, Germany

Bruce Sherin (bsherin@northwestern.edu)
School of Education and Social Policy, Northwestern University
Annenberg Hall, 2120 Campus Drive, Evanston, IL, 60208, USA

Motivation

Qualitative reasoning research creates computational models that capture aspects of reasoning about continuous systems, including space, time, and dynamics. It has tackled problems ranging from understanding human mental models of everyday systems to creating systems that can do engineering design and create scientific models. The original motivations for the field came mostly from cognitive science: Creating accounts of human causal reasoning about physical systems. However, over time the fields have grown more separate. This symposium is part of a bridge-building effort, to create more dialogue between these two communities for their mutual benefit.

Gautam Biswas: We have been developing computer-based learning systems where students teach computer agents. These teachable agents provide important structures to help shape the thinking of the *learner-as-teacher*. Each agent manifests a visual structure that is tailored to a specific form of knowledge organization, and has related underlying qualitative reasoning mechanisms that helps the agent interact with the learner, and provide explanations on how well it has understood the material it has been taught. This framework lets us build on well-known teaching interactions that organize student activity (e.g., teaching by "laying out," teaching by example, teaching by telling, teaching by modeling), and keep the start-up costs of teaching the agent very low (as compared to programming). We illustrate the effectiveness of our approach through Betty's Brain, a teachable agent that makes her qualitative reasoning visible through a concept map.

Bert Bredeweg: Conceptual models are an important means for developing and communicating knowledge, particularly concerning the behavior of (physical) systems. But how can the use of such models be facilitated and adapted to the specific needs of people via software? Part of the answers lies in the use of qualitative models and their simulations. Such models provide a rich ontology to capture conceptual models of how humans explain system behavior. Now that these techniques are reasonably well understood interesting questions emerge concerning the automated use of them in aiding learners to develop, share, and communicate knowledge. This requires the design of graphical workbenches to work with such models and the development of smart agents that support a learner in "doing the right thing."

Ronald W. Ferguson: Diagrammatic reasoning is cognitively interesting because it involves the interaction of our powerful and seemingly task-independent visual system with knowledge about culturally-specific cognitive artifacts. Interpretation of new diagram types must be learned, but once learned, take on the character and ease of perception. Recent research in Qualitative Spatial Reasoning (QSR), which examines the development and inferential power of spatial relationships, may help explain why diagrammatic reasoning works this way, and even allow us to characterize what makes a diagram effective. We explore this claim using our GeoRep system as an example of integrating QSR with cognitive models of vision and problem-solving.

Peter Struss: Helping to understand ecological systems, to analyze the reasons for the deterioration of our environment and climate and to propose counteractions provides an important challenge to knowledge-based systems. Modeling artifacts and diagnosing why and how man-made devices fail to perform as expected is an important application area of model-based and qualitative reasoning which is based on a rigorous logical theory. Extending this foundation to natural systems turns out to be not straightforward, emphasizes the needs for conceptual and qualitative modeling formalisms, and the problem solving techniques face a different level of complexity. We will discuss the challenges and directions of potential solutions using an example at the transition between technical and natural systems, namely water treatment processes.

Bruce Sherin: Bruce Sherin's research includes investigations of conceptual change in science and external representations in science and mathematics. Bruce will serve as the discussant for the Symposium.

Language and Thought

Lera Boroditsky (lera@psych.stanford.edu)
Department of Psychology, Stanford University
Jordan Hall, 450 Serra Mall, Stanford, CA 94305, USA

John Lucy (j-lucy@uchicago.edu)
Department of Psychology, University of Chicago
5848 S. University Ave, Chicago, IL 60637

Jill DeVilliers (jdevil@science.smith.edu)
Department of Psychology, Smith College
Northampton, MA 01063 USA

Phil Wolff (pwolff@emory.edu)
Emory University Department of Psychology
532 N. Kilgo Cir., Atlanta, GA 30322

Motivation

After decades of neglect, the language and thought hypothesis—that the language we speak may influence the way we think—has recently enjoyed a resurgence of interest. This new wave of language and thought research capitalizes on gains in our linguistic knowledge of cross-linguistic semantic patterns and in the range and subtlety of psychological techniques now available. This new research has tackled a wide range of content areas, including space, time, motion, causality, the nature of the object concept, and theory of mind. This symposium presents methods and recent findings in this arena.

Lera Boroditsky

Different languages divide the color spectrum in different ways. Does this lead speakers of different languages to perceive colors differently? Results of several experiments suggest that color language can influence people's color judgments even in conditions when all color stimuli are present at the same time and need not be stored in memory. Language appears to be involved online during simple color-discrimination tasks such that effects of language can be selectively disrupted by verbal interference (but not spatial interference). Further, color discrimination performance across a boundary that exists in one language but not another can be altered by linguistic interference only for the language group that codes that linguistic distinction. Finally, as the color discrimination tasks become simpler and faster, effects of linguistic interference disappear. These results suggest that language is involved online in a large number of low-level perceptual discriminations, but also that not all color discrimination is affected by language.

Jill DeVilliers

This paper will address the issue of possible connections between language development and thought using the case of Theory of Mind. There are multiple ways to construe connections between language and false belief reasoning, only some of them causal. The argument will be that the data are compatible with a causal connection for a specific aspect of language knowledge linked to recursion of sentences, but that this does not rule out other potential facilitatory effects of language.

John Lucy

Researchers in language and thought must constantly deal with the problem of devising truly matched instruction procedures. This paper discusses the efforts of Lucy and his colleagues to develop a nonverbal triads procedure: that is, a procedure that avoids using words like “same,” “similar,” “(more) like’” etc. The resulting procedure then does not depend on using translation equivalents in other languages, and offers a useful alternative for working with children and with deaf subjects. Some preliminary empirical results from new work will be described.

Phil Wolff

This research is on causal verbs and reasoning across languages. The concept of CAUSE has frequently been treated as a conceptual primitive in the linguistic, philosophical and psychological literatures. Given this assumption, one might expect that the meaning of words encoding the concept of CAUSE should be relatively consistent across languages. In contrast to this assumption, This research shows how the meaning of the verb “cause” and related verbs may differ significantly across languages, e.g., English, Russian, and German. In addition, it suggests that these differences in meaning might reflect underlying differences in the way causal events are categorized non-linguistically.

Discussant TBA

Symposium: The Diversity of Conceptual Combination.

MODERATOR

Fintan Costello (Fintan.Costello@ucd.ie),
Department of Computer Science, University College Dublin,
Dublin, Ireland.

SPEAKERS

Fintan Costello (Fintan.Costello@ucd.ie),
Department of Computer Science, University College Dublin,
Dublin, Ireland.

Christina Gagne (cgagne@ualberta.ca),
Department of Psychology, University of Alberta,
Edmonton, Alberta.

Zachary Estes (estes@uga.edu),
Psychology Department, University of Georgia,
Athens, Georgia.

Edward Wisniewski (edw@uncg.edu),
Department of Psychology, University of North Carolina,
Greensboro, North Carolina.

Introduction

A fundamental aspect of everyday language comprehension is the interpretation of novel compound phrases through conceptual combination: a mechanism that is engaged whenever people interpret phrases like "sand gun", "cactus fish" or "pet shark". Conceptual combination is a diverse and complex cognitive process: people are able to combine concepts in a variety of different ways (for example, a "sand gun" is a tool that sprays sand, while a "cactus fish" is a fish with prickly spines, and a "pet shark" is a shark which is also a pet). This diversity is reflected in the number of quite different theories of conceptual combination that have recently been proposed by, for example, Wisniewski (Wisniewski, 1997), Gagné (Gagné & Shoben, 1997), Estes (Estes & Glucksberg, 2000), and Costello (Costello & Keane, 2000). The aim of this symposium is to gather current researchers on conceptual combination to discuss both the diversity of ways in which concepts can combine, and the diversity of theories that have been put forward to account for conceptual combination.

Diversity of Combination Types

Combined concepts are often divided into three types: relational combinations (such as "sand gun"), which assert a relation linking the two concepts being combined; property combinations (such as "cactus fish"), which transfer a property from one concept to the other; and conjunctive combinations (such as "pet shark"), which describe something that is an example of both combining concepts. These types are quite loose, however, and are by no means definitive or all-inclusive. In this symposium, speakers will address questions such as

- Why do concepts combine in different ways?
- How significant are the different combination types?
- Are some combination types more important than others?

Relationship between Theories of Combination

Recent theoretical accounts of conceptual combination are strikingly different from each other, ranging from Gagne's CARIN theory (which uses a standard set of 16 relational templates such as X-HAS-Y or X-ABOUT-Y to interpret compound phrases), to Wisniewski's Dual-Process theory (which suggests that compound interpretation involves both a scenario-construction mechanism and a structural-alignment mechanism similar to that used in analogies), to Costello's Constraint theory (which describes conceptual combination as a process of constraint satisfaction subject to the pragmatic requirements of communication using compound phrases). Symposium speakers will address questions such as

- Why are the various theories of combination so different?
- What common ground do these theories share?
- How do these theories relate to each other?
- Can we come up with an integrating framework to unite these theories?

Conclusion

By bringing together researchers taking different approaches to conceptual combination, this symposium will give a useful synthesis of the current state of conceptual combination research. By directly addressing the diversity of concept combination, the symposium may provide the basis for a more unified view of this important and fascinating part of human thought and language.

References

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: Constraint guided conceptual combination. *Cognitive Science,24(2).*

Estes, Z. & Glucksberg, S. (2000). Interactive property activation in conceptual combination. *Memory & Cognition, 28,* 28-34.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23* (1), 71-87.

Wisniewski, E. J. (1997). When concepts combine. *Psychnomic Bulletin & Review*, 4, 167-183.

Cognitive Processing Effects of 'Social Resonance' in Interaction.

Susan Duncan (deng@uchicago.edu), Amy Franklin (alfrankl@midway.uchicago.edu), Fey Parrill (feyparr@uchicago.edu), Haleema Welji (haleemaw@uchicago.edu)
Psychology Department, University of Chicago
5848 S. University Avenue, Chicago, IL 60637

Irene Kimbara (ikimbara@uchicago.edu)
Linguistics Department, University of Chicago
1010 E. 59th Street, Chicago, IL 60637

Rebecca Webb (webb@bcs.rochester.edu)
Department of Linguistics, University of Rochester,
Lattimore Hall, Rochester, NY 14534

This symposium will consist of four presentations of recent work that examine the cognitive aspects of 'social resonance' in interaction. By social resonance, we mean the state of being that pertains when individuals engaged in face-to-face communication feel strongly connected. The studies we will present elaborate and extend a wealth of findings, primarily from social psychology and anthropology, on a set of closely related concepts; e.g., 'interpersonal sensitivity' (Hall & Bernieri, 2001), 'social glue' (Lakin, et al., 2003), 'interactional synchrony' (Bernieri & Rosenthal, 1991), empathy (Sonnby-Borgstrom, et al. 2003, Nishio, 2002), and 'socially distributed cognition' (DuBois, 2000). In choosing the term SOCIAL RESONANCE, we seek to emphasize the *dyad* or *group* as a focal level of analysis and theory. Resonance also evokes the dynamic interplay of behaviors, multi-modally, between participants in real time interactions. The social psychological notion of 'rapport,' as well as recent-era reformulations of anthropologist Malinowki's (1923) construct, 'phatic communion,' resemble the sense of social resonance we employ here:

> "… all the different communicative strands, speech, gesture, posture, body movements, orientation, proximity, eye contact, and facial expressions … woven together to form the fabric of conversation," (Laver, 1975).

Social resonance has been the focus of research across quite a range of scientific disciplines. Its presence in interaction is ascertained on the basis of behavioral observables that occur singly and in clusters, or in patterns of alternation or sequential unfolding across an interval of interaction. Examples include mimicry or 'mirroring' of various nonverbal behaviors, interactional rhythm in conversation, and syntactic priming. The work presented here takes up where many social psychological and anthropological treatments of resonance phenomena leave off, by examining its impact on real-time cognitive processing; specifically, on processes of language production and comprehension. Our studies illuminate, through close analysis of the multi-modal (speech and gesture) aspects of human interaction, what psycholinguistic processing mechanisms may be triggered by, facilitated by--or, conversely---remain inactive, or even be inhibited by, respectively, presence or absence of social resonance.

Kimbara and Parrill present recent work on the interaction between gestural mirroring (specifically, cases in which a speaker appears unconsciously to adjust her use of gesture space, in accord with her interlocutor's use of space, to create matching nonverbal expression of spatial layouts) and 'structural priming' (Bock, 1986), or, the ways in which one speaker's choice of syntactic construction influences the subsequent choice of construction on the part of her interlocutor. Webb examines types of metaphoric gestures (McNeill, 1992), spontaneously produced by speakers addressing an audience, with respect to their roles in managing the interactive context (Bavelas, 1992). Welji and Duncan examine gestures in spontaneous narrative discourse, comparing interlocutors who are strangers (less socially resonant) with interlocutors who are friends (more resonant). Franklin and Duncan demonstrate perturbation of social resonance, as a function of cognitive load, in dyads where one member is instructed to deceive the other about the visual and narrative content of a cartoon story eliciting stimulus, viewed on video.

Selected References

Bavelas, J.B., et al. (1992). Interactive gestures. Discourse Processes, 15:469-489.

Bernieri, F.J. & Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. In R.S. Feldman & B. Rimé (eds.), Fundamentals of Nonverbal Behavior. Cambridge: Cambridge Univ. Press.

Bock, J.K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355-387.

Lakin, J.L., Jefferis, V.E., Cheng, C.M., & Chartrand, T.L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *J. Nonverbal Behavior, 27(3)*, 145-162.

McNeill, David (1992). Hand and Mind: What Gestures Reveal about Thought. Chicago: Univ. of Chicago press.

Vrij, et al. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *J. Nonverbal Behavior, 24(4)*, 239-263.

CogSci2004 Symposium

Abduction and Creative Inferences in Science

Lorenzo Magnani (lorenzo.magnani@unipv.it) - Organizer, University of Pavia, Italy
Atocha Aliseda (atocha@filosoficas.unam.mx), UNAM. México City, México
Thomas Addis (tom.addis@port.ac.uk**)** and **David Gooding** (hssdcg@bath.ac.uk),
University of Portsmouth, Portsmouth, UK and University of Bath, Bath, UK
John Woods (jhwoods@interchange.ubc.ca) and **Dov Gabbay** (dg@dcs.kcl.ac.uk),
University of British Columbia, CA and King's College London, UK
Joke Meheus (joke.meheus@rug.ac.be), Ghent University, Ghent, Belgium
Matti Sintonen and **Sami Paavola** (matti.sintonen@helsinki.fi,sami.paavola@helsinki.fi,
- Discussants, University of Helsinki, Finland

The symposium aims to explore abduction (inference to explanatory hypotheses), an important but neglected topic in scientific reasoning. The aim is to integrate philosophical, cognitive, and computational issues. The main thesis is that abduction is a significant kind of scientific reasoning, helpful in delineating the first principles of a new theory of science. The status of abduction is very controversial. When dealing with abductive reasoning misinterpretations and equivocations are common. What are the differences between abduction and induction? What are the differences between abduction and the well-known hypothetico-deductive method? What did Peirce mean when he considered abduction a kind of inference? Does abduction involve only the generation of hypotheses or their evaluation too? Are the criteria for the best explanation in abductive reasoning epistemic, or pragmatic, or both? How many kinds of abduction are there? The symposium aims to increase knowledge about creative and expert inferences. The study of these high-level methods of abductive reasoning is situated at the crossroads of philosophy, epistemology, artificial intelligence, cognitive psychology, and logic; that is at the heart of cognitive science.

More than a hundred years ago, the great American philosopher Charles Sanders Peirce coined the term "abduction" to refer to inference that involves the generation and evaluation of explanatory hypotheses. The study of abductive inference was slow to develop, as logicians concentrated on deductive logic and on inductive logic based on formal calculi such as probability theory. In recent decades, however, there has been renewed interest in abductive inference from two primary sources. Philosophers of science have recognized the importance of abduction in the discovery and evaluation of scientific theories, and researchers in artificial intelligence have realized that abduction is a key part of medical diagnosis and other tasks that require finding explanations. Psychologists have been slow to adopt the terms "abduction" and "abductive inference" but have been showing increasing interest in causal and explanatory reasoning.

Thus abduction is now a key topic of research in philosophy of science. First, this symposium ties together the concerns of philosophers of science and logicians, showing, for example, the connections between formal models and abduction (Meheus, Woods and Gabbay). Second, it lays out a useful general framework for discussion of various kinds of abduction (Magnani), such as model-based and manipulative abductions. Third, it develops important ideas about aspects of abductive reasoning that have been relatively neglected in philosophy of science, including the role of testing in abductive inference (Aliseda), and the interrogative model of inquiry and the role of different kinds of why-questions and strategic principles employed in attempts to find and construct answers also at the computational level (Sintonen and Paavola, Addis and Gooding). The clarification of these topics aims to increase knowledge about some aspects of explanatory reasoning and hypothesis formation very relevant in many epistemic tasks.

1. If we stress the concept of *model-based and manipulative abduction* (Magnani), creative inferences in science can be seen as formed by the application of heuristic (strategic) procedures that involve all kinds of good and bad inferential actions and both internal and external representations, and not only the mechanical application of rules.

2. Recent *logical models* can illustrate in a rigorous way how these (strategic) abductive steps are combined with deductive steps (Meheus, Woods and Gabbay).

3. Common to all abduction problems is a cognitive target that cannot be hit on the basis of what the abducer presently knows. Abductive hypotheses do not enhance a reasoner's knowledge. Abduction, accordingly, is ignorance-preserving inference. These abductive processes are dynamical (Woods and Gabbay).

4. The "abductive steps" are also analyzable in terms of responses to surprising singular or general facts, showing a connection to *explanation-seeking why-questions* (Sintonen and Paavola).

5. The importance of *experimental verification for hypotheses evaluation* in science is stressed by the relationship between abduction and pragmatism in Peirce (Aliseda).

6. Abduction cannot be thought of in isolation from the two other type of inference (deduction and induction/validation) identified by Peirce. Computer models of scientific behaviour and music conversation suggest that in simulation of abduction requires the use of mixed strategies using random actions as suggested by game theory (Addis and Gooding).

Cognitive Neuroscience: What does it tell us about high-order cognition?

Jay McClelland (jlm@cnbc.cmu.edu)
Carnegie Mellon University

Ken A. Paller (kap@northwestern.edu)
Northwestern University

Paul J. Reber (preber@northwestern.edu)
Northwestern University

Mark Jung-Beeman (mjungbee@northwestern.edu)
Northwestern University

Andrew Ortony (ortony@northwestern.edu)
Northwestern University

Motivation

Cognitive neuroscience has generated considerable excitement among cognitive scientists. It offers a new route to understanding the mind. But some have argued that the current techniques of cognitive neuroscience apply mainly to perceptual-motor and attention tasks. This symposium considers the contribution cognitive neuroscience can make to the study of high-order cognition.

Jay McClelland

Formulating cognitive theories in terms of neural computations

Cognitive Neuroscience is a relatively new field defined differently by different people. One activity that might fall under the rubric of cognitive neuroscience is the effort to formulate cognitive theories in terms of the underlying neural computations. This effort, I will suggest, may ultimately lead to theories that are not only grounded in neuroscience but also more satisfactory as accounts of cognitive phenomena. I will exemplify the approach with my own work on the development of the complementary learning systems theory of learning and memory. The work draws on insights from neurophysiology, neuroanatomy, neuropsychology, and connectionist modeling to develop an overall theory of learning and memory

Ken Paller

A cognitive neuroscience perspective on memory with and without awareness

Patterns of memory impairment in patients with amnesia suggest that memory for facts and events depends on a process of "cross-cortical storage" that is not required for other forms of memory. This neuropsychological evidence provides a theoretical foundation for understanding memory phenomena like recollection (the conscious experience of remembering facts and events) and priming (a type of item-specific implicit memory). Building on this foundation, measures of brain activity have revealed distinct brain potentials and brain activations that are associated with recollection of episodic memories versus certain types of priming. This cognitive neuroscientific approach thus constitutes an appropriate way to investigate the neurocognitive events that make "memory-with-awareness" so different from "memory-without-awareness."

Paul J. Reber

Cognitive (neuro)science: Models of recognition and categorization

Categorization and recognition have been investigated with a variety of approaches that have led to very different conclusions. Dissociations reported in the patient literature suggest separate representational systems. A resolution is proposed based on integrating key computational features of Nosofsky & Zaki's (1998) exemplar-based model with the Complementary Learning Systems (CLS) theory, a neural network model of memory system organization (McClelland, McNaughton & O'Reilly, 1995). Functional neuroimaging studies of categorization and recognition with healthy participants (Reber et al., 1998a,b, 2003) provide evidence supporting the two-system CLS theory. Integrating computational approaches with both behavioral and neuroscience data thus leads to a better account of recognition and categorization than any single approach.

Mark Jung-Beeman

Imaging higher-order language comprehension and insight problem solving: What and how from where?

Neuroimaging of higher-order cognition offers two advantages: First, it can provide a relatively covert measure of processing; and second, cortical location information can constrain or expand cognitive theories regarding component processes. In one set of studies, we used fMRI signal in specific cortical regions as markers for two component processes in drawing inferences: semantic integration, and semantic selection. We thus observed the engagement of these processes while people comprehended stories, without requiring a concomitant "probe task." In another line, we studied insight processes in verbal problem solving and found involvement of an area of the right temporal lobe, similar to that observed in the inference experiments.

Andrew Ortony (Discussant)

Andrew Ortony's work has ranged from computational modeling to philosophy. His current research is on emotion and cognition.

Large-scale Knowledge Representation Resources for Cognitive Science Research

George A. Miller
Department of Psychology, Princeton University
1-S-5 Green Hall, Princeton, NJ 08544 USA

Charles J. Fillmore
International Computer Science Institute, University of California at Berkeley
1947 Center St., Suite 600, Berkeley, CA 94704-1198 USA

Martha S. Palmer
Department of Computer and Information Science, University of Pennsylvania
3330 Walnut Street, Philadelphia, PA 19104-6389 USA

Doug Lenat
Cycorp, Inc.
Suite 100, 3721 Executive Center Drive, Austin, TX 78731 USA

Pat Hayes
Institute for Human and Machine Cognition
40 South Alcaniz Street, Pensacola, FL 32502

Motivation

One of the unique features of cognitive science has been its emphasis on understanding how people represent and use knowledge. Unfortunately, knowledge representation can be quite difficult if one is starting from scratch. Having "off the shelf" representation systems that can be used for investigations can make new types of cognitive modeling efforts possible, at a scale that would otherwise be impossible. The participants in this symposium each represent a different approach to building such representational resources.

George A. Miller: WordNet is a lexical database that contains 146,900 English nouns, verbs, adjectives, and adverbs that are now organized by semantic relations into 117,500 meanings, where a meaning is represented by a set of synonyms (a synset) that can be used to express that meaning. An entry in WordNet consists of a synset, a definitional gloss, and (sometimes) one or more sentences illustrating usage. The semantic relations used to organize words and entries are synonymy and antonymy, hyponymy, troponymy and hypernymy, meronymy and holonymy. A currently active project, the disambiguation of definitional glosses, will be discussed.

Charles J. Fillmore: The FrameNet project is morphing from a lexicon-building project to a system capable of providing a layer or two of semantic annotation for full sentences. My remarks will summarize the kinds of information the FrameNet database can provide now, and what it should be able to offer in the not too distant future, for researchers in language engineering and cognitive science. FrameNet is moving toward greater coverage of the lexicon, adaptation to specialist vocabularies, more systematic treatment of multiword expressions, and provisions for incorporating a wide variety of (non-core) grammatical constructions.

Martha S. Palmer: Recently, a consensus has been achieved as to a task-oriented level of semantic representation to be layered on top of the existing Penn Treebank syntactic structures: The Proposition Bank, or PropBank. PropBank consists of argument labels for the semantic roles of individual verbs and similar predicating expressions such as participial modifiers and nominalizations. This talk will describe the PropBank verb semantic role annotation being done at Penn for both English and Chinese. The annotation process will be discussed as well as the use of existing lexical resources such as WordNet, Levin classes and VerbNet. Comparisons with similar projects, including the FrameNet Project at Berkeley and the Prague Tectogrammatics project, will be made.

Doug Lenat: Cognitive Science research could benefit from large-scale representational building blocks. One such building block is a broad ontology of, well, everything. Another one, resting on that, is a formal axiomatization of most of the meaning of most of those concepts; i.e., millions of axioms about those hundreds of thousands of terms. These two pieces have been worked on for twenty years -- and the better part of a person-century of effort -- as Cyc. It's been highly proprietary so far, with a small tip of its ontology exposed as OpenCyc. Starting this summer, though, Cyc is being made available in its entirety for R&D purposes for free, courtesy of Ron Brachman at DARPA's IPTO. In this presentation I will briefly describe this ResearchCyc ontology and KB, and some initial utilities provided with it (inference engine, English-Cyc lexicon, interfaces), how this might fit in with the other infrastructure elements being described in this panel, and how these might leverage your work.

Pat Hayes

Pat Hayes, the discussant for this symposium, has the unique distinction of having inventing two of the most widely-used representations for change, the *situation calculus* (with John McCarthy) and *histories*. He is a Fellow of the American Association for Artificial Intelligence and the Cognitive Science Society.

The inter-relationship between spatial cognition and gestures

J. Gregory Trafton	Mary Hegarty	Barbara Tvesky
trafton@itd.nrl.navy.mil	hegarty@psych.ucsb.edu	bt@psych.stanford.edu
Naval Research Lab	UC Santa Barbara	Stanford University
Chris Schunn	**Justine Cassell**	**Martha Alibali (Discussant)**
schunn@pitt.edu	justine@northwestern.edu	mwalibali@wisc.edu
LRDC, Univ of Pittsburgh	Northwestern University	University of Wisconsin

Introduction

One of the most interesting and frequent questions within cognitive science is "What type of mental representation is used when people solve problems/make decisions/think/etc.?" This symposium will explore the issue of mental representation within the area of spatial cognition by examining how people gesture within the context of a spatial task (e.g., mental animation, scientific analysis, construction, etc.).

Spatial Transformations and Gestures
(Greg Trafton, Susan Trickett, and Cara Stitzlein)

How do experts think about complex data spaces? We examined experts in three domains (meteorology, submarine, fMRI) as they solved a difficult domain problem. We found that, consistent with previous research, they created complex spatial mental representations (e.g., mental models) and used spatial transformations to mentally manipulate these representations. When we examined the gestures that these experts spontaneously made, we found that many of their gestures occurred while performing spatial transformations. We suggest that these gestures were outward manifestations of what they were thinking that facilitated their problem solving.

Gesture use while thinking about Machines
(Mary Hegarty, Sarah Mayer and Sarah Kriz)

We examined the use of gestures while people solved "mental animation" problems in which they have to predict the motion of a mechanical device from static diagrams. In "think aloud" experiments, participants gestured on more than 90% of problems (although they were not instructed to gesture), and their gestures communicated information that was not stated in words.

In another experiment, participants were asked to think aloud while solving problems (communication group), just solved the problems (control group), or solved the problems while tapping a spatial pattern (dual task group). Although the communication group gestured the most, gestures were also frequent in the non-communicative situation experienced by the control group.

The dual-task group had poorer performance than the control group, suggesting that prevention of gesturing impaired performance. These preliminary results suggest that gestures function both in the process of inferring motion from static diagrams and in communicating the results of this inference process.

Gestural Models for Self and Others (B. Tversky, H. Taylor, K. Emmorey, J. Heiser, & S. Lozano)

Three paradigms show that gestures can convey spatial information effectively both for self and for others. People's oral spatial descriptions include gestures that reflect the perspective taken on the space; they also provide a model of the space for the listener. The information in the words is not sufficient without the gestures. Listeners, too, gesture while hearing spatial descriptions from unobservable speakers; these gestures appear to help establish a spatial mental model for the listener. People's explanations of how to put something together include gestures that convey the structural relations of the object and actions needed for assembly, information that facilitates performance of communicator and receiver.

The role of gestures in a theory of spatial representation
(Chris Schunn, Lelyn Saner and Tony Harrison)

In many complex problems, there is a significant visual or spatial component to the problem solver's representations. Neuropsychological and cognitive work has suggested that there are several fundamentally different ways in which problem solvers can represent visual/spatial information---flat vs. 3-dimensional, near vs. far, approximate vs. detailed. Each of these differences can have a strong influence on problem-solving behavior. We have a theory, ACT-R/S (Harrison & Schunn, 2001), for how these features are strongly correlated in one of three possible visual/spatial representations, how the representations are selected, and how the representations influence behavior. We have been using analyses of gestures to diagnosis their representation choice. Our talk will illustrate how our theory of spatial representation influences our use of expert and novice problem solvers's gesture data to infer representations beyond previous analyses and how our theory of spatial representations has been changed by the gesture data.

The Consequences of Spatial Gestures (Justine Cassell)

Evidence from house descriptions, route descriptions, and the description of complex objects jives with earlier studies showing that roughly 50% of gestures convey content that is not redundant with the content of speech. These complementary gestures have consequences for later speech in that they may be referred back to – both via gesture and via speech – by both the speaker and the listener. I address a number of questions about the role of gesture in the semantics of ongoing talk, and their role in spatial cognition: how do we predict what aspects of spatial scenes will be described in gesture vs. speech, the form that these spatial descriptions will take in gesture, and how these features are represented by the participants in a discourse in such a way as to serve as the context for later speech.

Publication-Based Talks

Situating Abstract Concepts

Lawrence W. Barsalou (barsalou@emory.edu)
Department of Psychology
Emory University, Atlanta, GA 30322 USA

Katja Wiemer-Hastings (katja@niu.edu)
Department of Psychology
Northern Illinois University, DeKalb IL 60115 USA

Roughly speaking, abstract concepts such as *TRUTH* refer to entities that are neither purely physical nor spatially constrained (Wiemer-Hastings, Krug, & Xu, 2001). Such concepts pose a classic problem for theories that ground knowledge in modality-specific systems (e.g., Barsalou, 1999, 2003a,b). Abstract concepts also pose a significant problem for traditional theories that represent knowledge with amodal symbols. Surprisingly, few researchers have attempted to specify the content of abstract concepts using feature lists, semantic networks, or frames. It is not enough to say that an amodal node or a pattern of amodal units represents an abstract concept. It is first necessary to specify the concept's content, before beginning the task of identifying how this content is represented.

Hypotheses

A common assumption is that abstract and concrete concepts have little conceptual content in common, if any. Alternatively, we propose that concrete and abstract concepts share important similarities. In particular, we propose that they share common situational content, namely, information about agents, objects, settings, events, and mental states (Hypothesis 1). Where concrete and abstract concepts differ is in their specific foci within background situations. Whereas concrete concepts focus on objects and settings, abstract concepts focus on events and mental states (Hypothesis 2). As a result of these different foci, the representations of abstract concepts are more complex, being less localized in situational content (Hypothesis 3). Finally, because the content of abstract concepts is grounded in situations, modality-specific simulations could, in principle, represent it (Hypothesis 4).

Method

In an exploratory study, we assessed the content of three abstract concepts: *TRUTH, FREEDOM,* and *INVENTION*. These concepts were compared to three concrete concepts—*BIRD, CAR,* and *SOFA*—and also to three intermediate concepts—*COOKING, FARMING,* and *CARPETING*. We first asked participants to produce properties typically true of these concepts. We then analyzed these properties using two coding schemes, one that coded small protocol units, and a second that coded large ones.

Results

For the complete results, see Barsalou and Wiemer-Hastings (in press). Both coding analyses offered support for Hypothesis 1, namely, common situational content was produced across concrete, intermediate, and abstract concepts. For all concepts, participants tended to describe background situations, including information about entities, settings, events, and mental states. Indeed the similarities between concepts were more striking than the differernces.

Both analyses also offered support for Hypothesis 2, namely, concrete and abstract concepts differed in their situational foci. Whereas concrete concepts focused more on objects, locations, and behaviors, abstract concepts focused more on social aspects of situations (e.g., people, communication, social institutions) and mental states (e.g., beliefs. complex relations). Intermediate concepts lay in between.

Consistent with Hypothesis 3, conceptual structures were most complex for abstract concepts. Abstract concepts were most likely to contain deep hierarchies of large conceptual clusters organized by complex relations.

Regarding Hypothesis 4, we see no reason that the content of abstract concepts cannot be represented in simulations. Because their content is perceived in the situations that involve abstract concepts, it could, in principle, be reenacted later when representing them. Clearly, much further research beyond this exploratory study is necessary.

Acknowledgement

This work was supported by National Science Foundation Grants SBR-9905024 and BCS-0212134 to Lawrence W. Barsalou.

References

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577-609.

Barsalou, L.W. (2003a). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 358,* 1177-1187.

Barsalou, L.W. (2003b). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18,* 513-562).

Barsalou, L.W., & Wiemer-Hastings, K. (in press). Situating abstract concepts. In D. Pecher and R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought.* New York: Cambridge University Press.

Wiemer-Hastings, K., Krug, J., & Xu, X. (2001). Imagery, context availability, contextual constraint, and abstractness. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 1134-1139. Mahwah, NJ: Erlbaum.

Notes on the Negative Side of Rationality: Critical Principles

Mark H. Bickhard (mark@bickhard.name http://www.bickhard.ws/)
Lehigh University, 17 Memorial Drive East, Bethlehem, PA 18015

A central but neglected aspect of rationality is its negative aspect: knowledge of error. Knowledge of error warrants and motivates criticism, so it constitutes knowledge of principles of criticism, or *critical principles* (Bickhard, 1991, 2001a, 2002; Bickhard & Campbell, 1996a). I will outline a model of rationality, grounded essentially in an interactive, agent based model of representation and cognition (Bickhard 1993, 1996, 1998a, 1998b, 2001b; Bickhard & Campbell, 1996b; Bickhard & Terveen, 1995), that gives central place to such negative knowledge, and show how it solves and dissolves multiple problems. It puts creative interaction and problem solving at the center of the nature of mind, rather than portraying reason trying to rule the passions. It is a non-foundationalist model, thus avoiding the problem of the rational warrant for the foundations of rationality. Nevertheless it is self-consistent in the sense that being rational is itself rational, but without having to demonstrate that rationality leads closer to truth. It accounts for logic as a natural development of rationality given reasonable additional assumptions, such as that of language and of the possibility of conscious reflection, and, thus, renders logic as a rational creation rather than the essence of rationality.

In this talk, I will focus on logic as rational construction rather than as the center of rationality. Historically, logics have been developed of greater and greater power, but no logic can construct a logic more powerful than itself. If logic were the essence of rationality, therefore, the history of logic would necessarily be arational. In this model, logic emerges as an inherently wide, natural domain of possible error and of means of avoiding those errors, and thus avoids that problem, as well as the many other problems of foundationalism. This model also naturally situates higher order logics and modal logics within the broader framework. I will not be addressing the technical details of these points, but will outline the basic perspective.

If there is time, I will also outline another realm of advantages of this model of rationality: It dissolves several problems in the philosophy of science, such as the perplexities involved in notions of progress, of realism and truth, and of induction.

References

Bickhard, M. H. (1991). A Pre-Logical Model of Rationality. In Les Steffe (Ed.) *Epistemological Foundations of Mathematical Experience.* New York: Springer-Verlag.

Bickhard, M. H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 285-333.

Bickhard, M. H. (1996). Troubles with Computationalism. In W. O'Donohue, R. F. Kitchener (Eds.) *The Philosophy of Psychology*. London: Sage.

Bickhard, M. H. (1998a). Levels of Representationality. *Journal of Experimental and Theoretical Artificial Intelligence, 10*(2), 179-215.

Bickhard, M. H. (1998b). Whither Representation? In M. A. Gernsbacher, S. J. Derry (Eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, University of Wisconsin-Madison, August 1 - 4, 1998, 150-155. Erlbaum.

Bickhard, M. H. (2001a). Error Dynamics: The Dynamic Emergence of Error Avoidance and Error Vicariants. *Journal of Experimental and Theoretical Artificial Intelligence, 13*, 199-209.

Bickhard, M. H. (2001b). Why Children Don't Have to Solve the Frame Problems: Cognitive Representations are not Encodings. *Developmental Review, 21*, 224-262.

Bickhard, M. H. (2002). Critical Principles: On the Negative Side of Rationality. *New Ideas in Psychology, 20*, 1-34.

Bickhard, M. H., Campbell, R. L. (1996a). Developmental Aspects of Expertise: Rationality and Generalization. *Journal of Experimental and Theoretical Artificial Intelligence, 8*(3/4), 399-417.

Bickhard, M. H., Campbell, R. L. (1996b). Topologies of Learning and Development. *New Ideas in Psychology, 14*(2), 111-156.

Bickhard, M. H., Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution.* Elsevier Scientific.

Role of pattern recognition and search in expert decision making

Fernand Gobet (fernand.gobet@brunel.ac.uk)
Department of Human Sciences, Brunel University
Uxbridge, Middlesex, UB8 3PH UK

The study of expert behaviour has attracted widespread attention since the seminal work of de Groot (1965) and Chase and Simon (1973). Of particular interest is how experts, even under time pressure, can make relatively good decisions in spite of strong limits in their computational capacities. A substantial amount of research has focused on chess playing, as this domain offers a well-validated and ecological measure of expertise (the Elo rating). Obviously, this question has important repercussions beyond game playing, and extensive research has been carried out about decision making in domains such as fire fighting, medical diagnosis, and aviation (e.g., Zsambok & Klein, 1997).

While the importance of both recognition and search mechanisms is generally accepted, researchers disagree as to their relative importance. De Groot (1965) showed that even chess grandmasters seldom look at more than 100 possible continuations of the game before choosing a move. This number is vastly smaller than the number of legal moves (on average, for a middlegame position, the number of legal continuations six ply deep is about 1.8 billion, and increases exponentially for greater depths). De Groot (1965) also found that top-level grandmasters do not search reliably deeper than candidate masters, although more recent data suggest that masters search slightly deeper, on average, than weak amateurs (e.g., Gobet, 1998). De Groot (1965) as well as Chase and Simon (1973) propose that recognition, by allowing knowledge to be accessed rapidly, enables look-ahead search to be highly selective. Holding (1985), by contrast, argued that the main determinant of chess skill is the ability to plan ahead by search, rather than reliance on recognition of positional patterns.

Support for the role of pattern recognition in expert behaviour comes from two main lines of research: (a) perception and memory, and (b) decision making. Evidence from perception and memory indicates that experts can rapidly recognize the key features of a problem, and that there are important differences between experts' and non-experts' eye-movements (de Groot & Gobet, 1996; Gobet, de Voogt & Retschitzki, in press). Research has also shown that experts have a remarkable memory for domain-specific material (Chase & Simon, 1973; de Groot, 1965; de Groot & Gobet, 1996). Interestingly, their superiority extends to the recall of random positions, although the skill difference is then much smaller than with game positions. CHREST, a detailed computer model of pattern recognition, has accounted for these results (de Groot & Gobet, 1996; Gobet & Simon, 2000; Gobet & Waters, 2003).

The second line of evidence comes from rapid decision making (e.g., Zsambok & Klein, 1997). In particular, research with chess players suggests that grandmasters can play at a high level even under severe time pressure (e.g., Gobet & Simon, 1996). SEARCH, a computational model based on CHREST, accounts for several data from expert problem solving, such as how average depth of search increases as a function of skill (Gobet, 1997).

Recently, proponents of the predominant role of search processes have collected data aiming at undermining the importance of pattern recognition. In particular, Chabris and Hearst (2003), using data from rapid chess and blindfold chess, have questioned Chase and Simon's (1973) and Gobet and Simon's (1996) account. In this talk, I'll show that Chabris and Hearst's (2003) data, far from invalidating theories based on pattern recognition and selective search, actually support them.

References

Chabris, C. F., & Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: Effects of playing speed and sight of the position on grandmaster chess errors. *Cognitive Science*, *27,* 637–648.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.

de Groot, A. D. (1965). *Thought and choice in chess* (1st ed.). The Hague: Mouton Publishers.

de Groot, A. D. & Gobet, F. (1996). *Perception and memory in chess. Studies in the heuristics of the professional eye.* Assen: Van Gorcum.

Gobet, F. (1997). Roles of pattern recognition and search in expert problem solving. *Thinking and Reasoning, 3,* 291-313.

Gobet, F. (1998). Chess players' thinking revisited. *Swiss Journal of Psychology, 57,* 18-32.

Gobet, F., de Voogt, A., & Retschitzki, J. (in press). *Moves in mind.* Hove, UK: Psychology Press.

Gobet, F. & Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grandmaster level chess. *Psychological Science, 7,* 52-55.

Gobet, F. & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24,* 651-682.

Gobet, F., & Waters, A. J. (2003). The role of constraints in expert memory. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 29, 1082–1094.

Holding, D. H. (1985). *The psychology of chess skill*. Hillsdale, NJ: Erlbaum.

Zsambok, C. E., & Klein, G. A. (1997). *Naturalistic decision making*. Mahwah, N.J.: Erlbaum.

Group Path Formation

Robert L. Goldstone (rgoldsto@indiana.edu)
Andy Jones
Michael E. Roberts
Cognitive Science Program, Indiana University
Bloomington, IN. 47405

Communal Path Construction

When people make choices within a group, they are frequently influenced by the choices made by others. One reason for this is that the actions of other people changes the environment in which a person makes their choices. In many domains, initial pioneers reduce the costs for followers who subsequently pursue the same path.

Our concrete instantiation of this situation is group path formation where people travel between destinations with the travel cost for moving onto a location inversely related to the frequency with which the location has been visited by others. In this situation, people may detour from straight paths connecting destinations to take advantage of frequently visited, hence inexpensive, paths. At a group level, the mathematics of "Minimal Steiner Trees" describes optimal path systems for connecting a set of destinations. A Minimal Steiner Tree (MST) is the set of paths that fully connects a set of destinations using the minimal amount of total path length. Finding minimal MSTs is a notorious NP-complete problem, with all known, provably optimal algorithms requiring an exponential increase in computation as the number of destination points increases linearly (Garey, Graham, & Johnson, 1977). However, analog devices such as soap films over wire frames have been shown to spontaneously create MSTs. Do groups of people create path systems that approximate MSTs as well?

Path Formation Experiment

59 Indiana University undergraduates were divided into 8 groups. The participants in each group were given the task of traveling between randomly sampled cities from a set of 3-4 cities. Participants were told to try to maximize their total number of points. Points were earned by successfully reaching destination cities, and were deducted for travel costs associated with each visited square on the map. The travel cost for a square was inversely related to the number of times all individuals previously visited the square, with recent visits reducing travel costs more than older visits.

As participants moved between cities, they saw their own locations as red triangles, other participants' locations as yellow circles, and the moment-by-moment cost of each map square color-coded with brightness representing ease of travel. Cities were arranged in triangular or rectangular configurations.

Results and Modeling

There were systematic deviations from beeline pathways in the direction of MSTs for all of the configurations of cities, however the participants' paths never converged upon MSTs. Greater deviations of beeline paths (see the solid lines in Fig. 1) toward MSTs (dashed lines) were found for the isosceles than the equilateral triangle arrangement of cities. In Fig. 1, the more often a square is visited, the brighter it appears. Furthermore, greater pro-MST deviations were found for a small rectangle than a large rectangle possessing the same proportions. Finally, asymmetric pro-MST deviations were observed, with greater deviations for participants traveling from City A to City B than traveling from City B to City A.

All three of these deviations from beeline pathways can be explained by Helbing, Keltsch, and Molnár's (1997) "Active Walker" model of pedestrian motion. This agent-based model includes equations for environmental changes produced by walking on paths that make the paths more accessible for subsequent walkers. At every time step, each walker in a group moves in a direction that compromises between moving toward the destination and moving toward heavily trafficked locations. This model and our experimental groups both establish path systems that lie between beeline and MSTs, with the deviation from beeline paths influenced by the topology of the destinations, the duration of travel, and the absolute scale of the world.

Acknowledgments

This research was funded by NIH grant MH56871 and NSF grant 0125287.

References

Garey, M. R., Graham, R. L, & Johnson, D. S. (1977). The complexity of computing Steiner Minimal Trees. *SIAM Journal on Applied Mathematics*, *32*, 835-859.

Helbing, D., Keltsch, J., & Molnár, P. (1997). Modeling the evolution of human trail systems. *Nature*, *388,* 47-50.

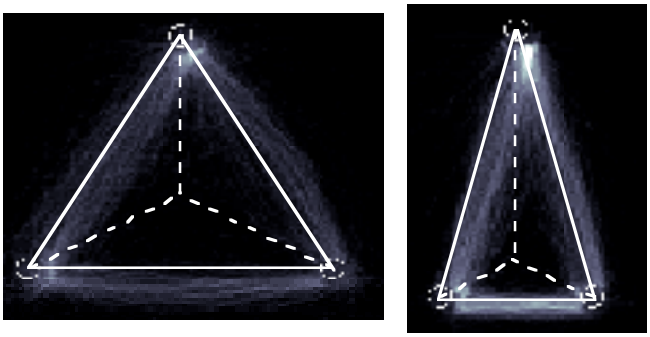


Figure 1: Group pathways for equilateral and isosceles triangles

Why Is Word Learning Related to List Memory? Empirical and Neuropsychological Tests of a Computational Account

Prahlad Gupta (prahlad-gupta@uiowa.edu)
Department of Psychology; University of Iowa
Iowa City, IA 52242 USA

Introduction

A growing body of evidence indicates that human word learning, nonword repetition, and immediate serial recall (ISR) abilities are related in some way (e.g., Baddeley et al., 1998). It seems clear why word learning might be related to nonword repetition: every known word was once a nonword to a particular learner, so greater facility in processing nonwords should lead to greater facility in eventually learning them. But why might nonword repetition and list recall be related?

Nonwords as Lists

One possibility is that a nonword is processed like a list when first encountered, and is thus directly dependent on list sequencing mechanisms. If this were the case, we would expect to observe serial position effects in repetition of the sequence of sounds comprising nonwords, just as in ISR of lists of words. We conducted three experiments to examine syllable serial position effects in repetition of individual polysyllablic nonwords, and obtained significant primacy and recency in all three experiments (Gupta, 2004).

Testing Alternate Models

There are two possible explanations of these results in terms of our previous computational work (Gupta, 1996). The two formulations can be distinguished by differing predictions with regard to correlations between ISR, nonword repetition, and word learning. The original formulation explains the fact that these correlations arise developmentally but predicts their absence in adults. An alternative formulation predicts that such correlations obtain not only developmentally, but also in adults. We conducted two experiments to examine whether ISR, nonword repetition, and word learning are correlated in adults (Gupta, 2003). The results indicated that the developmental relationships between all three abilities also exist in adults, thus supporting the revised model over the original model.

Neuropsychological Investigation

The revised version of our model incorporates the view that there is a functional relationship between the abilities, all of which invoke the same sequencing mechanisms. If this is really the case, we would expect relationships between these abilities to obtain even following early neurological injury across a variety of lesion sites. This is because in the case of early lesions, there is a real possibility for remission of deficits as a result of neural reorganization; persistence of correlations would thus support the hypothesis of an underlying functional relationship. We examined this question by administering tests of word learning, nonword repetition, and ISR to 5-10 year old children who had suffered perinatal brain injury across a variety of sites, and to age-matched controls (Gupta et al., 2003). The results indicate that the relationships between ISR, nonword repetition, and word learning are exhibited even under conditions of early brain injury. These findings thus provide further support for the functional architecture of our revised model.

These various lines of evidence together clarify how common sequencing mechanisms may underlie both nonword processing and immediate list recall, thereby offering an explanation of why word learning is related to list memory.

References

Baddeley, A. D., Gathercole, S. E., and Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105:158–173.

Gupta, P. (1996). Word learning and verbal short-term memory: A computational account. In Cottrell, G. W., editor, *Proceedings of the Eighteenth Annual Meeting of the Cognitive Science Society*, pages 189–194. Lawrence Erlbaum, Mahwah, NJ.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *Quarterly Journal of Experimental Psychology (A)*, 56:1213–1236.

Gupta, P. (2004). Primacy and recency in nonword repetition. *Memory*. In press.

Gupta, P., MacWhinney, B., Feldman, H., and Sacco, K. (2003). Phonological memory and vocabulary learning in children with focal lesions. *Brain and Language*, 87:241–252.

Does Gesture Play a Special Role in the Brain's Processing of Language?

Spencer D. Kelly (skelly@mail.colgate.edu)
Department of Psychology—Neuroscience Program, Colgate University
13 Oak Dr., Hamilton, NY 13346 USA

Corinne Kravitz (ckravitz@mail.colgate.edu)
Department of Psychology—Neuroscience Program, Colgate University
13 Oak Dr., Hamilton, NY 13346 USA

Background

People of all ages, cultures, and backgrounds gesture when they speak. What function do these hand movements serve? Although there is consensus that gesture plays an important role in language production, there is considerable theoretical debate as to what role gesture plays in language comprehension (Clark, 1996; Kelly, 2001; Kelly, Barr, Church, & Lynch, 1999; Krauss, 1998; Krauss, Morrel-Samuels, & Colasante, 1991; McNeill, 1992). At the core of this debate is the fact that previous research has relied on indirect behavioral measures that do not provide access to the underlying neurocognitive processing of speech and gesture. The present research addresses this issue by using a more direct neurocognitive measure: event-related potentials (ERPs). ERPs measure electrical brain activity and have been used successfully in previous research on the neural processing of language.

Methods

In Study 1, adult participants watched videos of speech and gesture, in which the gesture conveyed the same, complementary or different information as the accompanying speech. ERPs were recorded to the speech in these different gesture contexts. In Study 2, participants watched videos similar to Study 1, but the gesture was replaced with digitally-inserted visual information (a vertical or horizontal line representing different dimensions of the objects) that either conveyed the same, complementary or different information as the speech. The goal of Study 2 was to determine whether the results from Study 1 were due to gesture *per se*, or simply any visual information that preceded speech.

Results

The results from Experiment 1 have been published (Kelly, Kravitz & Hopkins, in press) and reveal that gestures not only influence ERPs to speech, but also that the gesture influence is late (N400) and early (sensory, P1-N1 and P2 components) in the brain's processing of speech. This suggests that gestures may affect not only high-level semantic processing of speech, but also low-level phonological processing as well.

Study 2 is currently in progress to determine whether the results from Study 1 are unique to gesture, or whether any meaningful visual information will have a similar impact on the brain's processing of speech.

Discussion

Researchers know very little about the neural time course of how gestures influence speech comprehension. This question bears on important theoretical issues in language research: are there aspects of language processing that are impervious to contextual influence? When does a gestural context influence the neural processing of language? In addition, the results will address a debate in the literature about the "specialness" of the relationship between speech and gesture in language processing (McNeill, 1992). Finally, by taking ERPs to speech using real-time, multimodal videos, this project makes an important methodological contribution to the neuroscientific investigation of language processing.

Acknowledgments

The authors thank Colgate University for their generous support of undergraduate research.

References

Clark, H. H. (1996). *Using Language*. Cambridge, GB: Cambridge University Press.

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language, 28,* 325-349.

Kelly, S. D., Barr, D., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40,* 577-592.

Kelly, S. D., Kravitz, C., & Hopkins, M. (in press). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language.*

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7(2),* 54-60.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality & Social Psychology, 61(5),* 743-754.

McNeill, D. (1992). *Hand and mind: What gesture reveals about thought*. Chicago, IL: University of Chicago Press.

Phonology without Phonemes

James L. McClelland (jlm@cnbc.cmu.edu)
Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon
115 Mellon Institute, 4400 Fifth Avenue, Pittsburgh, PA 15213 USA

Overview

Theories of language structure generally provide two things: a list of allowable units of various types, and rules or constraints that determine how the various units may be combined to create larger units. Together the units and rules determine which larger units are part of the language and which are not. In general, acceptability is an all-or-nothing proposition: A given proposed larger unit either is or is not acceptable according to this approach.

In this talk I will argue instead that at there are neither rules nor units and that the acceptability of particular possible utterances is a matter of graded constraint satisfaction. The particular utterances I will consider are the word-forms of English. The argument will draw heavily on the prior work of Joan Bybee (2001) as well as on work with Karalyn Patterson (McClelland and Patterson, 2002a,b) and Gary Lupyan (Lupyan and McClelland, 2003) and continuing work with Bybee, Lupyan, Catherine Harris, and Brent Vanderwyk. Two aspects of this effort that will be discussed in this talk are described below.

Constraints on English Word Forms

The idea that candidate word forms have graded goodness values has been introduced by Kessler and Treiman (1997). Work currently in progress with Brent Vanderwyk extends this idea, accounting for the relative probabilities of the different word bodies that occur in stressed monomorphemic monosyllabic words. More traditional work by Harris (1994) describes a set of rules that dictate which phoneme sequences are legal in English and which are not. However, the analysis seems incomplete in that there are many sequences that are legal but which occur relatively infrequently (e.g., sequences of the form *_vpt* where *v* is a short vowel and the '_' indicates the missing word onset), while there are other sequences that occur very much more frequently (e.g. sequences such as *_vnd*, *_vnt*, *_vld*, *_vlt*, and *_vst*). Harris himself allows that some forms are less preferred than others but offers no systematic way to address this.

According to our analysis, word bodies may involve a (partial or complete) closure with the lips (as in *cuff* or *cup*), with the tip of the tongue (as in *bass* or *cat*), or with the back of the tongue (as in *tack*), or no closure at all (as in *bee*). Each closure adds a complexity cost (with labial and back closures adding more than tip closures), but embellishments of a closure already paid for are relatively cheep. Thus, adding nasality to a tongue tip closure (as in *hint* as compared to *hit*) costs relatively little while the combination of a labial closure and a tip closure is more expensive, thus explaining the relative prevalence of *_vnt* compared to *_vpt*. At the time of this writing we have been able to use these ideas to account for nearly 90% of the variance in the frequencies of occurrences of different English word bodies.

Language Change and Morphology

Bybee (2001) has argued that language change transcribed as a shift from one phoneme to another reflects an underlying continuity of gestural change. Gary Lupyan and I have been exploring models of language change that capture the ways in which a graded constraint on the length of a word form gradually results in the creation of quasi-regular past tense forms like did, said, had, and made. Such forms are treated as exceptions listed in the lexicon by rule-based approaches such as Pinker (1999); in our model, they arise from fully regular forms from a graded constraint that tends to result in the gradual shortening of very frequent forms. What is lost in shortening is not completely arbitrary; some sign of the regular past tense inflection is preserved in nearly all cases, while the vowel is shortened or a consonant is deleted from the regular form. Recent work extending our approach to account for a variety of aspects of the evolution of the English past tense system, building on Lupyan and McClelland (2003), will be presented.

References

Bybee, J. (2001). *Phonology and language use.* Cambridge: Cambridge University Press.

Harris, J. *English sound structure*. Oxford: Blackwell.

Kessler, B. and Treiman, R. (1997). Syllable Structure and the Distribution of Phonemes in English *Syllables. Journal of Memory and Language, 37*, 295-311.

Lupyan, G. and McClelland, J. L. (2003). Did, Made, Had, Said: Capturing the Quasi-Regularity in Exceptions. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

McClelland, J. L., & Patterson, K. (2002a). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*, 465-474.

McClelland, J. L., & Patterson, K. (2002b). 'Words or Rules' cannot exploit the regularity in exceptions. Reply to S. Pinker & M. T. Ullman (2002b). *Trends in Cognitive Sciences, 6*, 464-465.

Pinker, S. (1999). Words and Rules. New York: Basic Books.

A Multi-Modal Study of Cognitive Processing under Negative Emotional Arousal

Lilianne R Mujica-Parodi (lilianne.strey@stonybrook.edu)
Laboratory for the Study of Emotion and Cognition, Departments of Biomedical Engineering and Psychiatry, HSC T10; State University of New York at Stony Brook, School of Medicine; Stony Brook, NY 11794-8101
Tsafrir Greenberg (tsafrirg@aol.com)
Laboratory for the Study of Emotion and Cognition, Departments of Biomedical Engineering and Psychiatry, HSC T10; State University of New York at Stony Brook, School of Medicine; Stony Brook, NY 11794-8101
John F Kilpatrick (jfk2001@columbia.edu)
Laboratory for the Study of Emotion and Cognition, Departments of Biomedical Engineering and Psychiatry, HSC T10; State University of New York at Stony Brook, School of Medicine; Stony Brook, NY 11794-8101

Abstract

It is a truism of everyday life that anger and fear affect cognition. In high-risk perceptually complex contexts, such as air combat, the effects of negative arousal on performance can be significant and potentially catastrophic. To better understand the interaction between emotion and cognition, we studied the effects of negative emotional stimuli on pre-attentive sensorimotor gating and selective attention in 39 healthy adults, as well as their relationship to neural, cardiac, and endocrine variables associated with the arousal response. Subjects were tested for pre-pulse inhibition under neutral and arousal conditions, as well as on emotionally-valent Flanker and Stroop tasks. Physiological arousal reactivity was measured using functional MRI, 24-hour EKG, electrodermal activity, cortisol testing, and dexamethasone suppression. Subjects were clinically assessed for levels of anger, anxiety, and perceived stress. Affect-valent conditions were induced using the International Affective Picture Scale, the Morphed Eckman Facial Stimuli, and affect-valent words matched for length and frequency. All conditions were counter-balanced for order. Our results indicate that even under relatively mild emotional challenge, the introduction of negative emotion significantly affected nearly all components of our cognitive battery, and correlated with changes in heart rate and electrodermal activity. Pre-attentive sensory gating and habituation were diminished, which may reflect the underlying neural conditions necessary for an increased orienting response. On tasks that required selecting a target in the presence of distractors, such as the Flanker Task, arousal had the effect of reducing both response time and accuracy. Our results were also consistent with our previous research on the higher-order effects of arousal on reasoning, indicating that individuals make decisions with less information under emotional arousal. On tasks such as the Stroop, in which orienting to the source of arousal conflicts with selective attention to a target, response time was lengthened. Importantly, the effects of negative arousal were widely variable across individuals, falling roughly into classes of individuals who showed strong physiological arousal response with strong cognitive effect, individuals who showed little physiological arousal response with little cognitive effect, and individuals who showed strong physiological arousal response with little cognitive effect. It is the third group that we are investigating most closely with fMRI, to determine which limbic feed-back mechanisms produce the most efficient cognitive performance under stress. This information, in turn, will permit more effective screening for high-risk environments to select only those individuals that are "hard-wired" for neural aptitude during fear.

Background

Emotional arousal primes the organism for imminent danger by increasing the orienting response, which permits the organism to find and focus on the source of danger. Once oriented to the source of danger, emotional arousal strengthens attention to the source of danger and diminishes attention to stimuli unrelated to its source, narrowing the amount of peripheral information simultaneously accessible with the target. This two-pronged approach has both costs and benefits: cognition is limited with respect to breadth, with the individual attending to less information at a time, but is more flexible in terms of the ability to switch attention from one target to another. Under most dangerous conditions in our evolutionary past, these costs and benefits were appropriate for survival: in the presence of a predator, it makes sense to focus on the predator, to ignore peripheral information such as ambient noise, and to be able to quickly switch attention between two or more predators that together present a collective threat.

While the cognitive changes associated with arousal in humans are appropriate for predator/prey contexts, most states of arousal (fear, stress, anxiety) in modern societies today occur under far different circumstances, in which the source of arousal is often not a concrete palpable entity to which one can readily orient. Even individuals in actually dangerous situations, such as fighter pilots in combat, protect themselves by defying their instincts: a fighter pilot needs to attend not only to the "predator" shooting at him, but equally to the myriad of dials and instruments that keep his plane aloft and his artillery engaged. Thus, while emotional arousal can benefit cognitive performance by increasing focused attention on a target and decreasing attention to distracting irrelevant information, emotional arousal today can just as often wreck havoc on cognitive performance by triggering the orienting response in the absence of an appropriate target and by disregarding potentially relevant peripheral stimuli ("tunnel vision").

Easterbrook, in 1959, seems to have been the first to fully articulate the hypothesis that arousal produces attentional narrowing, while Bacon (1974) was instrumental in relating attentional narrowing to the orienting response. Their hypotheses have since been

supported by a wide range of studies on humans and animals that used induced arousal by reward (Bruner et. al, 1955), electric shock (Cornsweet, 1969), loud noise (Hockey, 1970a), threatening words (Combs & Taylor, 1952), test anxiety (Rockett, 1956), and pre-parachuting anxiety (Hammerton & Tickner, 1967) on various tests of information processing. Research on selective attention in actual dangerous environments, in which simulation acted as the control for arousal, demonstrate that the tendency to overlook incidental (peripheral) cues in real-life situations can have severe implications for actual performance. Significant decline in performance has been shown for complex tasks that were performed during combat (Walker & Burkhardt, 1965), during deep-sea diving (Baddeley, 1972), as well as during realistically-simulated experiments in which subjects thought they were in mortal danger and were required to perform selective attention tasks (Berkun et. al, 1962) (Weltman & Egstrom,1971).

While useful and informative, these early experiments had several limitations, the most prominent of which was that they investigated mean performance effect without considering the effects of individual variability. Yet the factors that predict vulnerability or resilience to the cognitive effects of arousal have tremendous practical importance, particularly in screening for occupations that require complex cognitive processing under dangerous conditions. Other limitations were the failure to discriminate between selective attention and orienting responses, two processes that are presumed to be linked but nonetheless distinct, as well as the failure to distinguish between the effects of arousal on pre-attentive sensory gating versus the effects of arousal on attentive selective attention, two processes that intuitively might be linked but whose relationship has not been extensively studied.

The purpose of our study was therefore threefold. Our first aim was to establish or replicate findings on mean cognitive changes that occur in the general population in the context of mild emotional arousal; specifically pre-attentive sensory gating (emotional pre-pulse inhibition), selective attention (emotional flanker task), and orienting (emotional Stroop task). Our secondary aims were to compare the role that emotional arousal plays in selective attention and the orienting response, and to evaluate the interaction of sensorimotor gating and selective attention. Our third aim was investigate the effects of individual variability, specifically relating to neural, endocrine, and subjective assessment of baseline stress, anxiety, and anger, on task performance.

Methods

Subjects: We tested 39 adults (18 male, 21 female) between the ages of 18 and 50 (mean = 30.92; SD = 9.103). All subjects were screened and shown to be free from neurological and DSMIV Axis I &2 psychiatric illness using the Schedule for Affective Disorders – Lifetime Version (Endicott & Spitzer, 1978).

Tasks: We used pre-pulse inhibition (PPI) as a pre-attentive measure of sensorimotor gating. PPI, which has been well validated under non-emotional conditions, measures the inhibition of the startle response when an ordinarily startling stimulus is immediately preceded by a "pre-pulse," and is thought to reflect pre-attentive thalamic "gating" of non-novel stimuli. We used standard acoustic methods, as per Blumenthal (1993). During the task, subjects viewed stimuli presented on a 21-inch computer screen in a completely dark room. The visual protocol consisted of a two-minute orienting cross on a black background, followed by 48 pictures from the International Affective Picture Scales (IAPS) (Lang, Bradley, & Cuthbert, 1995), counter-balanced for order. These IAPS pictures were Neutral during one session and Negative Arousal during the other session. Sessions were also counter-balanced for order, with a duration of 4 minutes, 58 seconds each. The sessions were separated by a 15 minute unrelated task, to avoid habituation or "bleeding" between the conditions. Inter-trial intervals were calculated to prevent a trial commencing less than 2s before or after a picture change, to avoid the picture change itself acting as a pre-pulse. Data was produced separately for Neutral and Negative Arousal conditions, allowing for comparison between the two conditions.

We used a modified Flanker (Fan et al., 2001) to investigate affect-valent selective attention, as well as an emotional Stroop task (D'Alfonso et al., 1999) to measure affect-valent orienting response. During all three tasks, the subject was monitored for EKG and respiration. Physiological data were collected and recorded via the Biopac Systems MP150 module. Subjects performed the tasks under two conditions, Neutral and Negative Arousal, at the same time on two consecutive days, counter-balanced for order. All three tasks used emotionally-valent words (Times New Roman 66 pt. Font) to induce the two conditions. Words for the two conditions were matched for frequency and word length, and came from lists used in previous studies (Dalgleish, 1995, John, 1988, McKenna, & Sharma, 1995). Subjects performed the tasks on a computer, sitting 24 inches from the screen. The tasks were scored for both accuracy and average response time, calculating total score as well as scores for the first and last thirds to measure habituation effects.

The Flanker Task was adapted from the Attentional Network Task (Fan et al., 2001). For each affect conditions, the subject was presented with a series of 48 stimulus pairs. The first screen of the pair was an emotionally-valent word, presented for 1s. The second screen of the pair presented a series of 5 white arrows on a black background. The subject was instructed to identify the direction of the middle arrow by pressing a right or left button on a keypad. The subject's response immediately advanced the task to the next stimulus pair. There were 12 variations for arrow appearance, relating to position on screen, congruence, and direction of arrows, which were programmed to present randomly.

For the Stroop Task, during each affect condition the subject was presented with 60 words that were printed in one of four different colors: Red, Green, Yellow and Blue (15 for each group). The words were presented pseudo-randomly such that no color was repeated twice in a row. The subject was instructed to press the key corresponding to the color of the word shown. A practice run made up of symbols instead of words was presented before the task in order to get the subject comfortable using the keypad without looking down at the keys.

Neuroimaging data was acquired with a 3T Siemens system at the Nathan S. Kline Institute for Psychiatric Research in Orangeburg, New York. During scanning, subjects viewed a series of facial stimuli with negative (angry and fearful) and neutral expressions. Passive viewing of an orienting cross was used as a control condition. The subject's head was secured in a custom-made head-holder and headphones were provided for magnet noise attenuation and for experimenter/subject communication. 198 T2*-weighted coronal echoplanar images (EPI) were acquired covering the frontal and temporal lobes, with TR=3000ms, TE=40ms, Flip angle = 90º, Matrix=64x64, and a FOV=224mm. Our voxel size was 3.5 mm^3 and 31 contiguous coronal slices were obtained.

Following the EPI, we collected 31 T2*-weighted gradient-echo (GE) images which were used in the data-analysis process to correct for distortion found in the EPI images. The parameters for the gradient-echo sequence were TR=3000ms, TE=40ms, Matrix=64x64, with a FOV=224mm. Again our voxel size=3.5mm^3, and 31 contiguous coronal slices were acquired.

Anatomic information for regions of interest (ROI) analysis was obtained with an MP-RAGE sequence. T1-weighted images were collected with TR=3000ms, TE= minimal, Flip angle = 18º, Matrix=256x192 (zero filled to 256), and a FOV=250mm. The voxel size was .9 mm x .9 mm x 1.3 mm and 120 contiguous sagittal slices (zero filled to 128) were obtained. For our image processing we used the 198 EPI images, 31 GE images, and 128 MP-RAGE images collected during the scanning session. The primary steps of the image processing were: motion correction, distortion correction, spatial normalization, smoothing, and statistical analysis. First, the EPI images were realigned, using an estimation of head movement, relative to the last image (the last image is used because it immediately precedes the GE sequence). The EPI files were co-registered and re-sliced using a sinc interpolation to generate a mean EPI image as well as registered EPI images. The mean EPI image was registered to the GE, a distortion-free image, in order to generate a warp file that will be applied to all the EPI images to correct for distortion. The next step was spatial normalization in which the images were transformed to a standard anatomical space (Talairach and Tournoux, 1988) using a T1 brain template. This procedure facilitates intersubject comparison. Finally, images were smoothed with a 7 mm Gaussian kernel (twice the voxel size) so that they were appropriate for statistical analyses.

Salivary samples were obtained at 10am and 4pm for cortisol levels; in addition, we administered 1mg of dexamethasone and measured 10am salivary cortisol the following morning.

Self-report of baseline State/Trait Anxiety and Perceived Stress were obtained using scales by Spielberger (1970) and Kuiper (1986), respectively.

Analysis

To evaluate the effects of arousal on the entire group's mean accuracy and response time, we performed a repeated measures analysis of variance, with arousal and difficulty level as the independent measures, and performance accuracy and response time as the dependent measures. Because of consistent order effects, described below in the Results section, we included testing order (whether the neutral or the arousal condition occurred first) as a covariate. To evaluate differences between tasks, we performed a bivariate correlational and linear regression analyses; task*condition*physiological variables interactions were assessed using MANOVA. To further evaluate differences between subjects, and their relationships to physiological variables, we first separated subjects into K-mean clusters, based on their fMRI activation of the left amygdala in response to neutral and aversive visual stimuli. Clusters were defined as non-responders, who showed minimal activation of the left amygdala in response to both neutral and aversive stimuli, selective high responders, who activated in response to aversive but not neutral stimuli, and non-selective high responders, who responded highly to both aversive and neutral stimuli. Using these clusters, we then performed a between-group analysis of variance to determine whether different clusters corresponded with significantly different task performance.

Results

Mean Performance Under Arousal

As shown in Table 1, arousal had a significant impact on mean cognitive performance. Prepulse inhibition was reduced an average of 3% ($p = 0.057$, $F = 3.88$) under the arousal condition, particularly in the second half, and showed diminished habituation under arousal. Increase in baseline-corrected skin conductance ($p = 0.000$, $F = 22.418$), a measure of sympathetic nervous system activation, and decrease of baseline-corrected heart-rate ($p = 0.041$, $F = 4.519$), a measure of parasympathetic nervous system activation, confirmed validity of the visual stimuli in causing an emotional response. Response time was significantly shortened for the Flanker Task ($p = 0.000$, $F = 18.022$), with accompanying decrease in accuracy; Flanker Task Efficiency (accuracy/response time) was reduced an average of 3% during the arousal condition. On the Flanker Task, congruence had a significant impact on performance: the incongruent condition had lower accuracy and longer response times than the congruent condition, which in turn had lower accuracy and longer response times than

the control condition. Response time was significantly lengthened for the Stroop task ($p = 0.000$, $F = 18.271$) in the arousal condition, with accuracy virtually unaffected; Stroop Task Efficiency (accuracy/response time) was also reduced by an average of 3% during the arousal condition.

Selective Attention versus Orienting Response

We found that efficiency on the Stroop Task, which measures the strength of the orienting response to an emotionally-valent stimulus, and efficiency on the Flanker Task, which measures the ability to focus on a (neutral) target and ignore distractors under emotionally-valent conditions, were related, as predicted. The correlation was stronger for the neutral condition ($r = 0.386$; $p = 0.015$) and weaker for the arousal condition ($r = 0.284$; $p = 0.068$). The difference between the two conditions likely resulted from subjects' performance on the orienting task being more affected by arousal ($F = 18.088$; $p = 0.000$) than their performance on the selective attention task ($F = 13.020$; $p = 0.001$). Since both tasks used similar emotionally-valent stimuli (words), these results suggest that arousal is more directly tied to the orienting response than to selective attention.

Sensorimotor Gating versus Selective Attention

While both PPI and performance on the Flanker Task were affected by arousal, the arousal condition had a stronger effect on the Flanker Task ($F = 13.020$; $p = 0.001$) than on pre-pulse inhibition ($F = 3.767$; $p = 0.061$). Our correlation and cluster analyses did not show either a direct, inverse, or hierarchical relationship between their pre-pulse inhibition and performance on the Flanker task, suggesting that pre-attentive and attentive cognitive filtering are distinct processes, mediated by different neural networks.

Individual Variability

We found a large range of variability on all cognitive variables between our healthy test subjects. For example, on the Pre-Pulse Inhibition (PPI) Task, 58% showed a relative decline in PPI under the arousal condition (ranging from 2% to 9% decreased PPI), while 42% showed a relative increase in PPI under the arousal condition (ranging from 1% to 2% increased PPI). On the Flanker Task, while 47% of all subjects showed a relative decline in efficiency under the arousal condition (ranging from 3% to 58% decline), another 51% showed a relative increase in efficiency under the arousal condition (ranging from 1% to 34% improvement). Two percent had identical scores on the neutral and arousal conditions. On the Stroop Task, 60% of all subjects showed a relative decline in efficiency under the arousal condition (ranging from 1% to 36% improvement), another 38% showed a relative increase in efficiency under the arousal condition (ranging from 3% to 27% improvement. Two percent had identical scores on the neutral and arousal conditions. Thus results, including our own, that report a mean decrease in sensorimotor gating and cognitive efficiency for selective attention and orienting tasks under arousal are statistically correct but are missing an interesting and potentially important part of the picture.

As shown in Figures 1, 2, and 3, our cluster analyses indicate that neural reactivity was a significant factor in predicting whether individuals' cognition (specifically sensory gating and selective attention) was positively or negatively affected by arousal (for overall, trend: $p = 0.106$, $F = 3.000$; for Flanker Task: $p = 0.037$, $F = 6.250$). For pre-attentive sensorimotor gating, selective attention, and orienting, individuals who showed selective high activation of the amygdala to aversive visual stimuli showed improvement on the arousal condition, while individuals who showed non-selective high activation of the amygdala to both neutral and aversive visual stimuli showed strong decline on the arousal condition. Non-responders, those individuals who showed minimal amygdala activation to either condition, showed small decline on the arousal condition.

Endocrine and subjective perception of baseline perceived stress were predictive of mean performance and physiological reactivity (electrodermal activity and heart rate). Afternoon (4pm) cortisol levels had significant between-subject effects for heart rate ($p = 0.043$, $F = 4.270$) and electrodermal activity ($p = 0.051$, $F = 4.270$), as well as a trend between-subjects effect for PPI ($p = 0.100$, $F = 2.947$). Post-dexamethasone cortisol, a measure of endocrine negative feedback loops, was related to mean PPI for both conditions ($p = 0.029$, $F = 5.432$). Mean PPI decreased with subjective perception of baseline perceived stress ($p = 0.087$, $F = 44.000$). For example, individuals with Perceived Stress Scale scores that clustered around 4.50 ("Low Stress," SD = 2.393) on the Perceived Stress Scales showed mean PPI of 50.993, individuals with scores that clustered around 13.18 ("Moderate Stress," SD = 2.538) showed mean PPI of 42.222, and individuals with scores of 21.33 ("Pronounced Stress," SD = 2.739) showed mean PPI of 26.314. On the Stroop Task, Low Stress individuals showed mean response times of 924.809 ms, Moderate Stress individuals showed mean response times of 935.104 ms, and Pronounced Stress individuals showed mean response times of 979.127 ms. This pattern was not observed for the Flanker Task, although Low Stress individuals still showed shorter response times than Pronounced Stress individuals ($p = 0.009$, $F = 5.650$).

There were no prominent age or gender effects.

Conclusions

Our study of emotional arousal's impact on cognition demonstrates that even the mild arousal induced in a controlled laboratory setting is sufficient to show consistent changes under two conditions, both of which are common in actually dangerous contexts. For tasks performed under conditions of arousal but without the possibility of orienting to the aversive stimulus (such as the Flanker Task), we recorded decrease in response time

and accuracy. For tasks with the possibility of orienting to the aversive stimuli but which require attending away from the aversive stimulus (such as the Stroop Task), we saw no loss of accuracy, but recorded significant increase in response time. Our analysis indicates that selective attention, which is fully attentive, and the orienting response, which is only semi-attentive, are related. Prepulse inhibition, which is wholly pre-attentive, was not correlated with either the Flanker or the Stroop Tasks. Finally, our analysis of variability indicates that, while most individuals are negatively influenced by arousal, others are not. Our use of neural clusters suggests that it is the *selectivity* of the neural arousal response, rather than its amplitude, that corresponds with the direction of impact; future studies examining neural activation and performance under more severe stress may shed light on the practical implications of these results.

Figures

Table 1: Estimated Marginal Means for Cognitive/Physiological Measures During Neutral & Arousal Conditions

Measure (N = 39)	Condition	Mean	Std. Error
Prepulse Inhibition	neutral	40.486	4.481
	arousal	39.173	4.854
Flanker Task Acc	neutral	.970	.010
	arousal	.966	.012
Flanker Task RT	neutral	652.208	43.341
	arousal	637.963	37.253
Stroop Acc	neutral	.959	.011
	arousal	.966	.009
Stroop RT	neutral	984.073	48.350
	arousal	1007.698	46.796
Baseline Corr HR	neutral	-1.216	.307
	arousal	-2.315	.509
Baseline Corr EDA	neutral	.131	.055
	arousal	.189	.075

Evaluated at covariate: TSTORDER = 1.4595.

Table 2: Tests of Within-Subjects Contrasts for Cognitive/Physiological Changes During Neutral vs. Arousal Conditions

Measure (N = 39)	F	Sig.
Prepulse Inhibition	3.883	.057
Flanker Task Acc	2.307	.138
Flanker Task RT	18.022	.000
Stroop Acc	.197	.660
Stroop RT	18.271	.000
Baseline Corrected Chge in HR	4.519	.041
Baseline Corrected Chge in EDA	22.418	.000

Figure 1: Preattentive Sensorimotor Gating for NonResponder, Selective High Responder, and Nonselective High Responder Groups Defined by Activation of the Left Amygdala

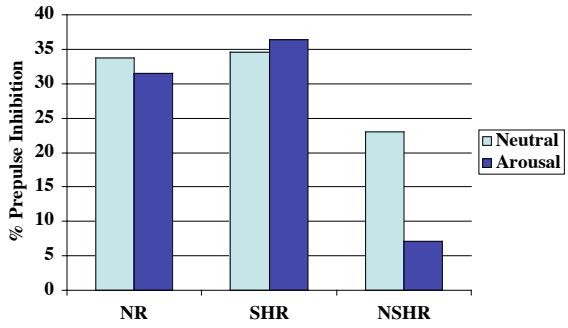


Figure 2: Selective Attention for NonResponder, Selective High Responder, and Nonselective High Responder Groups Defined by Activation of the Left Amygdala

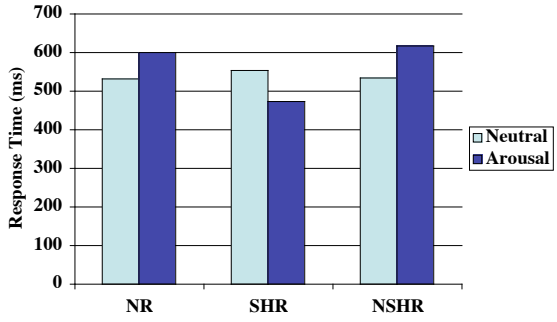
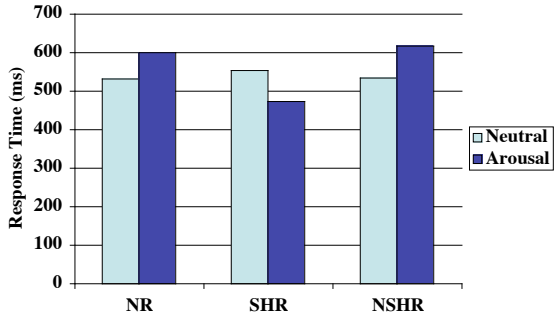


Figure 3: Selective Attention for NonResponder, Selective High Responder, and Nonselective High Responder Groups Defined by Activation of the Left Amygdala



Acknowledgements

This study was funded by the Office of Naval Research, Grant #N00014-04-1-0051

References

Bacon, S.J. (1974). Arousal and the Range of Cue Utilization. *Journal of Experimental Psychology, 102*(1), 81-7.

Baddeley, A.D. (1972). Selective attention and performance in dangerous environments. *British Journal of Psychology, 63*(4), 537-546.

Berkun, M.M., Bialek, H.M., Kern, R.P. & Yagi, K. (1962). Experimental studies of psychological stress in man. *Psychol. Monogr. 76*(15); (Whole no. 534).

Blumenthal, T.D. (1993). Prepulse Inhibition of the startle eyeblink as an indicator of termporal summation. *Perception*, Vol 57(4), p. 448-494.

Bruner, J.S., Matter, J., & Papnek, M.L. (1955). Breadth of learning as a function of drive level and mechanization. *Psychological Review, 62*, 1-10.

Combs, A.W., & Taylor, C. (1952). The effect of the perception of mild degrees of threat on performance. *Journal of Abnormal Social Psychology, 47*, 420-424.

Cornsweet, D.M. (1969). Use of cues in the visual periphery under conditions of arousal. *Journal of Experimental Psychology, 80*, 14-18.

Dalgleish, T. (1995). Performance on the emotional stroop task in groups of anxious, expert, and control subjects: a comparison of computer and card presentation formats. *Cognition and Emotion, 9*, (4), 341-362.

D'Alfonso, A.A.L., Van Honk, J., Herman, E., Posta, A., & de Han, E.H.F. (1999). Laterality effects in selective attention to threat after repetitive transcranial magnetic stimulation at the prefrontal cortex in female subjects. *Neuroscience Letters*, 280, (3), 195-198.

Endicott, J. & Spitzer, R.L. (1978). A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch Gen Psychiatry*. Jul; 35 (7): 837 844.

Endicott, J., Spitzer, R.L., Fleiss, J.L., & Cohen, J. (1976). The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance. *Arch of Gen Psychiatry, 33,* 766-771.

Fan, J., McCandliss, B.D., Sommer, T., Raz, A., & Posner, M.I. (2001). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience* (in press).

Hammerton, M. & Tickner, A.H. (1967). Tracking under stress. Medical Research Council Report no. APRC 67/CS 10(A).

Hockey, G.R.J. (1970a). Effect of loud noise on attentional selectivity. *Quarterly Journal of Experimental Psychology, 22*, 28-36.

John, C. H. (1988). Emotionality ratings and free-association norms of 240 emotional and non emotional words. *Cognition & Emotion. Vol 2(1),* 49-70.

Kuiper NA, Olinger LJ, Lyons LMJ (1986). Global perceived stress level as a moderator of the relationship between negative life events and depression. *Human Stress*. Winter;12(4):149-53.

McKenna, F. P. &Sharma, D. (1995). Intrusive cognitions: An investigation of the emotional Stroop task. [Journal Article] *Journal of Experimental Psychology: Learning, Memory, & Cognition. Vol 21*(6), 1595-1607.

Rockett, F.C. (1956). Speed of form recognition as a function of stimulus factors and test anxiety. *Journal of Abnormal Soc Psych, 53,* 137-202.

Spielberger, C.D., Gorsuch, R.L., Lushene, R. (1970). Manual for the State-Trait Anxiety Inventory: STAI ("Self-Evaluation Questionnaire"). Palo Alto, CA: Consulting Psychologists Press.

Walker, N.K. & Burkhardt, J.F. (1965). The combat effectiveness of various human operator controlled systems. *Proc 17th U.S. Military Operations Research Symposium.*

Weltman G., Smith J.E., & Egstrom G.H. (1971). Perceptual narrowing during simulated pressure-chamber exposure. *Human Factors. 13*(2)*:*99-107.

Embodied Cognition and The Nature of Mathematics: Language, Gesture, and Abstraction

Rafael E. Núñez (NUNEZ@Cogsci.Ucsd.Edu)
Department of Cognitive Science, 9500 Gilman Drive
La Jolla, CA 92093-0515 USA

The Cognitive Science *of* Mathematics

Mathematics is a highly technical domain, characterized by the fact that the very entities that constitute it are idealized mental abstractions. These entities cannot be perceived directly through the senses. Even the simplest entity in, say, Euclidean geometry (i.e., a point, which only has location but no extension,) can't actually be *perceived*. This is obvious when the entities in question involve infinity (e.g., limits, least upper bounds, mathematical induction, infinite sets, points at infinity in projective geometry and so on) where, by definition, no direct experience can exist with the infinite itself. Lakoff and Núñez (1997, 2000) and Núñez (2000a, 2000b, in press), inspired by theoretical principles of embodied cognition and using mainly techniques from cognitive linguistics (especially cognitive semantics) have suggested that these idealized abstract technical entities in mathematics are created by the human imaginative mind via a very specific use of everyday bodily-grounded cognitive mechanisms such as conceptual metaphors, conceptual blends, analogical reasoning, fictive motion, aspectual schemas, and so on (see also Núñez & Lakoff 1998, in press). Mathematics is, according to this view, a specific powerful and stable product of human imagination. The claim is that a detailed analysis of the inferential organization of mathematical concepts, theorems, definitions, and axioms (Mathematical Idea Analysis) provide cognitive foundations of mathematics itself. From this perspective, mathematics *is* the network of bodily-grounded inferential organization that makes it possible. The study of these foundations and their extended inferential organization constitutes one of the most important goals of the cognitive science of mathematics.

Towards Convergent Empirical Evidence: Gesture and Conceptual Mappings

So far the work by Lakoff & Núñez on the cognitive science of mathematics has been based mainly on cognitive semantics, focusing on the conceptual mappings (conceptual metaphors, blends, metonymies, frames, etc.) that model the inferential organization of mathematical concepts. Some important questions, however, remain open:

1. Are the mathematical concepts considered by Lakoff & Núñez to be metaphorical (e.g., least *upper* bound, space-*filling* curve, point *at infinity* in projective geometry, etc.) simply cases of "dead" metaphors with no actual metaphorical semantic content? In other words, is the meaning and inferential organization of these concepts fully characterized by their *literal* formal mathematical definition (as it is often claimed in mathematics proper)?

2. If the answer to (1) is negative, what then is the psychological reality of the suggested conceptual metaphors involved?

In this presentation I intend to address these two questions by:

a) Focusing on cases in mathematics where *dynamic* language is used to refer to mathematical objects that, within mathematics proper, are completely defined in *static* terms via the use of universal and existential quantifiers and set-theoretical entities. For example, when treating limits of infinite series classic mathematics books often make statements like this one: "We describe the behavior of s_n by saying that the sum s_n *approaches* the limit 1 as n *tends* to infinity, and by writing $1 = 1/2 + 1/2^2 + 1/2^3 + 1/2^4 + \ldots$" (Courant and Robbins, 1978, p. 64). This statement refers to a sequence of discrete and static partial sums of s_n (real numbers), corresponding to successive discrete and static values taken by n. Technically, numbers, as such, don't move, therefore no dynamic language should provide any literal meaning in cases like this one.

b) Providing evidence from gesture studies, supporting the claim that the conceptual metaphorical nature of these mathematical linguistic expressions is indeed psychological real, operating under strong real-time and real-world constraints. I will build on the increasing evidence showing the extremely close relationship between speech, thought, and gesture production at a behavioral (McNeill, 1992), developmental (Iverson & Thelen, 1999; Bates & Dick, 2002), neuropsychological (McNeill & Pedelty, 1995; Hickok, Bellugi & Klima, 1998), psycholinguistic (Kita, 2000), and cognitive linguistic level (Lidell, 2000; Núñez & Sweetser, 2001).

I will argue that the dynamic component of many mathematical ideas is constitutive of fundamental mathematical ideas such as limits, continuity, and infinite series, providing essential inferential organization for them. The formal versions of these concepts, however, neither generalize nor fully formalize the inferential organization of these mathematical ideas (i.e., ε-δ definition of limits and continuity of functions as framed by the arithmetization program in the 19th century). I suggest that these deep cognitive incompatibilities between dynamic-wholistic entities and static-discrete ones explain important dimensions of the great difficulties encountered by students when learning the modern technical version of these notions (Núñez, Edwards, and Matos, 1999). In order to support my arguments I will analyze converging linguistic and gestural data involving infinite series, limits and continuity of functions, showing the crucial role played by conceptual metaphor and fictive motion (Talmy, 1996) in constituting the inferential organization of these fundamental concepts.

References

Bates, E. & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychology*, *40*(3), 293-310.

Courant, R. & Robbins, H. (1978). *What is mathematics?* New York: Oxford.

Hickok, G., Bellugi, U. and Klima, E. (1998). The neural organization of language: evidence from sign language aphasia. *Trends in Cognitive Science*, *2*(4), 129-136.

Iverson, J. & Thelen, E. (1999). Hand, mouth, and brain: The dynamic emergence of speech and gesture. In R. Núñez & W.J. Freeman (Eds.), *Reclaiming cognition: The primacy of action, intention, and emotion*. Thorverton, UK: Imprint Academic.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture*. Cambridge, UK: Cambridge University Press.

Lakoff, G. & Núñez, R. (1997). The metaphorical structure of mathematics: Sketching out cognitive foundations for a mind-based mathematics. In L. English (Ed.), *Mathematical reasoning: Analogies, metaphors, and images*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lakoff, G. & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.

Lidell, S. (2000). Blended spaces and deixis in sign language discourse. In D. McNeill (Ed.), *Language and gesture*. Cambridge, UK: Cambridge University Press.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: Chicago University Press.

McNeill, D. & Pedelty, L. (1995). Right brain and gesture. In K. Emmorey & J.S. Reilly (Eds.), *Gesture, sign, and space*. Hillsdale, NJ: Erlbaum.

Núñez, R. (2000a). Mathematical idea analysis: What embodied cognitive science can say about the human nature of mathematics. Opening plenary address in *Proceedings of the 24th International Conference for the Psychology of Mathematics Education* (pp. 3–22). Hiroshima, Japan.

Núñez, R. (2000b). Conceptual Metaphor and the embodied mind: What makes mathematics possible? In F. Hallyn (Ed.), *Metaphor and analogy in the sciences*. Dordrecht, The Netherlands: Kluwer Academic Press.

Núñez, R. (in press). Creating mathematical infinities: Metaphor, blending, and the beauty of transfinite cardinals. *Journal of Pragmatics*.

Núñez, R., Edwards, L., Matos, J.F. (1999). Embodied Cognition as grounding for situatedness and context in mathematics education. *Educational Studies in Mathematics*, *39*(1-3): 45-65.

Núñez, R. & Lakoff, G. (1998). What did Weierstrass really define? The cognitive structure of natural and ε-δ continuity. *Mathematical Cognition*, *4*(2): 85-101.

Núñez, R. & Lakoff, G. (in press). The cognitive foundations of mathematics. In J. Campbell (Ed.), *Handbook of mathematical cognition*. New York: Psychology Press.

Núñez, R. & Sweetser, E. (2001). Spatial embodiment of temporal metaphors in Aymara: Blending source-domain gesture with speech. *Proceedings of the 7th International Cognitive Linguistics Conference* (pp. 249-250). Santa Barbara, CA: University of California.

Talmy, L. (1996). Fictive motion in language and "ception." In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space*. Cambridge: MIT Press.

Conceptual muddles: Truth vs. Truthfulness, Logical vs. Psychological Validity, and the non-monotonic vs. defeasible nature of human inferences.

Walter J. Schroyens (Walter.Schroyens@psy.kuleuven.ac.be)
Laboratory of Experimental Psychology, University of Leuven
Tiensestraat 102, Leuven, B-3000, Belgium

Deductive logic is concerned with logical validity (henceforth 'L-validity'). An inference is L-valid only when it must necessarily be true *if* the premises are true. This classic definition has a built-in truth-assumption. The truth of L-valid inferences is always a hypothetical truth, not a factual truth. This means that the factual truth of the premises and/or conclusion does not affect L-validity and that a L-valid inference cannot become L-invalid by adding new information. Technically speaking, deductive logics are *monotonic*. This stands in apparent contrast with the psychological evidence to the contrary, showing that the premise or conclusion believability affects the reasoning process. Common-sense reasoning is *defeasible* and *non-monotonic* in nature. A conclusion that is believed to be true at one point can be considered false later, and an inference that is considered valid at one point can later be revoked and re-evaluated as invalid. (Note that we used the unqualified term 'valid', not 'L-valid'). Consider, e.g., the classic benchmark example: "If it is a bird, then it can fly. Tweety is a bird. Hence, Tweety can fly." This is a L-valid argument. Assuming it is true that 'if something is a bird, then it can fly' and assuming that Tweety is a bird, it follows necessarily that it would be true that Tweety can fly. People will nonetheless retract the conclusion that 'Tweety the bird can fly' when being given the information that Tweety is in fact an ostrich.

Some theorists have created polemics between what they call 'logic theories' and their own probabilistic theories of human reasoning. The core argument against theories of human deduction is their presumed incapability of dealing with the defeasibility of common-sense reasoning. I will argue that there is only an *apparent* contrast between logic's monotonicity and common-sense reasoning's defeasibility. It is only when we are sure that people are reasoning hypothetically that defeating an inference would show that the monotonicity of deductive logics is problematical.

Let us assume that people aim to establish L-validity. If so, people are abandoning the truth-assumption when defeating an inference. The existence of ostriches falsifies the claim that 'if something is a bird, then it can fly. It is not always true that when something is a bird, it can fly. If people abandon the truth-assumption when confronted with the added information, they are shifting from one notion of validity (i.e., L-validity) to another notion of validity (let's call it P-validity). This means that they are not changing an *L-valid* inference into a *L-invalid* inference, but are changing an L-valid inference into a *P-valid* inference. This example indicates that though defeasible, common-sense reasoning is not necessarily non-monotonic.

Theorists who argue against logic theories contest that questioning the literal truth of, e.g., 'if it is a bird, then it can fly' is involved in defeasible reasoning: "surely [this] mischaracterizes people's cognitive attitude towards this and a million other commonsense generalizations" (Oaksford & Chater, 1998, p. 5). This claim as regards the psychological 'truthfulness' of a logically false conditional is not congruent with reality. We asked 150 first-year psychology undergraduates to judge whether the conditional is strictly speaking false when the context either did or id not include TF cases. These cases reflect situations where the antecedent is satisfied while the consequent is not (e.g., birds that do *not* fly). When there were TF cases, 83% of them said it is strictly speaking false. In case there were no such falsifying TF cases, 89% said the conditional was true. Moreover, with the false conditionals, 91% selected a conditional of the form 'if p then possibly q' as the best description of the situation. With a true conditional, 93% effectively preferred 'if p then q' as the best description. This first study used abstract materials (coloured figures). In a second study we asked 44 first-year psychology students to "think about the fact that for instance ostriches and penguins are also birds (and can not fly)." Thirty-eight (86%) of them judged the conditional to be false. In short, the falsity of the conclusion 'Tweety flies' in real everyday inference, license the conclusion that "if it is a bird, then it can fly" is a false utterance.

To ground their intuition pump, Oaksford and Chater (1998) appeal to the comforting idea that there is true commonsense knowledge. "If our commonsense descriptions of the world and of ourselves are not candidates for truth then precious little else of what we call our commonsense knowledge of the world will be candidates for truth. We would then be in the paradoxical position of having to provide a system of human inference that is always based on false premises but which is nonetheless apparently capable of guiding successful action in the world!" (Oaksford & Chater, 1998, p. 5). There is really only an apparent contradiction (a paradox), not a contradiction. It is not problematical that there is preciously little (if any) knowledge that is strictly true. The induction problem still exists: every generalization is a potential over-generalization. However, the fact that some birds do not fly does not make it senseless to use the generalization that birds fly. An absolute truth is universally applicable, but if something is not universally applicable then this does not imply that it is inapplicable. It might be inapplicable (applicable to none) or applicable to some (but not all). The demonstrable fact that most of our commonsense generalizations are false (i.e., not strictly true), marks that they only have a certain degree of truth: They are false, but applicable. Verity is not verisimilitude.

References

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.

The Relations Between Causal (x2) and Counterfactual Reasoning, the Hindsight Bias and Regret (and the kitchen sink)

Barbara A. Spellman (spellman@virginia.edu)
Department of Psychology, University of Virginia, P.O. Box 400400
Charlottesville, VA 22904-4400 USA

The research areas of causal and counterfactual reasoning, hindsight bias and regret, have often been studied in isolation, sometimes studied in pairs, and occasionally studied in triads. I suggest that there are common mechanisms shared by these judgments that explain how, when, and why they will (a) be similarly or differently affected by information and (b) influence each other.

To start, I distinguish two types of causal reasoning: the types of judgments we make in science when we have multiple examples of causes and effect and the types of judgments we make in law when we want to figure out the cause of a one-time only event. In the former, an important cue to causality is covariation -- a cause is something that increases the probability of an effect above its usual probability. I then draw an analogy to the latter -- and assume that a causality judgment about a person or event is a function of how much that person or event increases the probability of the eventual outcome above its "baseline" probability (i.e., its natural probability of occurring).

$$C \approx p(Oafter) - p(Obefore)$$

The equation above represents how a causality judgment is a function of the estimated probability of the eventual outcome occurring after the target cause has occurred [p(Oafter)] and the estimated probability of the eventual outcome occurring before the target cause has occurred [p(Obefore)].

But for one-time events, how can people make probability judgments? I suggest that such judgments rely on pre-existing knowledge -- especially of previous covariations and causal mechanisms -- and counterfactual reasoning. The equation below expands the one above by putting each estimate over 1 (i.e., p(Oafter) + p(~Oafter) = 1).

$$C \approx \frac{p(Oafter)}{p(Oafter) + p(\sim Oafter)} - \frac{p(Obefore)}{p(Obefore) + p(\sim Obefore)}$$

That causality relies on counterfactual information in this manner explains the "if-only" and "even-if" effects -- ways in which considering counterfactuals affects causal judgments. For example, if someone takes an unusual route home, and then is in a car accident, she might think "If only I had taken my usual route." That counterfactual thought would increase the estimate of p(~Obefore), decrease the fraction on the right, and increase the causality assigned to the decision to take the unusual route.

The relation to the hindsight bias is clear: When do people make these probability estimates? Typically after events have unfolded. Thus, the hindsight bias is implicit in causality judgments. However, these equations also suggest ways in which the hindsight bias can be de-biased and, in particular, which kinds of counterfactuals should be most effective in doing so.

Finally, I argue that regret is both a counterfactual and causal emotion -- it depends on knowing that what you might have done could have changed an outcome. Our studies compare measures of causality with measures of regret. We find that regret depends on the difference between an actor's perceived causality for the (negative) outcome given his actual decision and the imagined causality for that outcome had an alternative decision been made. (Again, such causality judgments are made in hindsight.) Our experiments use this relation to explain "action" and "inaction" effects in regret judgments.

I hope to relate these analyses to other types of reasoning.

Acknowledgments

This research was supported by an NIMH Grant.

Relevant Publications

Spellman, B. A., Kincannon, A., & Stose, S. (in press). The relation between counterfactual and causal reasoning. Invited chapter to appear in D. R. Mandel, D. J. Hilton, & P. Catellani (Eds.), *The psychology of counterfactual thinking*. London: Routledge Research.

Spellman, B. A., & Mandel, D. R. (2003). Causal reasoning, psychology of. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (Vol 1, pp. 461-466). London: Nature Publishing Group.

Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems: Causation in Law and Science, 64(4),* 241-264.

Spellman, B. A., Price, C. M., & Logan, J. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition, 29,* 193-208.

Spellman, B. A., & Mandel, D. R. (1999). When possibility informs reality: Counterfactual thinking as a cue to causality. *Current Directions in Psychological Science, 8,* 120-123.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126,* 323-348.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7,* 337-342.

Does the practice of meta-cognitive description facilitate acquiring expertise?

Masaki Suwa (suwa@sccs.chukyo-u.ac.jp)
PRESTO, JST & School of Computer and Cognitive Sciences, Chukyo University
101 Tokodachi, Kaizu, Toyota, Aichi 470-0348, Japan

Expertise as Differentiation

How do people acquire expertise? Studies on machine learning in artificial intelligence have captured some aspects of human learning. But its limitation lies in that the representation of the target domain needs to be given in advance and fixed during the process in order for a learning mechanism to work. Contrarily, the ways people represent the external world evolve as they become experts. Experts are able to differentiate and perceive some features and relations in the world that would be meaningless to novices. This means that acquisition of expertise can be regarded as a process of becoming able to perceive what was not evident before. Gibson and Gibson (1955) described a similar process, in relation to expertise, in their case, wine-tasting: "Perceptual learning, then, consists of responding to variables of physical stimulation not previously responded to" (p. 34). The expertise-as-differentiation view was not necessarily discussed in studies on chess in which expert performance was explained by chunks (e.g. Chase and Simon, 1973).

Our previous studies on expert-novice differences in design showed similar findings. Designers draw sketches in the early design phase. Sketches are not only a record of generated ideas but also a stimulus for new ones. The success of a design process hinges on differentiately perceiving features and relations in sketches and generating interpretations of them. We found that expert designers were more capable of associating features and relations with functional issues (Suwa and Tversky, 1997). Further, perceiving features and relations unheeded before, which is difficult to do for novices, was the major driving-force for the generation of ideas for an expert designer (Suwa, 2003).

Hypothesis: The Practice of Meta-cognitive Description will Foster Differentiation

Then, what kind of cognitive practices in a target domain help become able to differentiately perceive features and relations in the external world and interpret them? We have made a hypothesis that self-awareness of and thereby meta-cognitive descriptions of what one has perceived and conceived of will foster the ability of differentiation and thus facilitate acquisition of expertise (Suwa and Tversky, 2003). Anecdotal evidence comes from the domain of sports. A Japanese player in Major Baseball League, named Ichiro, said in a TV interview that it is through a persistent effort of describing how he has perceived the ball and how his body has reacted and hit it that he had become one of the most productive hit-maker. Our finding (Suwa and Tversky, 2003) that expert designers are better at a meta-cognitive skill of reorganizing perception and generating interpretations than novices is also supportive.

We have recently obtained empirical evidence from the case study of singing a song. A participant in the experiment continued to sing a song for 4 months, recording his voice in every trial of singing. During the period, he continued describing meta-cognitively, in the form of writing in a notebook, how he was utilizing his throat, breath and tongue and how that helped express his feeling and emotion. Evaluation of all the trials of singing by three musicians after 4 months revealed that his performance exhibited a U-shape learning curve, i.e. getting better in the beginning, then turning worse sharply and gradually getting better and better toward the end. This indicates that he was climbing steps toward acquiring expertise. An interesting finding is that the evaluation scores of the recorded songs correlated in a statistically significant manner with the amount of meta-cognitive descriptions accumulated for about one month up to the time of every trial of singing. This suggests that the practice of meta-cognitive description had a latent effect, not an instant one, on his performance.

We interpret this in the following manner. A meta-cognitive description is a kind of narrative created by self. Its validity is not assured anyhow. Important is, however, the practice of meta-cognition itself, not its validity. A meta-cognitive description as a narrative will effectively drive differentiation of features and relations in the external world, and thereby encourage the next cycle of meta-cognitive description. This cycle will lead up to acquiring expertise.

Acknowledgments

I am grateful to Barbara Tversky for insightful discussions.

References

Chase, W. G. & Simon, H. A. (1973). Perception in chess, *Cognitive Psychology*, 4, 55-81.

Gibson, J. J. & Gibson, E. J. (1955). Perceptual learning: differentiation or enrichment?, *Psychological Review, 62*, 32-41.

Suwa, M. (2003). Constructive perception: coordinating perception and conception toward acts of problem finding in a creative experience, *Japanese Psychological Research*, 45, 221-234.

Suwa, M. & Tversky, B. (1997). What do architects and students perceive in their design sketches?: a protocol analysis, *Design Studies*, 18, 385-403.

Suwa, M. & Tversky, B. (2003). Constructive perception: a meta-cognitive skill for coordinating perception and conception, *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1140-1144). Cognitive Science Society.

Papers

Task Interruption: Resumption Lag and the Role of Cues

Erik M. Altmann (ema@msu.edu)
Department of Psychology
Michigan State University
East Lansing, MI 48824

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
Naval Research Laboratory, Code 5513
4555 Overlook Avenue S. W.
Washington, DC 20375

Abstract

The consequences of interrupting someone in the middle of a complex task are of considerable practical and theoretical interest. We examine one behavioral measure of the disruption caused by task interruption, namely the *resumption lag*, or the time needed to "collect one's thoughts" and restart a task after an interruption is over. The resumption lag (in our task environment) was double the interval between uninterrupted actions (3.8 s vs. 1.9 s), indicating a substantial disruptive effect. To probe the nature of the disruption, we examined the role of external cues associated with the interrupted task, finding that cues available immediately *before* an interruption facilitate performance immediately *afterwards* (reducing the resumption lag). This *cue-availability* effect suggests that people deploy preparatory perceptual and memory processes, apparently spontaneously, to mitigate the disruptive effects of task interruption.

Introduction

For better or worse, interruptions are part of everyday life. For better, interruptions are an essential part of efficient communication, among people and between people and machines. For worse, interruptions can be annoying, and can seem disruptive. For example, the annoyance of unwanted telephone solicitations drove the recent overwhelming popularity of "do not call" registries in the United States. Similarly, consider the "software assistant" included in Microsoft products in the late 1990s. If Word, for example, detected what it thought was a letter being drafted, it would freeze the keyboard and demand to know if the user needed "help" — a feature that in more recent editions of the software is no longer enabled by default.

There are many parameters to how an interruption is structured — including interruption duration, for example, or whether the interrupted person has control over interruption timing (McFarlane & Latorella, 2002) — and there is also a range of different behavioral measures on which the assess the impact of an interruption. In one classic result, Zeigarnik (1927/1938) found that interrupted tasks were actually remembered better, in terms of recalled detail, than tasks that were allowed to run to completion. In more applied work, however, interruptions have been found to have detrimental effects on situational awareness in dynamic task environments like aviation (Latorella, 1996), where losing one's place in a checklist during takeoff, for example, can have catastrophic results (NTSB, 1988).

The current study examines the disruptive effects of interruption in terms of the time needed to resume the primary (interrupted) task after the secondary (interrupting) task is complete. In less formal terms, we examine the time needed to collect one's thoughts, or pick up the thread again, when an interruption is over and we can return to what we were doing before. There is surprisingly little research focused on this measure, and what there is is distributed across a variety of domains and paints no clear picture of whether interruptions are disruptive or not. For example, in a study of interruption of administrative and clerical workers, disruptive effects of interruption were difficult to detect (Zijlstra, Roe, Leonora, & Krediet, 1999), and interruption in a simple question-answering task can actually improve performance, measured in terms of overall accuracy and time-on-task (Speier, Vessey, & Valacich, 2003). In the mainstream cognitive psychology literature, research on "task switching" (Monsell, 2003) would seem to be relevant, as studies in this domain typically focus on the "switch cost" associated with shifting from one task to another. However, this literature is perhaps not so aptly named; the "tasks" used in task-switching studies take a few hundred milliseconds to complete, with switch cost a small and not particularly relevant fraction of that (Altmann, in press). We are interested in higher-level tasks with greater ecological validity, where switch cost is measured not in tens of milliseconds, but in seconds or longer.

In operational terms, the dependent measure in the current study is the *resumption lag*, illustrated in Figure 1. The resumption lag is the time interval separating the end of the secondary task and the first subsequent action taken by the human operator in the primary task. We report first a comparison of this resumption lag to an estimate of the time interval that usually separates actions in the primary task, to give a sense of the absolute magnitude of the disruptive effect; to preview, the mean inter-action interval, in the highly interactive primary task we are using, is roughly 2 s, and the resumption lag is roughly 4 s, indicating a substantial disruption both in absolute and relative terms.

We then report on two factors that have the potential to reduce the resumption lag. Both focus on the *interruption lag*, also illustrated in Figure 1. The interruption lag is a brief transitional interval immediately preceding an interruption, during which the operator knows of the pending interruption but is not yet engaged by it. An example is the time between the phone starting to ring and the act of actually taking the call; during this interval, there is a brief opportunity to complete a thought, for example, or negotiate quickly with a conversation partner (physically present) how and when to resume after the call is over. Many real-world interruptions, even more urgent ones, afford a brief interruption lag; even when a fire alarm sounds, one is still likely to take time to save changes to an electronic document, for example, before evacuating.

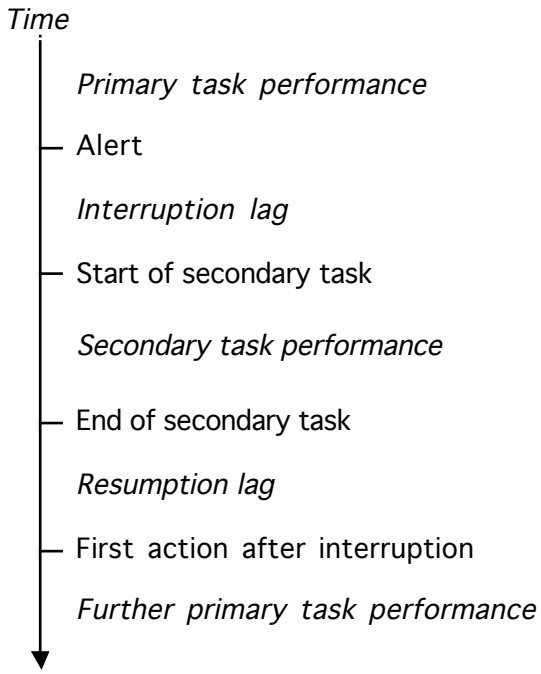


Figure 1: Time course of an interruption. For example, if the alert is the phone starting to ring, then the secondary task is the ensuing phone conversation.

Two characteristics of the interruption lag are examined here: (1) whether or not the primary task display is perceptually available during this brief period, and (2) the actual duration of this period. The relevance of these two factors is predicated on a memory model of what makes interruptions disruptive (Altmann & Trafton, 2002; Trafton, Altmann, Brock, & Mintz, 2003). The basic premise of the model is that during an interruption, the cognitive representations that support performance of the primary task will decay, in particular relative to the cognitive representations that support performance of the secondary task. Thus, when resuming the primary task, retrieval cues will be necessary to re-activate the relevant representations.

This memory analysis predicts, qualitatively, that the interruption lag — the brief interval *before* an interruption — has a crucial role to play in facilitating resumption *after* the interruption. During the interruption lag, when the operator is aware that he or she will soon be interrupted but can still focus mentally on the primary task, there is an opportunity to "prepare to resume," for example by prospectively encoding goals to accomplish at resumption (Trafton et al., 2003). To the extent that people do engage in such preparatory processing, it should help to have the primary-task display perceptually available during the interruption lag, to allow retrieval cues to be quickly accessed and accurately encoded. Thus, to build on earlier evidence that people do prepare to resume (Trafton et al., 2003), we asked here whether cue availability is a factor in this process. In the *cue* condition the primary task display was preserved during the interruption lag, whereas in the *no-cue* condition the screen went blank during the interruption lag (starting with onset of an alert signaling the pending interruption). Moreover, because processes like perceptual search and memory encoding take time, we varied the duration of the interruption lag across experiments, to examine what length of interruption lag would render cue availability effective in reducing the resumption lag.

Experiments

We conducted four experiments, with interruption lags of two, four, six, and eight seconds respectively. These values were based on evidence that an 8-second interruption lag is enough to allow people to (at least partially) prepare to resume (Trafton et al., 2003). The primary task involved planning and resource allocation subject to constraints, and thus involved a substantial amount of state information to be represented cognitively. The secondary task was less complex but nonetheless involved a series of tightly-spaced forced-choice decisions unfolding over a 30- to 45-second period. Interruption timing was under system, rather than operator, control, a factor that tends to aggravate the disruptive effects of interruption (McFarlane & Latorella, 2002).

The independent variable within each experiment was whether or not cues were available during the interruption lag. The main dependent variable was the resumption lag (Figure 1), but we also compare the resumption lag to the mean interval between primary-task actions, to estimate the overall disruptive effect of an interruption.

Method

Participants Ninety-six Michigan State University undergraduates participated in exchange for partial credit toward a course requirement. Each of the four experiments involved 24 participants, randomly assigned to the cue and no-cue conditions (described below).

Materials The primary task was a complex resource-allocation task (Trafton et al., 2003) in which participants were asked to defeat a set of simulated destinations using simulated tanks. Participants selected which destinations to attack and in what order, and issued tanks with appropriate amounts of fuel and munitions. Fuel was consumed in reaching a destination, and munitions were consumed in engaging it, but tank payload was limited, as was the total number of tanks and other resources available. Points were awarded for defeating destinations and subtracted for consuming resources.

Figure 2 shows a view of the primary-task display as it normally appears when the participant is doing the task. There is a central window with buttons for allocating tanks to missions, choosing destinations, and displaying a map with distances between destinations. There is also an area for displaying mission outcomes (whether a destination was defeated, whether a tank ran out of fuel, etc.) To the left are windows showing the supply pool (available fuel, munitions, and tanks) and windows for outfitting heavy and light tanks with varying amounts of fuel and munitions. To the right is a window showing the participant's scoring

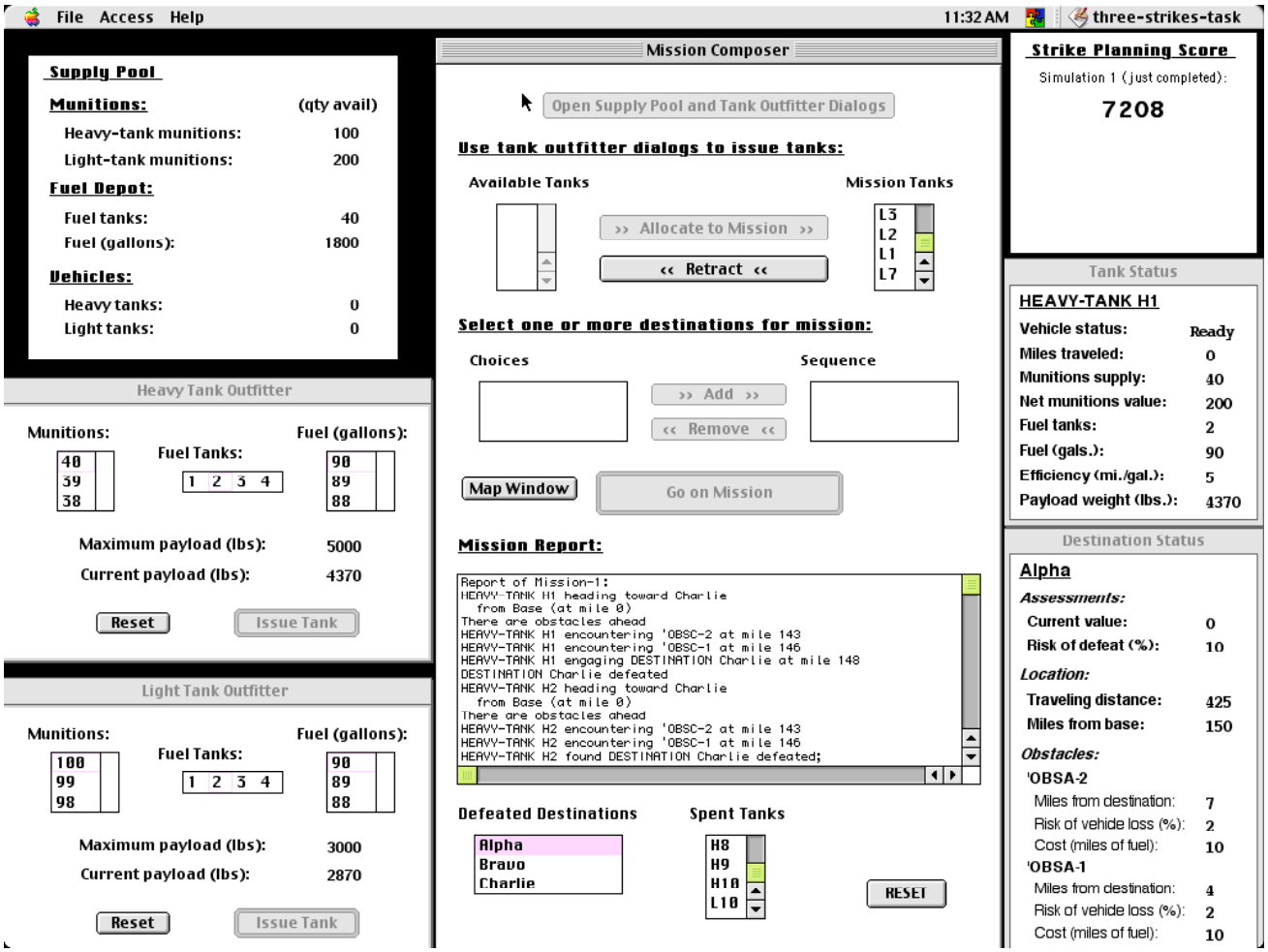


Figure 2: Screen shot of the primary task display during normal performance (see text for summary).

history from past simulations, and windows showing the status of selected tanks and destinations.

Figure 3 shows (at a more reduced scale) the state of the display during the interruption lag in the no-cue condition. To signal the pending interruption (and thus mark the start the interruption lag), an "eyeball alert" appeared in the top-right corner of the display. In the no-cue condition, the primary-task display was blanked out simultaneously with alert onset, whereas in the cue condition the primary-task display was preserved. In both conditions, with the start of the secondary task (and thus the end of the interruption lag), the primary-task display (whatever its state) was completely erased and replaced with the secondary-task display.

During the interruption lag, the cursor was hidden and disabled, so that all physical interaction with the primary task ceased. After an interruption, the primary-task display was reinstated in the same form it was in at the moment the eyeball alert appeared, with the following exceptions: The window that was active then — that is, the window that the participant was working in at the moment the alert appeared — was de-activated at task resumption, and the cursor was moved to the top-left corner of the screen. The effect was to eliminate the active window and the mouse cursor as

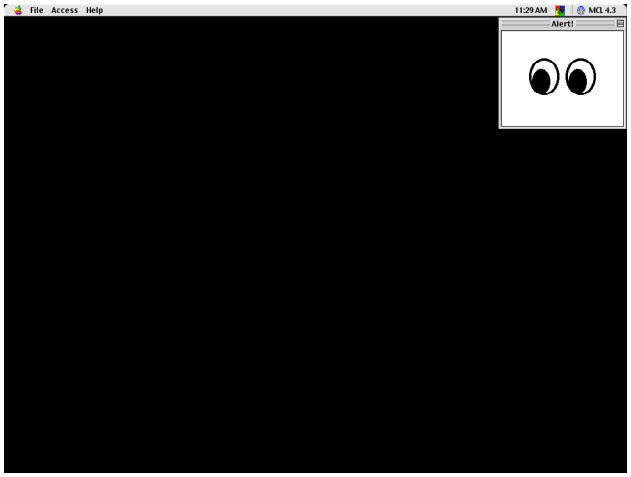


Figure 3: Screen shot of the primary task display during the interruption lag in the no-cue condition. The "eyeball alert" in the top right corner onsets at the start of the interruption lag and remains until start of the secondary task. In the cue condition the eyeball alert is identical but the primary task display (as in Figure 2) is preserved during the interruption lag.

retrieval cues that participants could deploy strategically to remind themselves of what they had been doing before the interruption, and therefore make resumption lag more sensitive to our experimental manipulations.

The secondary task involved evaluating "tracks," or targets on a radar screen, as friendly or hostile, according to attributes like speed and shape of icon (Trafton et al., 2003). A screenshot of the secondary-task display appears in Figure 4. Each instance of the secondary task lasted 30 to 45 seconds; afterwards, the participant was returned directly back to the primary task, as described above.

Design Interruption lag varied between experiments, as described above. Within each experiment, the independent variable was cue availability during the interruption lag (cue, no-cue), which was manipulated between subjects. The main dependent variable was resumption lag, the time from the end of the secondary task to the first subsequent action (mouse click) in the primary task. The other measure of interest, for comparison with resumption lag, was the inter-action interval, the mean time between actions in the primary task. Reported values for these measures are means of participant medians. For the inter-action interval, values below 1 s were discarded first, to eliminate anticipation errors, as well as ballistic components of motor plans, such as the second click of a double-click action.

Procedure Participants were tested individually, in sessions lasting roughly 90 minutes. A session began with a training period, in which participants learned to perform both tasks separately and were shown an example of how the computer would switch them from one task to the other and back again. After training, there were three 20-minute blocks of actual task performance. Within each block there were 10 interruptions, each triggered by a mouse click selected randomly to occur within a time window with quasi-random boundaries, to make interruption timing difficult for participants to predict.

At no point was the hypothetical function of the interruption lag in facilitating resumption mentioned to participants; thus, any effects of cue availability can be attributed to spontaneous use of preparatory strategies.

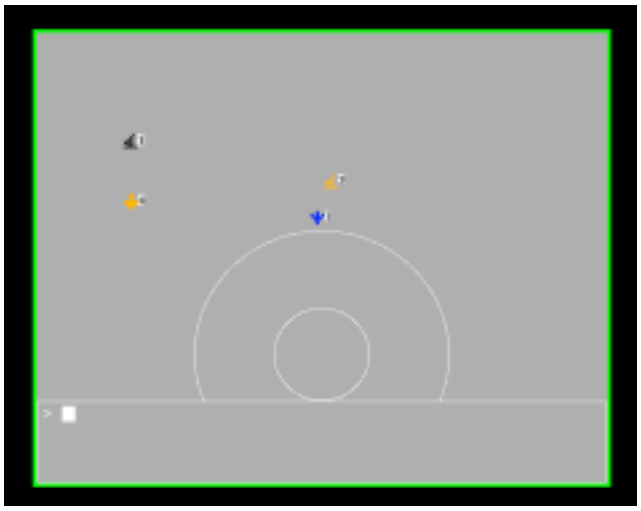


Figure 4: Screen shot of the secondary task display. Participants classified objects moving across a simulated radar display, according to color, shape, and speed.

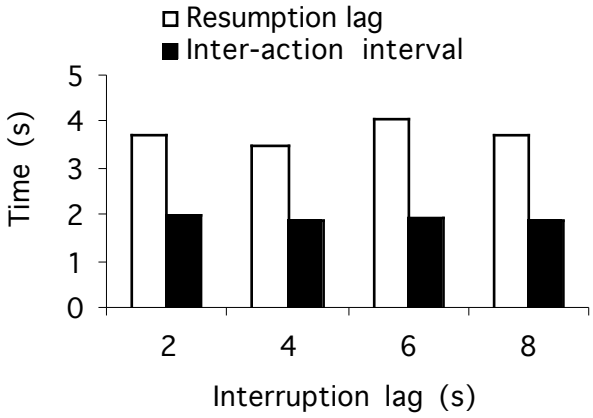


Figure 5: Resumption lag compared to inter-action interval, across experiments. The difference is an estimate of the disruption caused by task interruption.

Results We conducted two analyses of variance (ANOVAs) for each experiment. The first ANOVA, for which the data appear in Figure 5, compared mean resumption lag to mean inter-action interval; for all four experiments the difference was highly reliable, ps < .0001. The second ANOVA, for which the data appear in Figure 6, compared cue resumption lags to no-cue resumption lags. For 2- and 4-second interruption lags, there was no effect of cue availability, Fs<1. For the 6-second interruption lag, the cue-availability effect was marginal, $F(1,22)=4.1$, $p=.056$. For the 8-second interruption lag, the cue-availability effect was reliable, $F(1,22)=5.7$, $p<.03$.

We also conducted an omnibus ANOVA to compare across experiments, with cue availability and interruption lag as factors. The cue-availability effect was marginal, $F(1,88)=3.6$, p=.060, the interruption-lag effect was not reliable, $F(3,88)=1.2$, $p>.30$, and the two factors did not interact, $F(3,88)=1.4$, $p>.25$.

Discussion

The first empirical finding was that resumption lag is substantially longer than the mean interval between actions (Figure 5). This affords one measure of the disruptive effect of interruptions, at least in this highly interactive task in which, without interruption, actions occur at a rapid pace: The first action after an interruption took longer to execute than the first action after another primary-task action. In absolute terms, the resumption lag was 3.8 s – double the 1.9 s inter-action interval, which was measured rather conservatively by excluding all inter-action intervals under 1 s. This difference would appear to be of considerable practical interest in dynamic task environments, for example involving real airplanes or even automobiles traveling at highway speed, in which the world can look substantially different after an additional few seconds have elapsed.

The disruptive effect of interruption, as illustrated in Figure 5, was large and robust, which may agree with our intuitions about interruptions but doesn't necessarily agree

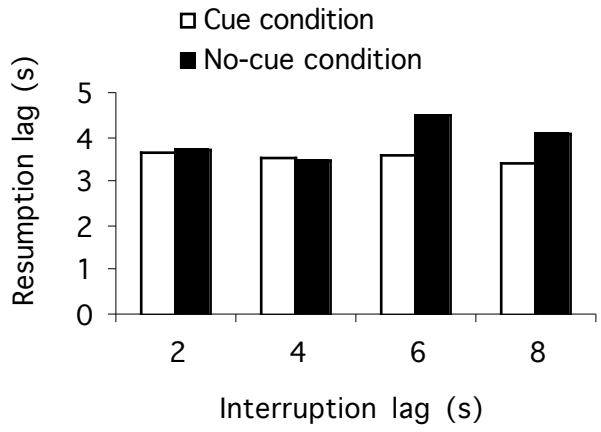


Figure 6: Resumption lag for cue and no-cue conditions, across experiments. The difference is is marginally reliable with a 6-second interruption lag and reliable with an 8-second interruption lag (see text).

with other controlled studies (e.g., Speier et al., 2003; Zijlstra et al., 1999). Three factors may have contributed to the robust effect measured here in terms of resumption lag. First, resumption lag is a local measure, taken immediately after every interruption; in contrast, other studies have reported global measures, such as overall time on task (Gillie & Broadbent, 1989; Speier et al., 2003). In our study, one global measure is a participants' total score over a session, but this was highly variable and showed no interpretable trends (so we did not report it). One specific problem with global measures is that they allow for compensatory strategies to work against the disruptive effects of interruption. Zijlstra et al. (1999), for example, speculate that their administrative workers compensated for interruption by using the time in between interruptions more efficiently.

A second factor may have been the relatively substantial cognitive state required to perform our primary task. In many scenarios in this task, beyond the resource-allocation tradeoffs involved, it was a challenge simply to piece together missions that would actually succeed in defeating destinations. In other studies, the primary task was simpler (Speier et al., 2003) and may have been more automated (Zijlstra et al., 1999), and therefore placed a smaller premium on maintaining complex representations in working memory.

Finally, a third factor may have been our implementation decision to trigger interruptions using mouse clicks, rather than strictly on the basis of time passage. Our rationale was that motor actions are often selected and programmed with the intention of achieving specific goals, so we reasoned that action-triggered interruptions would be more likely to disrupt these goals, which are one critical element of cognitive state; even in mundane tasks it's not uncommon to have "Now what was I doing?" moments, and it may be that these are more effectively induced by linking interruptions to actions rather than leaving interruption timing entirely to chance. Indeed, in Zeigarnik's (1927/1938) classic study, the experimenter was charged with judging when the participant was engrossed, in order to interrupt with the greatest impact.

The second empirical finding was that cue availability during the interruption lag (*before* the interruption) affected performance at task resumption (*after* the interruption, 30 to 45 s later), at least for longer interruption lags (Figure 6). One interpretation of this result, consistent with our memory analysis earlier, is that the various cognitive operations required to locate and encode retrieval cues during the interruption lag take somewhere between 6 and 8 seconds to complete (in our task environment). In other words, longer interruption lags afford enough time to link cognitive representations to external cues to facilitate retrieval later. However, this interpretation would also seem to predict that resumption lag in the cue condition should decrease at longer interruption lags, because cues are facilitating resumption. Instead, though, Figure 6 suggests that longer interruption lags drove an *increase* in resumption lag in the no-cue condition. Cue availability and interruption lag did not interact in the cross-experiment ANOVA, so the increase in no-cue resumption lags could be spurious, but given the exploratory nature of this work it seems useful to consider alternative accounts of why the cue-availability effect was limited to longer interruption lags.

One possible explanation of the increase in no-cue resumption lags might implicate changes in alertness or arousal – participants might simply have gotten bored, staring at a blank screen for 6 or 8 seconds. Some studies suggest that task interruption serves to increase arousal and stress, and thus improve overall (globally-measured) performance, at least on simple tasks (Speier et al., 2003); perhaps a long interruption lag, without visual information to focus on, moderates this effect. However, if arousal were to play a role in the cue-availability effect, it would remain to explain how a change in arousal *before* the interruption could affect performance *after* the interruption, tens of seconds later. Perhaps a drop in arousal caused participants' minds to wander in a way that activated irrelevant thoughts that in turn interfered with relevant cognitive representations. In such an account, however, memory would again play a central role in mediating the effect of pre-interruption variables on post-interruption performance.

One could explain the cue-availability effect without reference to memory processes if changes in arousal during the interruption lag persisted across the entire length of the interruption, to influence performance directly at task resumption. Secondary task performance was basically at ceiling for all subjects, so offers little evidence on this possibility. However, if arousal effects were to persist for the entire 30 to 45 seconds of the interruption, one might expect them to persist somewhat beyond as well, and then only gradually dissipate. This would predict that the time between the first and second action after the interruption would also reflect the cue availability effect. Revisiting our data, however, we found no difference, as a function either of interruption lag or cue availability, in the duration of the interval between the first and second actions after an interruption; this measure appears in Figure 7. It seems most likely, then, that even if cue availability and interruption lag interact to affect arousal before an interruption, memory

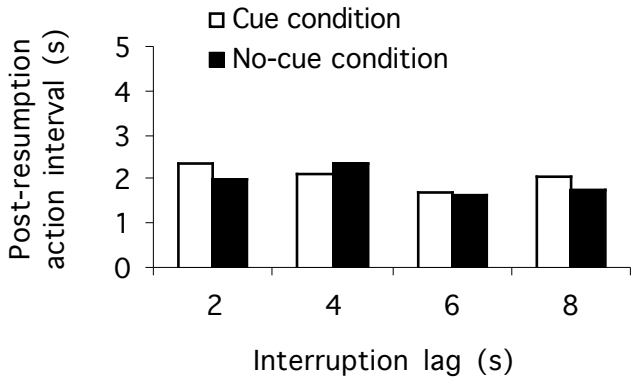


Figure 7: Inter-action interval between the first and second actions after an interruption, for cue and no-cue conditions, across experiments.

and/or perceptual processes mediate the delayed effect on task resumption.

At least two avenues of future work seem indicated to clarify the effects of our interruption-lag manipulations on speed of task resumption. First, it will be important to repeat these manipulations in context of a factorial design in which interruption lag and cue availability are fully crossed; here, one objection is that the cross-experiment comparison is potentially confounded by changes in the subject population.

Second, although the model that motivated these experiments emphasizes memory processes (Altmann & Trafton, 2002), there are alternative characterizations of why task resumption is time consuming. In particular, one account of automation deficit (Ballas, Kieras, Meyer, Brock, & Stroup, 1999) — like resumption lag, but measured in terms of accuracy — is that it reflects encoding of perceptual information (Kieras & Meyer, 1997) rather than memory retrieval. Thus, in our task environment it could be that the difference between the resumption lag and the baseline inter-action interval (Figure 5) simply reflects the cost of re-encoding the display, and that this re-encoding is what is facilitated by cue availability during the interruption lag. To distinguish between these accounts, one could vary the extent of the cognitive representations required to perform the primary task on one hand, and the perceptual complexity of the display on the other. Under a memory-retrieval model, the cue-availability effect should be linked to complex cognitive states, whereas under a perceptual-encoding model the effect should be linked to complex external displays; in our task environment, these two factors are confounded.

Whatever the ultimate explanation, the cue-availability effect shows an interesting link between what happens before an interruption and what happens later, after tens of seconds of intervening behavior. In practical terms, the effect suggests that interface designs, and possibly training interventions, could exploit cue availability in some way to facilitate resumption in task environments in which interruptions are frequent and seconds matter. In theoretical terms, probing this effect should help us develop constraints on models of memory, perception, and cognitive control as these functions are deployed in complex dynamic task environments.

Acknowledgements

This research was supported by grants N00014-03-1-0063 (to EMA) and N0001400WX21058 (to JGT) from the US Office of Naval Research. Thanks to the conference reviewers for helpful suggestions for improvement.

References

Altmann, E. M. (in press). Advance preparation in task switching: What work is being done? *Psychological Science*.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science, 26*, 39-83.

Ballas, J. A., Kieras, D. E., Meyer, D. E., Brock, D. P., & Stroup, J. (1999). *Cueing of display objects by 3-D audio to reduce automation deficit*. Paper presented at the Fourth annual symposium on situational awareness in the tactical environment, Patuxent River, MD.

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research, 50*, 243-250.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*, 391-438.

Latorella, K. A. (1996). *Investigating interruptions: Implications for flightdeck performance*. PhD thesis, State University of New York, Buffalo, NY.

McFarlane, D. C, & Latorella, K. A. (2002). The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction, 17*, 1-61.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*, 134-140.

NTSB. (1988). *Aircraft accident report: NWA DC-9-82 N312RC, Detroit Metro, 16 August 1987* (No. NTSB/AAR-88/05). Washington, DC: National Transportation Safety Board.

Speier, C., Vessey, I., & Valacich, J. S. (2003). The effects of interruptions, task complexity, and information presentation on computer-supported decision-making performance. *Decision Sciences, 34*, 771-797.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies, 58*, 583-603.

Zeigarnik, B. (1927/1938). On finished and unfinished tasks. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 300-314). New York: Harcourt Brace.

Zijlstra, F. R. H., Roe, R. A., Leonora, A. B., & Krediet, I. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology, 72*, 164-185.

Linguistic diversity and the bilingual lexicon: The Belgian Story

Eef Ameel (eef.ameel@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Gert Storms (gert.storms@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Barbara Malt (bcm0@lehigh.edu)
Department of Psychology, Lehigh University, 17 Memorial Drive East
Bethlehem, PA 18015 USA

Steven Sloman (Steven_Sloman@brown.edu)
Cognitive & Linguistic Sciences, Brown University, Box 1978

Providence, RI 02912 USA

Abstract

Analogous to Malt, Sloman, Gennari, Shi and Wang (1999) we examined the relation between linguistic categorization and similarity of artifacts by Dutch-speaking and French-speaking monolingual Belgians. We replicated the dissociation between naming and sorting found by Malt et al. (1999) for speakers of English, Chinese and Spanish. We also investigated the relation between the two naming patterns of bilingual Belgians, raised simultaneously in French and in Dutch, and how these naming patterns can be linked to the naming of the monolinguals. The results showed that the French and Dutch naming pattern of the bilinguals didn't parallel the respective naming patterns of the monolinguals, but rather merged into a common naming pattern.

Introduction

Research from several different traditions concerns the coupling of similarity and naming. However, different studies have resulted in contradicting conclusions. Some studies found that categorization judgments paralleled similarity judgments, for example the study of Smith and Sloman (1994). Other studies have shown a clear dissociation between similarity judgments and preferred category labels for novel objects. Keil (1989) and Rips (1989) presented participants with artifacts described as physically resembling one type of object, but having been made to be used as another type, or with animals looking like one type of animal but said to have internal parts of a different species. They both found that although objects were rated as more similar to the former, they tended to be categorized as the latter (see also Rips & Collins, 1993). Also studies that look at well-established lexical categories and make comparisons across speakers of different languages find substantial differences in naming objects, but only small differences in perceived similarity among the objects. For example, Kronenfeld, Armstrong and Wilmoth (1985) looked at the names given to various drinking vessels and the similarity among them judged by American, Japanese and Israeli participants. They found a dissociation between naming and similarity. However, the sample of objects used by Kronenfeld et al. (1985) was small and they did not attempt to assess whether the observed differences in naming paralleled the differences in perceived similarity.

Malt et al. (1999) carried out a larger-scale evaluation of the relation of perceived similarity among objects to the names they are given. They presented data from speakers of three different language groups: American, Chinese and Argentinean participants, speaking respectively English, Chinese and Spanish. The participants performed two tasks: they named 60 common containers (all mostly called 'bottle' or 'jar' in English) and they provided similarity ratings, by sorting the objects into piles based on three types of similarity: physical, functional or overall similarity. Speakers of the three languages showed substantially different patterns of naming for the set of containers, but they saw the similarities among the objects in much the same way. Malt et al. claim that the linguistic differences arise from differences in language-specific conventions and differences in language history.

The imperfect relation between naming patterns of a language and non-linguistic knowledge of objects and between the naming patterns of two different languages raises questions about how bilingually-raised individuals build and maintain their two lexicons. Do they maintain two distinct and native-like naming patterns, each with its own language-specific conventions or do the two competing

patterns merge into a single pattern that may not be fully native-like for either language. The latter might be due to individual cognitive constraints on memory capacity. One way to address this issue is to examine the two naming patterns of bilinguals to see how they relate to one another and to the naming patterns of corresponding monolinguals. Belgium, a bilingual country where French- and Dutch-speaking monolinguals live alongside bilinguals, who are brought up simultaneously in French and Dutch, live together, provides us with a laboratory to investigate this issue.

Two hypotheses are suggested concerning the naming patterns of bilingual Belgians: First, the French and Dutch naming patterns are kept separate and thus parallel the naming patterns of respectively the French-speaking monolinguals and the Dutch-speaking monolinguals. Second, the two naming patterns merge into one naming pattern and the bilinguals use just one single naming pattern both for the French and the Dutch naming.

Method

Participants

Thirty-two native speakers of Dutch, all students or research assistants at the Psychology Department of the Leuven University, and 29 native speakers of French, students at the Faculty of Law of the University of Liège, participated both in a naming and a sorting task (to be described below). Five participants of the Dutch-speaking group were retested for the naming, to check for within subject reliability. The time span between the test and the retest was approximately six months. The bilingual subjects consisted of 25 people whose father is Dutch-speaking and whose mother is French-speaking (14 out of 25) or vice versa (11 out of 25) and who have been raised in both languages. All of them were students (except one research assistant) at the university of Leuven, Brussels or Louvain-la-Neuve. The bilingual subjects performed the naming task twice (once in French and once in Dutch) and the sorting task once. They also completed a language history questionnaire, used to determine the participants' language background. Five bilinguals renamed the objects in French and five other bilinguals renamed the objects in Dutch after a time span of about six months.

The Dutch- and French-speaking monolingual subjects received course credit or participated as unpaid volunteers. The bilinguals were systematically paid for their participation.

Materials

Objects. There were 2 sets of stimuli, one consisting of 73 pictures of storage containers (as in Malt et al.'s study (1999)), the other consisting of 67 pictures of housewares for preparing food and serving food and drink. The objects of the first set were selected to be likely to receive the name 'bottle' or 'jar' in American English, or else to share one or more salient properties with bottles and jars. Translated into Dutch and French, the objects are likely to be called respectively 'fles' or 'bus' and 'bouteille' or 'flacon'.[1] For the second set, the 'dishes set', objects had been selected to be likely to be called 'dish', 'plate' or 'bowl' in American English. In Dutch, the objects are mostly called 'bord', 'schaal' or 'kom', in French 'assiette', 'plat' or 'bol'.[2]

The objects were all found at home, work, or in stores frequented by the researchers. For both sets, we made an effort to include objects that would represent a wide range of respectively bottles, jars and other similar containers (Set 1) and of dishes, plates, bowls and other similar housewares (Set 2). The wide range of objects allows a sensitive comparison of the naming patterns of the Dutch-speaking monolinguals, the French-speaking monolinguals and the bilinguals.

All objects were photographed in color against a neutral background with a constant camera distance to preserve relative size. In front of each object a ruler was included to provide additional size information. Because the labels on the objects were mostly both in Dutch and in French, no additional information about the nature of the content (e.g. ketchup) was necessary.

Language history questionnaire. A questionnaire was used to determine the language background of the bilingual participants. Questions were asked about age and sex; where the participant was raised; what language her mother and father speaks; what language she speaks with her mother and father and whether she systematically speaks the same language (Dutch or French) with her mother and the same other language (French or Dutch) with her father; what language was used at primary and secondary school, during leasure activities; which language she currently uses most and estimated proficiency for both languages. Proficiency estimates were obtained by asking the participants for each language to encircle a number between 1 ('not at all fluent: you can barely speak the language') and 7 ('very fluent: you can speak the language like a native speaker').

Procedure

Naming task. In the naming task, participants were asked to name each object of two sets of pictures (the bottles and dishes sets), after looking through all the pictures of the set to be named to familiarize themselves with the variety of objects in the set. The instructions were the same as in the naming task of Malt et al. (1999): They were asked to give whatever name seemed like the best or most natural name, and they were told that they could give either a single-word name or a name with more than one word. The instructions emphasized that participants should name the object itself

[1] It should be noted that we do not claim 'jar', 'bus' and 'flacon' to be translation equivalents or to cover the same group of referents. Referring to a dictionary, 'bus' is translated as 'can', 'flacon' as 'bottle'. 'Fles' and 'bouteille' are however translated unambiguously as 'bottle'.

[2] As for the bottles set, the corresponding names ('dish', 'bord' and 'assiette'; 'plate', 'schaal' and 'plat'; 'bowl', 'kom' and 'bol') 'are not assumed to be perfect translation equivalents.

and not what it contained. Each participant named first all the objects of one set (bottles set or dishes set) and then all the objects of the other set (dishes set or bottles set). The order of the two sets was counterbalanced. The bilingual participants named both sets in French and in Dutch. Hence, besides the order of sets also the order of languages was counterbalanced. Between the Dutch and the French version of the naming task, the pictures were shuffled.

After participants completed the naming task (once for the Dutch- and French-speaking participants, twice for the bilinguals), the pictures were again shuffled. The second task to be performed was the sorting task.

Sorting task. The large number of objects prevented us from collecting direct pairwise similarity judgments. Instead we asked the participants to sort the objects into piles. Based on these sorting data, we can calculate a derived measure of similarity for each pair of objects. Sorting was based on overall similarity. First, participants were asked to look through the pictures. The instructions for the sorting were as follows: 'I would like you to focus on the overall qualities of each container. This means that you focus on any feature of the container including what it looks like, what it's made of, how it contains the substance that is in it (in a stack, in separate pieces, as a single solid, as a liquid, with pouring capability, etc.[3]) or any other aspect of the container that seems important or natural to you. I would like you to put together into piles all the containers that you think are very similar to each other OVERALL. Note that we are interested in how similar the containers themselves are overall, not what is in the containers. Only put two pictures together if the containers are like each other in an overall way. DO NOT put pictures together just because the containers hold things that tend to be found together. For instance, if several containers contain health products, DON'T put them together unless you really think the containers themselves are alike in an overall way.'

Further, the participants were instructed to use as many piles as they wanted, but at least two different ones. They were not allowed to make a pile of only one picture, unless they really could not classify the object in one of the existing piles. They could take as much time as they wanted to complete the sort. In general, the sorting task took about 30 minutes.

Due to space restrictions, we will focus on the results of the bottles set only. However, the results with the dishes set were perfectly parallel to those of the bottles set.

Results

Replication of Malt et al.'s study

Comparison of linguistic category boundaries. For each name produced for each object, we first calculated its frequency separately for each language group. Only the head noun of the response was considered as the name given to the object. Diminutive forms of names and additional adjectives were disregarded. The first analysis is restricted to the dominant category names for each object: i.e. the most frequently produced name for each object.

Table 1 shows the Dutch and French dominant category names for the bottles set together with the number of objects out of 73 for which each name was dominant. To gain an insight into the similarities and differences between the Dutch and French categories, the French categories are described in terms of their Dutch composition.

Table 1: Linguistic categories for the bottles set of the Dutch- and French-speaking monolinguals.

French bottles (monolinguals)	**N**	**Dutch Composition (monolinguals)**
bouteille	16	13 flessen, 3 bussen
flacon	16	10 flessen, 3 bussen, 2 potten, 1 roller
pot	10	9 potten, 1 fles
boîte	7	3 dozen, 2 brikken, 1 blik, 1 pot
tube	6	4 tubes, 1 pot, 1 stick
spray	5	5 bussen, 1 spray
bidon	3	3 bussen
brique	2	1 bus, 1 doos
berlingo	2	2 brikken
biberon	1	1 fles
bombe	1	1 bus
canette	1	1 blik
pannier	1	1 mand
poivrier	1	1 molen
salière	1	1 vat

For Dutch-speaking monolinguals, there were three main categories: 'fles', 'bus' and 'pot'[4]. The three categories together encompassed 74 per cent of the stimulus set. The remaining names were given to only a few objects. The French-speaking monolinguals used a total of 15 categories. Three category names were dominant for at least 10 objects out of 73: 'bouteille', 'flacon', 'pot'[5]. The other names were restricted to a smaller number of objects.

When we look at the Dutch composition of the French categories, we find some resemblance in how the two languages classify the objects into linguistic categories: the largest part of the objects called 'pot' in Dutch (9/13) are put into one single French category 'pot'. All Dutch 'tubes' are put together into the French category of objects called 'tube'. On the other hand, there are also prominent differences between the naming patterns of both languages: The objects called 'fles' (# 25) in Dutch are mainly split up

[3] This information is only provided for the sorting of the bottles, since it is not applicable to the dishes.

[4] 'Fles' is translated as 'bottle', 'bus' as 'can' and 'pot' as 'pot' or 'jar'.

[5] 'Bouteille' is translated as 'bottle', 'flacon' as well, 'pot' as 'pot' or 'jar'.

into two different categories in French: 'bouteille' (# 13) and 'flacon' (# 10). The Dutch category 'bus' is not represented in a corresponding French category, but the objects are spread over 6 different categories ('bouteille', 'flacon', 'spray', 'bidon', 'brique' and 'bombe').

By looking at the dominant names, a lot of information in the data is lost. For the bottles set, only 5 objects were called by the same name by every Dutch monolingual and the same was true for the French monolinguals. Hence, it would be more useful to include all the names used for each object in the analysis. Therefore, in a second analysis, we calculated for each object the name distribution which can be described as the number of times each name was allocated to each object. Our intention was to compare the linguistic categories of the different language groups by comparing the naming distributions across the language groups. However, the naming distributions cannot be compared directly across the language groups since different language groups use different sets of names ($Dutch_{monolingual}$ versus $French_{monolingual}$). As an alternative, for each language group we compared the similarity of each object's name distribution to every other object's name distribution by using a Pearson correlation. The similarity in name distribution between two objects was calculated as follows: for each pair of objects within a language group, the correlation was computed across all the names between the name frequencies for both objects. For each language group, this resulted in 2628 correlations (for 73*72/2 pairs of objects). These correlations indicated the name distribution similarity between each possible pairing of the objects. The next step consists in correlating the 2628 name similarity values for the Dutch-speaking monolinguals with the corresponding 2628 name similarity values for the French-speaking monolinguals. This correlation mirrors the extent to which the two language groups correspond in the pairs of objects that have similar name distributions. The correlation between both monolingual language groups is 0.63: a substantial correlation, but far from perfect. Both the analysis of the dominant names and the correlation between the name distribution similarities confirm that the French- and Dutch-speaking monolinguals named the objects differently.

Comparison of the perceived similarity. The data from the sorting task were used to obtain a measure of similarity for each pair of objects. Pairwise similarity was recovered by counting for each pair of objects how many participants of a language group placed that pair of objects in the same pile. For each of both language groups, these calculations gave us 2628 pairwise similarity judgments. The similarity judgments of both groups were correlated. The resulting correlation of 0.87 -comparable to the mean estimated reliability of .92- indicates that the French- and Dutch-speaking monolinguals agree to a considerable extent on their perception of similarities among the objects.

Conclusion. For the two monolingual language groups, we found substantial differences in naming and no differences in sorting. These results replicate the findings of Malt et al. (1999) for speakers of three different languages.

Naming in bilinguals: Two hypotheses

How did the French and Dutch naming patterns of the bilinguals interrelate and how are they linked to the naming patterns of the respective monolingual language groups? One possibility is that the naming of the bilinguals follows that of the corresponding monolinguals, i.e. the French bilingual naming pattern equals the naming pattern of the French-speaking monolinguals and the Dutch bilingual naming pattern equals that of the Dutch-speaking monolinguals. Another alternative possibility is that the bilinguals use just one naming pattern, or in other words, that their naming patterns of their two languages converge into a single naming pattern. To decide between these hypotheses, we analyzed the data both on a group level and on an individual level.

Group-level analysis. Correlations were calculated among all the language groups ($Dutch_{monolingual}$, $Dutch_{bilingual}$, $French_{monolingual}$, $French_{bilingual}$) between measures of name similarity (i.e. name distribution similarities). Figure 1 shows the pattern of correlations.

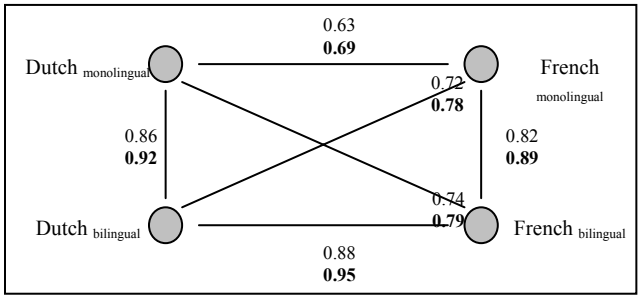


Figure 1: Pattern of correlations between the name distribution similarities of the language groups.[6]

When we compare the observed pattern of correlations with the patterns of correlations predicted by the two hypotheses (see Figure 2), we can conclude that the data are inconsistent with the two-pattern-hypothesis, since the correlation between the two naming patterns of the bilinguals (0.88) was significantly larger than the correlation between the naming patterns of both monolingual language groups (0.63), $Z = 21.82 > 1.96$. The data favor the one-pattern-hypothesis. Note however that some deviations from a single common naming pattern were observed. For example, it happens that a group of objects, with a single

[6] The upper r_{XY}'s are the Pearson correlations, the lower $\mathbf{r_{XY}}$'s (in bold) are correlations corrected for unreliability of the data ($r^*_{XY} = \frac{r_{XY}}{\sqrt{r_{XX}} * \sqrt{r_{YY}}}$ with r_{XX} the reliability of X and r_{YY} the reliability of Y).

name in Dutch ('fles') is subdivided into more than one category in French (e.g., 'bouteille' and 'flacon').[7]

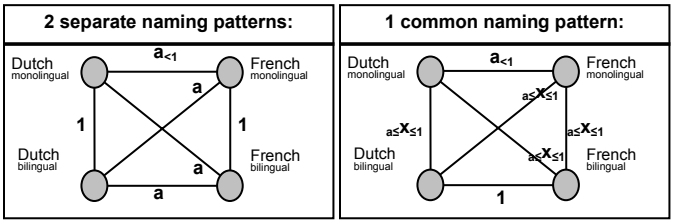


Figure 2: Patterns of correlations between the name distribution similarities of the language groups, predicted by the first and second hypothesis, respectively.

Individual-level analysis. On the individual level, object*object-matrices for each individual task, containing 0's and 1's with 0 indicating equal naming of both objects by the person performing the task and 1 indicating different naming of both objects- were correlated with each other. This resulted in 126*125/2[8] different correlations between all possible pairs of individual tasks. Next, the correlations were Z'-transformed to normalize the sampling distribution of the correlations. Then, the Z'-transformations of the correlations were analyzed in a randomized block factorial ANOVA design, with three factors: language (two levels: the subjects of the pair perform the naming task in the same language or in a different language), person (two levels: correlation between naming data of the same subject or of different subjects) and linguistic statute (three levels: both subjects are monolingual, one subject is monolingual, the other bilingual and both subjects are bilingual), resulting in a 2*2*3 design with unequal cell frequencies and three (structurally) empty cells (see Figure 3).

The results of the ANOVA confirmed the conclusions that were derived from the correlational group-level analysis. The three main effects –language, person, linguistic statute- were all significant, respectively $F(1,7866) = 23.29$, $p < .0001$, $F(1,7866) = 42.61$, $p < .0001$, $F(2,7866) = 8.15$, $p < .0005$) as was the interaction effect language*person*linguistic statute, $F(4,7866) = 25.05$, $p < .0001$. We tested the following crucial contrasts: μ_{221} versus μ_{223} (C1) and μ_{113} versus μ_{213} (C2). If C1 is significant, Hypothesis 1 is rejected, since according to the two-pattern hypothesis, the mean correlation between the naming of a French-speaking monolingual and the naming of a Dutch-speaking monolingual must be equal to the mean correlation between the French naming of a bilingual and the Dutch naming of a(-nother) bilingual. If C2 is significant, Hypothesis 2 is rejected, because the one-pattern hypothesis claims that the Dutch and French naming task of a (same) bilingual correspond equally well as the naming of that bilingual and the renaming (retesting) of the same bilingual in the same language. We found that C1 was significant, $F(1,7866) = 299.94$, $p < .0001$, and hence Hypothesis 1 is rejected. C2 was not significant, which means that Hypothesis 2 is retained.

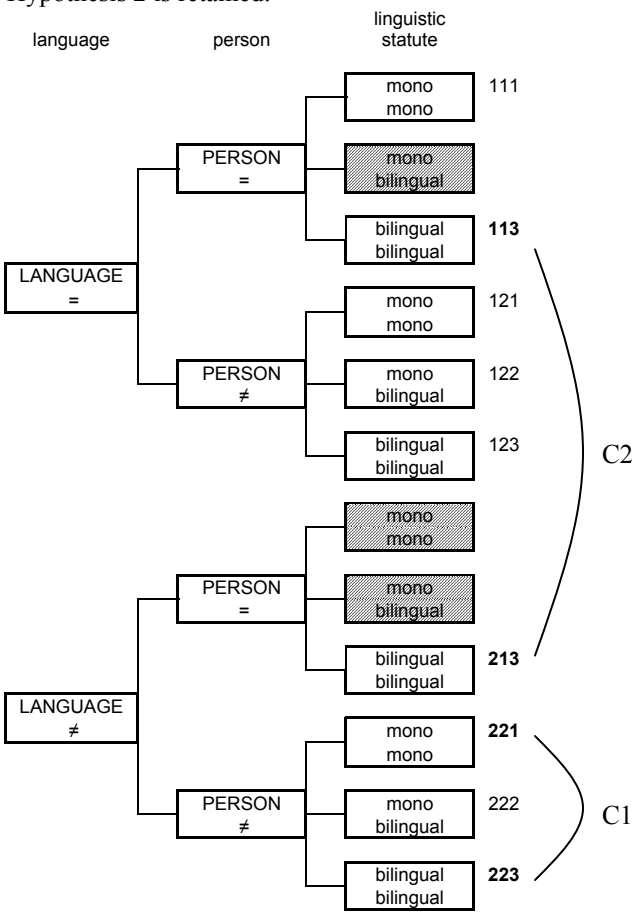


Figure 3: 2*2*3-factorial design with unequal cell frequencies and three empty cells.

General Discussion

The dissociation between naming and sorting, found by Malt et al. (1999) for three different language groups was replicated for the French-speaking and Dutch-speaking monolinguals: The analyses of the dominant names and of similarities among naming distributions revealed substantial differences in French-speaking and Dutch-speaking monolingual linguistic categories for the bottles set, while in contrast, no differences were found in their perceptions of the similarity among the objects, as revealed by the high correlation between the sorting data of both monolingual language groups. Hence, similarity cannot fully account for the observed naming patterns. Other factors must contribute to linguistic categorization. Malt et al. (1999) proposed that besides the contribution of similarity to naming choices,

[7] Remark that this kind of subdivisions occurs much more frequently between the monolingual naming patterns than between the bilingual naming patterns.

[8] 126 individual tasks: = 32 Dutch-speaking monolinguals + 5 retested Dutch-speaking monolinguals + 29 French-speaking monolinguals + 25 bilinguals (Dutch naming) + 5 retested bilinguals (Dutch) + 25 bilinguals (French naming) + 5 retested bilinguals (French).

mechanisms such as chaining, convention and pre-emption influence naming patterns.

Concerning the bilingual naming patterns, the data (at group and individual levels) reject the two-pattern hypothesis. So, we can conclude that the French and Dutch naming patterns of the bilinguals are not kept separate and hence don't parallel the naming patterns of the French and Dutch monolinguals, respectively. The data are more consistent with the one-pattern hypothesis, suggesting that the two naming patterns of the bilinguals merge into one. However, the data also show that the assumption of a perfect match between the naming patterns is too strong and that is should be attenuated, since bilinguals did not use the French and Dutch category names as perfect translation equivalents. Apparently, even if the two naming patterns of bilinguals deviate from the corresponding monolingual naming patterns, naming in each of both languages is still influenced by culture- and language-specific factors: bilinguals name the objects in a way that is consistent with the language in which they name the object. This is not so surprising, since language and culture can't be considered separately. On the other hand, the convergence of the two naming patterns on one naming pattern suggests that bilinguals do not only satisfy cultural and linguistic constraints, but also individual cognitive constraints: it is less demanding on the limited sources of memory to store only one set of mappings between objects and names. So, in a way, bilinguals do find a set of mappings between words and objects that meet linguistic, cultural and individual memory constraints.

References

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kronenfeld, D. B., Armstrong, J. D., & Wilmoth, S. (1985). Exploring the internal structure of linguistic categories: An extensionist semantic view. In J. W. D. Dougherty (Ed.). *Directions in cognitive anthropology* (pp. 91-113). Urbana: University of Illinois Press.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic Categorization of Artifacts. *Journal of Memory and Language,* **40**, 230-262.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.). *Similarity and analogical reasoning.* New York: Cambridge University Press.

Rips, L. J., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General,* **122**, 468-496.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus Rule-based categorization. *Memory and Cognition,* **22**, 377-386.

Cross-linguistic Semantic Differences Influence Recognition of Pictures

Florencia Anggoro (f-anggoro@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Rd., Evanston, IL 60208 USA

Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Rd., Evanston, IL 60208 USA

Abstract

We compared recognition memory for pictures of family interactions in Indonesian – in which sibling terms are based on *relative age* – and English – in which sibling terms are based on *gender*. In Experiment 1, participants saw a set of pictures of family interactions and gave a verbal description of each picture. They later received a recognition test that included variants that altered either seniority relations or gender relations. The same method was used in Experiment 2, except that the recognition variants were changed to be *similar* families (with parallel relationships). During study, participants were asked either simply to remember the pictures (Experiment 2a) or to provide a verbal description (Experiment 2b). Results from both experiments suggest effects of language on memory, particularly when non-identical transfer is involved.

Introduction

The Whorfian Question

Does the language we speak influence the way we think? This question, out of favor for many years, has had a resurgence of interest (Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996). There is evidence (Bowerman & Choi, 2003; Sera et al., 2002; Choi, McDonough, Bowerman, & Mandler, 1999; Levinson, 1998) that linguistic distinctions may influence non-linguistic similarity and memory for scenes (but see Munnich, Landau, & Dosher, 2001; Li & Gleitman, 2002 for contrary evidence).

The recent investigations have mostly centered around perceptual arenas such as space, time and motion. However, classic studies in anthropological linguistics suggest that there are also substantial differences in semantic categories in social arenas such as kinship (Romney & D'Andrade, 1964; Danziger, 2001; Foley, 1997). It is important to test whether these linguistic differences have cognitive consequences. There are direct studies of the cognitive effects of social semantics. Boroditsky and Schmidt (2000) found effects of linguistic gender on people's encodings of objects. For example, they taught Spanish-English and German-English bilinguals English names for objects (such as "Mary" for a table) and found that people retained the names better when the gender was consistent with the gender of the noun in their first language. In addition, bilinguals' English descriptions of the objects were consistent with the gender in their first language. Sera, et al. (2002) have also shown that gender retains semantic context in that cross-linguistic differences influence classification (Sera et al., 2002).

Our work explores an arena of social categories, namely kinship terms. As Malinowski (1930) noted, some dimensions seem likely to be universal in kinship systems—such as the gender of the person named, the age and/or generation relative to ego, and the gender of the linking relative. Nevertheless, kinship systems vary considerably in how these distinctions play out. Our study focuses on one pair of contrasting languages – English and Indonesian – which vary in the way they name sibling relations.

Indonesian makes a lexical distinction for whether a sibling is *older or younger*. The word *kakak* refers to older sibling while the word *adik* refers to younger sibling. For example,

(1)

Saya	mempunyai	seorang	kakak.
1st pers. sing.	have	one (person)	older sib

'I have an older sibling'

In contrast to English *brother* and *sister*, the Indonesian sibling terms *kakak* and *adik* are gender-neutral. When an Indonesian refers to his/her siblings, he/she speaks not in terms of sister and brother but rather of older and younger. Thus, "How's your older sibling (*kakak*)?" is as natural in Indonesian as "How's your brother?" is in English.

Of course, both languages can specify both gender and seniority if desired. An English speaker could refer to "your younger brother" and an Indonesian to "your male younger-sibling"

(2)

Adikmu	laki-laki
younger sib.-your	male

'your younger-sibling boy'

Thus, the Indonesian semantic system focuses on the relational seniority of siblings, whereas the English systems focus on gender. Our study investigates a) whether this linguistic difference matter to the way people think about family relations, and b) whether it affects the way people construe scenes involving families.

In our previous study (Anggoro & Gentner, 2003, Experiment 2) we used a recognition task to test whether the two languages induce different encodings. Indonesian and English speakers were shown a series of pictures: three kinship standards and their three corresponding family pictures, along with 21 other pictures (Figures 1 and 2). Participants were asked to remember the scenes for a later memory task. Recognition memory for the scenes was later tested using variants of the standard pictures. Memory for each standard was tested using two variants: the Seniority Variant, which preserved the seniority relation but altered the gender relation, and the Gender Variant, which preserved the gender relation but altered the seniority relation. There was a tendency for Indonesian speakers to make more false alarms to the Seniority Variants than to the Gender Variants, suggesting better memory for Seniority than for Gender. English speakers showed the reverse pattern. An ANOVA over language and variant type showed a marginally significant interaction between the two factors. Other results from the same set of studies also point to an influence of language on encoding and recognition. For example, relative to English speakers, Indonesian speakers showed greater sensitivity to changes in seniority than to changes in gender in a similarity task. These results suggest greater sensitivity to the dimension that is required in naming siblings in each language.

The Current Study

Slobin (1987) has suggested in his *thinking for speaking* hypothesis that "[when] constructing utterances in discourse, one fits one's thoughts into available linguistic forms." In our current work we seek to test whether verbally describing the pictures would strengthen the language effect. In addition, we further explored the effects of a more challenging task that involved nonidentical transfer.

Experiment 1

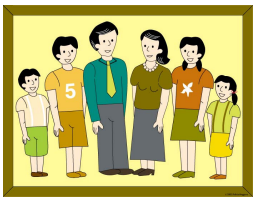
As in our previous study, Indonesian and English speakers were shown a series of pictures: the three kinship standards (as exemplified in the top picture in Figure 2) and their three corresponding family pictures (Figure 1), along with 21 other pictures (a total of 27 pictures). For each picture, participants were asked to describe the scenes and remember them for a later memory task. After a brief filler task, participants were given a recognition memory test. As in the previous study, the test included two variants: the Seniority Variant, which preserved the seniority relation but altered the gender relation, and the Gender Variant, which preserved the gender relation but altered the seniority relation

If describing the scenes leads participants' encodings to be influenced by the semantics of their kinship systems, then Indonesian monolinguals will be relatively more sensitive to changes in relative age than to changes in gender, as compared to English monolinguals. Specifically, if verbal description heightens the effects of semantic categories on encoding and recognition, then we should find a significant interaction between language and variant type.

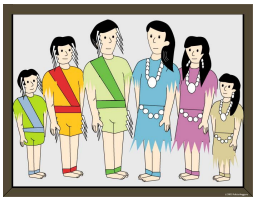
Method

Participants The participants were 15 Indonesian monolinguals and 13 English monolinguals, ranging in age from approximately 17 to 20 years old. Participants were either given credit or a small monetary compensation. Data from Indonesian speakers were collected in Jakarta, Indonesia. Data from English speakers were collected at Northwestern University and other areas near Chicago.

Materials The stimuli were three sets of pictures. One set (the Kitchen set) involved scenes of siblings performing a simple activity in the kitchen. The other two sets (the Ritual sets) involved ceremonies. Family pictures were used to introduce the 'characters' and make clear the sibling relations.



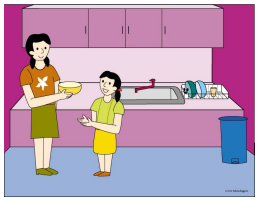
Family Picture for Kitchen Set



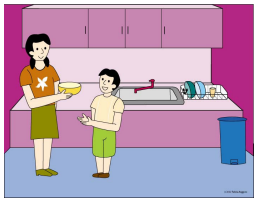
Family Picture for Ritual 1 Set

Figure 1: Family pictures used in the Kitchen and Ritual 1 sets.

The triad pictures consisted of one standard picture and two variants: the Seniority Variant, which preserved the seniority relation but altered the gender relation, and the Gender Variant, which preserved the gender relation but altered the seniority relation.



Standard
older sister gives bowl to younger sister



Seniority Variant
older sister gives bowl to younger brother

Gender Variant
younger sister gives bowl to older sister

Figure 2: The Kitchen set. In the Seniority Variant, the bowl still goes from the older to the younger sibling, but the gender of the younger sibling is altered. In the Gender Variant, the gender of both actors is the same as in the standard, but the bowl now goes from the younger to the older sibling.

Standard
younger brother gives crown to older sister

Gender Variant
older sister gives crown to younger brother

Seniority Variant
younger brother gives crown to older brother

Figure 3: Ritual 1 set. In the Gender Variant, the genders of both actors are the same as in the standard, but the crown now goes from the older to the younger sibling. In the Seniority Variant, the bowl still goes from the younger to the older sibling, but the gender of the older sibling is altered.

Procedure Participants were run individually in a quiet room. Instructions were given in Indonesian for the Indonesian speakers and English for the English speakers. For each set of stimuli, participants were first shown a family picture to ensure that they understood that the triad that followed only involved the children. (For the Ritual sets, the experimenter explained that the family lives on some island and they held a ritual each year. Then the standard was shown without further description.) For each standard, participants were asked to verbally describe what they saw in the picture. After participants had seen and described all of the standards, they were given a short break (approximately 10 minutes) during which they were asked to solve a few simple puzzles. Then they were given a yes-no recognition task. The two variants for each standard were intermixed in semi-random order among the fillers. The three standards that the participants had actually seen were given at the end of the test.

Results

As in the previous study, the key dependent measure is the mean proportion of times a participant responded 'yes' to each variant; i.e., the false alarm rates on the Gender Variants vs. the Seniority Variants. An ANOVA over Language and Variant Type showed a main effect of Variant Type ($F(1,26) = 7.21$, $p = .01$) such that participants made more seniority-preserving false alarms ($M = .29$) than gender-preserving ones ($M = .13$) and a main effect of language ($F(1,26) = 6.61$, $p = .02$), such that Indonesian speakers made more false alarms ($M = .29$) than English speakers ($M = .12$). As predicted, there was a significant interaction between the two factors ($F(1,26) = 4.90$, $p = .04$). Indonesian speakers made more false alarms to the Seniority Variants ($M = .42$, $SD = 30$) than to the Gender Variants ($M = .16$, $SD = .21$). English speakers showed no difference in false alarm rates (for Seniority Variants, $M = .13$, $SD = .17$, and for Gender Variants, $M = .10$, $SD = .21$).

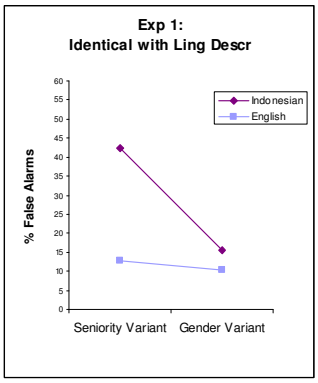


Figure 4: Results from Experiment 1.

Discussion

When participants described scenes verbally, they showed sensitivity to the dimension ensconced in their language on a subsequent recognition test. That is, Indonesians were significantly more able to reject variants that changed seniority relations than those that preserved seniority but changed gender. The results are stronger than those of our previous study (described above), in which linguistic descriptions were not elicited prior to the memory task. This suggests a 'thinking for speaking' effect whereby giving a linguistic description strengthens the effects of language on the encoding of the scenes. Interestingly, in this study the interaction appears to be driven by the Indonesian pattern; the English speakers showed roughly equal false alarms. However, the relative difference between the two language groups is as predicted: Indonesian speakers attended more to seniority than did English speakers. This pattern fits with our prediction of greater relative sensitivity to the dimension required in naming siblings.

One methodological point to note is that the nature of the design is limited in terms of the possible variants that we could devise for a given standard. This resulted in the Gender Variant being more perceptually similar to the Standard than was the Seniority Variant. In order to get around this problem, in the next study we decided to alter the identity match between the families in the pictures by

using a different family altogether for the variants. Thus, the variants did not retain the perceptual aspects of the Standard but were still designed to be relationally similar to the standard along either the seniority or gender dimension (see Figure 5). These new variants always embodied the same relationships as the previous variants. Thus the new variants fell further along the *literal similarity – analogical similarity continuum* (Gentner, 1989) than the original ones.

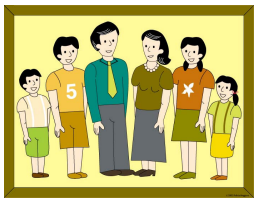
The analogous variants were used in a recognition task like the one described in Experiment 1. The idea was to test the strength of participants' "hold" of the relation and to see whether verbalization would influence the strength of the language effect. Of course, one might predict that participants would not false alarm at all to the analogous variants; after all, the variants depict different people altogether. On the other side, it seems possible that applying a linguistic description could invite a more abstract encoding, and that this could increase participants' propensity to recognize the same relation in different characters.

Indonesian and English speakers were shown a series of pictures: the three kinship standards and their three corresponding family pictures from Experiment 1, along with 23 other pictures. Participants were asked to remember the scenes for a later memory task. Recognition memory for the scenes was later tested using analogically similar Gender and Seniority Variants.

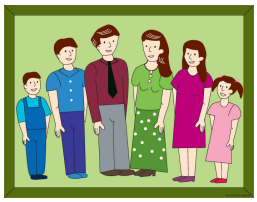
Experiment 2a

Participants Participants were 15 Indonesian monolinguals and 17 English monolinguals (not previously tested), ranging in age from approximately 17 to 20 years old. They were either given credit or a small monetary compensation. Data from Indonesian speakers were collected in Jakarta, Indonesia. Data from English speakers were collected at Northwestern University and other areas near Chicago.

Materials There were 29 study pictures (the three standard pictures from Experiment 1, their three corresponding family pictures, and 23 filler pictures). There were 67 test pictures (the three standards and all six of their variants, plus 58 fillers, as described below).

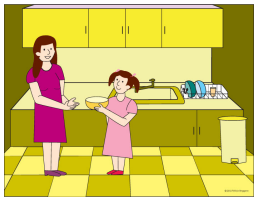


Identical Family Pict.

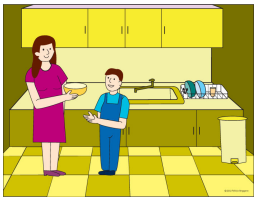


Analogous Family Pict.

Standard
older sister gives bowl to younger sister



Analogous
Seniority Variant
older sister gives bowl to younger brother



Analogous
Gender Variant
younger sister gives bowl to older sister

Figure 5: The Kitchen set, showing analogous variants used in Experiment 2.

Procedure As in Experiment 1, before each standard picture, participants were shown a family picture to ensure that they understood that the picture that followed involved only the children. (For the Ritual sets, the experimenter explained that the family lives on some island and held a ritual each year. Then the standard was shown without further description.) Participants were simply asked to remember the pictures. Then they were given a short break (approximately 10 minutes) during which they were asked to complete an unrelated paper-and-pencil task. Then they were given a yes-no recognition task. The two analogous variants for each standard were intermixed in semi-random order among the fillers. The three standards were given at the end of the test.

Results

An ANOVA over Language and Variant Type showed no significant main effects or interaction (all p's $> .1$). Qualitatively, there was a weak tendency for Indonesian speakers to make more false alarms to the Seniority Variants ($M = .36$, $SD = .32$) than to the Gender Variants ($M = .18$, $SD = .28$). The English speakers showed little difference between the two conditions (Seniority Variants, $M = .20$, $SD = .31$; Gender Variants, $M = .16$, $SD = .27$).

Experiment 2b

Participants Participants were 15 Indonesian monolinguals and 17 English monolinguals not previously tested, ranging in age from approximately 17 to 20 years old. They were either given credit or a small monetary compensation. Data from Indonesian speakers were collected in Jakarta, Indonesia. Data from English speakers were collected at Northwestern University and other areas near Chicago.

Materials Same as Experiment 2a.

Procedure Similar to Experiment 2b. The only difference is that during the study phase, participants were asked to verbally describe each scene, much like in Experiment 1.

Results

An ANOVA over Language and Variant Type revealed no significant main effects (all *p*'s >.1). As predicted, the ANOVA showed a significant interaction between the two factors ($F(1,30) = 5.16$, $p = .03$). There was a tendency for Indonesian speakers to make more false alarms to the Seniority Variants ($M = .18$, $SD = .28$) than to the Gender Variants ($M = .11$, $SD = .21$). The English speakers showed the reverse pattern; they made more false alarms to the Gender Variants ($M = .33$, $SD = .24$) than to the Seniority Variants ($M = .16$, $SD = .21$).

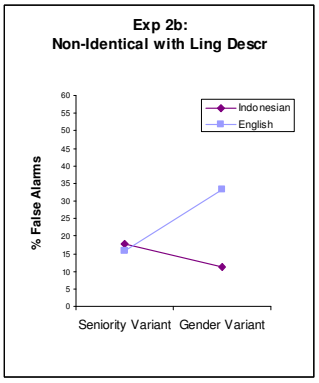


Figure 6: Results from Experiment 2b.

Discussion

The results from Experiment 2 suggest that for this more challenging task, cross-linguistic effects on encoding and recognition was only evident when participants were asked to verbally describe the scenes. In Experiment 2a, the effects of semantic categories on recognition seemed to be "masked" by the more challenging non-identical transfer task. When participants described the scenes verbally, however, as in Experiment 2b, this effect resurfaced. The pattern from the previous study was also found here: Indonesian speakers showed greater relative sensitivity to changes in seniority than to changes in gender in recognition memory, whereas English speakers showed the reverse pattern, as evidenced by the marginally significant interaction between language and variant type. As in Experiment 1, this pattern suggests greater sensitivity to the dimension that is required in naming siblings in each language.

Overall, we find a pattern of stronger results when people gave verbal descriptions of the scenes. Their descriptions, which typically included kinship terms, seem to heighten effects of language on encoding and memory. Using the identical family variants, the interaction between language and variant type (which was only marginal in our previous study) was significant in Experiment 1. Using the analogous variants (Experiment 2), the interaction between language and variant type was only significant when participants were instructed to use linguistic description. Our overall results suggest that the difference in the semantic patterns of the two languages may lead to differences in the way speakers encode situations – even nonlinguistic perceptual scenes. These results are consistent with the *thinking for speaking* idea in that actively using a language influences encoding and recognition memory.

Our thinking-for-speaking pattern of result is consistent with previous research showing that cross-linguistic differences influence judgments of spatial pictures (Gentner & Feist, submitted) and motion events (Malt, Sloman, & Gennari, 2003), but only when participants were asked to use linguistic description. However, it is also important to note that some studies have shown effects of cross-linguistic semantic differences on nonlinguistic performance without asking participants to first verbalize the scenes (e.g., Boroditsky, Ham, & Ramscar, 2002; Levinson et al., 2002). Indeed, in our own previous studies in this arena, we found cross-linguistic effects in similarity judgments and word extension without the use of verbalization. The similarity task was a simple triad judgment task and the word extension task was a novel name given to a standard picture and participants were asked to extend the novel name. In neither case were participants asked to describe the scenes themselves (Anggoro & Gentner, 2003). Interestingly, a comparison of the present results with our prior results (in which no verbal descriptions were elicited) (Anggoro & Gentner, 2003) suggests that in some cases, the predicted pattern are qualitatively stronger when prior verbalizations are elicited.

General Discussion

In conclusion, our findings suggest that linguistic differences in kinship terms (specifically, sibling terms) influence the way people encode and remember scenes and perceive similarities among them. Overall our results suggest that the actual verbalization of these semantic distinctions strengthens the cross-linguistic effects. Since in our previous work, cross-linguistic effects were found with or without linguistic description, the most intriguing aspect of our current findings is that this influence of verbalization appears stronger when the variants were *non-identical* to the standard[1]. The use of language may be particularly important in cases where the bridge between initial encoding and later experience is somewhat more abstract. Gentner (2003) has suggested that one role of language in

[1] Cross-experiment analyses of recognition results comparing participants who had initially given linguistic description of the scenes with those who had not (Anggoro & Gentner, 2003) showed a stronger pattern in the predicted direction for analogous pairs, especially among English speakers.

cognition is to facilitate memory access between situations that are not superficially similar but can be categorized in similar ways, as in the case of relational meanings.

Thus, our findings may provide a link between work in language and thought with work on analogical processing. Finally, our findings provide evidence that the use of language can influence encoding not only in spatial domains but also in the social arena.

Acknowledgments

This research was supported by NSF-ROLE award 21002/REC-0087516 and by the Culture, Language, and Cognition group at Northwestern University. We would like to thank Jeanne Arijanti, SMU PSKD in Jakarta, Ronnie Rios, Vini Putri, and Kathleen Braun for their assistance with data collection and other stages of this project. Special thanks to Felicia Anggoro for creating the materials.

References

Anggoro, F., & Gentner, D. (2003). Sex and seniority: The effects of linguistic categories on conceptual judgments and memory. *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*. Boston, MA.

Boroditsky, L., Ham, W. & Ramscar, M. (2002). What is universal about event perception? Comparing English and Indonesian speakers. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society.* Fairfax, VA

Boroditsky, L. & Schmidt, L. (2000). Sex, Syntax, and Semantics. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, Philadelphia, PA.

Bowerman, M. & Choi, S. (2003). Space under construction: Language specific spatial categorization in first language acquisition. In D. Gentner & S. Goldin-Meadow, Eds. *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Choi, S., McDonough, L., Bowerman, M., and Mandler, J.M (1999). Early sensitivity to language-specific spatial categories in English and Korean. *Cognitive Development, 14*, 241-268.

Danziger, E. (2001). Relatively speaking: Language, thought and kinship in Mopan Maya. *Oxford Studies in Anthropological Linguistics.* Oxford University Press.

Foley, W. (1997). *Anthropological Linguistics: An Introduction.* Blackwell Publishers.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199-241). London: Cambridge University Press. (Reprinted in *Knowledge acquisition and learning*, 1993, 673-694).

Gentner, D., & Feist, M. I. (submitted). Spatial language influences memory for spatial scenes.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Gumperz, J. J., & Levinson, S. C. (Eds.). (1996). *Rethinking linguistic relativity*. Cambridge, NY: Cambridge University Press.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*, 10-20.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*, 45-75.

Levinson, S. C. (1998). Studying spatial conceptualization across cultures: Anthropology and cognitive science. *Ethos, 26*, 7-24.

Levinson, S.C., Kita, S, Haun, D.B.M., & Rasch, B. J. (2002). Returning the tables: language affects spatial reasoning. *Cognition, 84*, 155-188.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition, 83*, 265-294.

Malinowski, B. (1930). Kinship. *Man, 30*, 19-29.

Malt, B. C., Sloman, S. A., & Gennari, S. (2003). Speaking vs. thinking about objects and actions. In D. Gentner & S. Goldin-Meadow, Eds. *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14,* 57-77.

Munnich, E., Landau, B., & Anne Dosher, B. (2001). *Cognition, 81, 171-207.*

Romney, A. K., & D'Andrade, R. G. (1964). Cognitive aspects of English kinship terms. *American Anthropologist, 66*, 146-270.

Sera, M. D., Elieff, C., Burch, M. C., Forbes, J., & Rodríguez, W. (2002). When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General, 131*, 377-397.

Slobin, D. I. (1987). Thinking for speaking. *Proceedings of the Berkeley Linguistic Society, 13*, 435-444.

Linking Rhetoric and Methodology in Formal Scientific Writing

Shlomo Argamon (argamon@iit.edu)
Illinois Institute of Technology
Dept. of Computer Science, 10 W 31st Street
Chicago, IL 60616, USA
Jeff Dodick (jdodick@vms.huji.ac.il)
The Hebrew University of Jerusalem
Department of Science Teaching, Givat Ram
Jerusalem 91904, ISRAEL

Abstract

Studying the communication patterns of scientists can give us insight into how science actually works. We argue that methodological differences between different scientific fields should lead to recognizable differences in how scientists in these fields use language to communicate with one another. This paper reports on a corpus-based study of peer-reviewed journal articles in paleontology and physical chemistry which used techniques of computational stylistics to compare the rhetorical styles used in the two fields. We found that indeed the two fields are readily distinguishable based on the stylistic character of their articles. As well, the most significant linguistic features of these distinctive styles can be connected directly to differences posited by philosophers of science between 'historical' (such as paleontology) and 'experimental' (such as physical chemistry) sciences.

Introduction

It has become clear in recent years that communication among different scientists working in a laboratory is critical for scientific success (Dunbar 1995). The particular uses of language by scientists serve to create a "collaborative space", whose worldview makes possible communication about complex observations and hypotheses (Goodwin 1994). Linguistic analysis has also been shown to elucidate features of scientific problem solving, as in Ochs et al.'s (1994) study of physicists' metaphoric talk of travel in a variety of graphical spaces.

At the same time, philosophers of science are increasingly recognizing that the classical model of a single "Scientific Method" (usually based on that of experimental sciences such as physics) does a disservice to sciences such as geology and paleontology, which are no less scientific by virtue of being historically oriented. Instead, it is claimed, differences in method may stem directly from the types of phenomena under study (Cleland, 2002). *Experimental* science (such as physics) attempts to formulate general predictive laws, and so relies heavily on repeatable series of controlled experiments which test hypotheses (Latour & Woolgar 1986). *Historical* science, on the other hand, deals with *contingent* phenomena, studying specific phenomena in the past in an attempt to find unifying explanations for effects caused by those phenomena (Mayr 1976). Because of this, reasoning in historical sciences consists largely of *reconstructive* reasoning, as compared to the *predictive* reasoning from causes to possible effects characteristic of experimental science (Gould 1986; Diamond 1999).

In this paper, we take some first steps towards analyzing the linguistic features of scientific writing in experimental and historical science, using several types of linguistically-motivated document features together with machine learning methods. Our goal is to examine if linguistic features that are indicative of different classes of scientific articles may be usefully correlated with the rhetorical and methodological needs of historical and experimental sciences. This paper describes a corpus-based study of genre variation between articles in a historical science (paleontology) and an experimental science (physical chemistry), with methodological differences as mentioned above. We hypothesize that corresponding rhetorical differences between articles in the respective fields will also be found. Standard methods of computational stylistics were used, confirming this hypothesis. Further, we defined a set of linguistically-motivated features for use in genre classification, based on systemic functional principles. These features enable a more nuanced examination of the rhetorical differences, allowing us to correlate these linguistic differences with the methodological differences posited by philosophers of science.

We note that the work reported here is only a first step, and more extensive studies of larger and more varied corpora of scientific papers will need to be undertaken in order to more firmly determine the links between scientific rhetoric, methodology, and cognition.

Hypotheses

Based on prior work in the philosophy and history of science we thus formulate our main hypothesis:

> ***H1*****:** *Stylistic features will distinguish more strongly between articles from different kinds (historical or experimental) of science than between articles from different journals in the same kind of science.*

We also formulate more detailed hypotheses regarding what sorts of rhetorical features we expect to be most significant in distinguishing articles in the different fields, based on posited methodological differences between historical and experimental sciences, as follows. First, a key element of *historical* reasoning is the need to differentially weight the evidence. Since any given trace of a

past event is typically ambiguous as to its possible causes, many pieces of evidence must be combined in complex ways in order to form a confirming or disconfirming argument for a hypothesis (termed *synthetic* thinking by Baker (1996)). Such thinking is, as Cleland (2002) argues, a necessary commitment of historical science (as opposed to experimental science), due to the fundamental asymmetry of causation. A single cause will often have a great many disparate effects, which if taken together would specify the cause with virtual certainty. Since all the effects cannot actually be known (as some are lost in the historical/geological record), evidence must be carefully weighed to decide between competing hypotheses (the methodology sometimes known as "multiple working hypotheses"). Experimental sciences tend, on the other hand, to adhere more or less to a "predict and test" methodology, in which manipulative experiments are used to confirm or disconfirm specific hypotheses (Cleland 2002). We therefore hypothesize:

***H2a*:** *Writing in historical science has more features expressing the weight, validity, likelihood, or typicality of different assertions or pieces of evidence*

***H2b*:** *Writing in experimental science has more features typical of explicit reasoning about predictions and expectations.*

Note that the presence or absence of linguistic features that can be linked to reasoning of a particular type is not by itself evidence of such reasoning. However, a consistent pattern of many of these features (as shown below) together aligned with the dichotomy proposed in *H2* strongly argues for such differences, which future research will attempt to elucidate in greater detail.

The Corpus

The study reported here was performed using a corpus of recent (2003) articles drawn arbitrarily from four peer-reviewed journals in two fields: *Palaios* and *Quaternary Research* in paleontology, and *Journal of Physical Chemistry A* and *Journal of Physical Chemistry B* in physical chemistry (chosen in part for ease of electronic access). *Palaios* is a general paleontological journal, covering all areas of the field, whereas *Quaternary Research* focuses on work dealing with the quaternary period (from roughly 1.6 million years ago to the present). The two physical chemistry journals are published in tandem but have separate editorial boards and cover different subfields of physical chemistry, specifically: studies on molecules (*J. Phys Chem A*) and studies of materials, surfaces, and interfaces (*J. Phys Chem B*). The numbers of articles used from each journal and their average (preprocessed) lengths in words are given in Table 1.

Table 1. Journals used in the studies with number of articles and average words per article.

Journal	# Art.	Avg. Words
Palaios	116	4584
Quaternary Res.	106	3136
J. Phys. Chem. A	169	2734
J. Phys. Chem. B	69	3301

Study 1: Distinctiveness

Methodology

We first test hypothesis *H1* by testing on our corpus whether paleontological and physical chemistry articles are stylistically distinctive from each other. The method was to represent each document as a numerical vector, each of whose elements is the frequency of a particular lexical feature of the text. We then applied the SMO learning algorithm (Platt 1998) as implemented in the Weka system (Witten & Frank 1999), using a linear kernel, no feature normalization, and the default parameters. (Other options did not appear to improve classification accuracy, so we used the simplest option.) SMO is a support vector machine (SVM) algorithm; SVMs have been previously applied successfully to text categorization problems (Joachims 1998). Generalization accuracy was measured using 20-fold cross-validation[1].

Features

For this first study, we used a set of 546 functions words taken en masse from the stop-word list of the popular research information retrieval system AIRE (Grossman & Frieder 1998); this procedure ensured task and theory neutrality. The set of function words used are similar to those used in many previous studies, such as Mosteller and Wallace's (1964) seminal stylometric work[2]. Each document was thus represented as a vector of 546 numbers between 0 and 1, each the relative frequency of one of the function words.

Results and Discussion

Table 2 shows results for binary classification between each pair of journals in our corpus, giving the percentage of test articles erroneously classified (in 20-fold cross-validation) using linear SMO learning and function-word frequencies as features. We first note that average accuracy on test documents from different fields (historical vs. experimental) was at least 97%, indicating excellent discriminability (far above chance). At the same time, the two physical chemistry journals are quite indistinguishable, as 34% is slightly *greater than* the error of always choosing the majority class (since 69/238=29% of those

[1] In *k*-fold cross-validation (Mitchell 1997) the data is divided into *k* subsets of equal size. Training is performed *k* times, each time leaving out one of the subsets, and then using the omitted subset for testing, to estimate the classification error rate; the average error rate over all *k* runs is reported. This gives quite a stable estimate of the expected error rate of the learning method for the given training size (Goutte 1997).

[2] Relative frequencies of function words, such as prepositions, determiners, and auxiliary verbs, have been shown in a number of studies to be useful for stylistic discrimination, since they act as easily extracted proxies for the frequencies of different syntactic constructs, and also tend not to covary strongly with document topic.

Table 2. Error rates for linear SMO using function word features for pairs of journals using 20-fold cross-validation.

	Historical		Experimental	
	P	***QR***	***PCA***	***PCB***
Palaios	--	10%	**0.4%**	**1%**
Quat Res	10%	--	**2%**	**3%**
Ph Ch A	**0.4%**	**2%**	--	34%
Ph Ch B	**1%**	**3%**	34%	--

articles are from *Phys. Chem. B*). In the case of *Palaios* vs. *Quat. Res.* we get an average error rate of 10%, an order of magnitude higher than any error rate in the cross-disciplinary case. Hence these results support *H1*, in that articles across disciplines are more easily distinguished than articles within a single discipline (from different journals). Of course, the 10% error rate obtained for distinguishing the two paleontology journals is far less than the 48% we would get by majority class classification, which points to a subsidiary distinction between these two journals. This is not unreasonable, given that *Quat. Res.* deals with a specific subset of the topics in *Palaios*[3]. We leave this question, however, for future research.

Study 2: Systemic Variation

Methodology

In order to more precisely analyze the rhetorical differences between articles in the two fields a follow-up study used as features the relative frequencies of sets of keywords and phrases derived from consideration of notions of systemic functional linguistics (Halliday 1994).

Systemic functional linguistics (SFL) construes language as a set of interlocking choices for expressing meanings: "either this, that, or the other", with more general choices constraining the possible specific choices. For example: "A message is either about doing, thinking, or being; if about doing, it is either standalone action or action on something; if action on something it is either creating something or affecting something pre-existent," and so on. A *system* is a set of options for meanings to be expressed, with *entry conditions* denoting when that choice is possible – for example, if a message is not about doing, then there is no choice possible between expressing standalone action or action on something. Each option has also a *realization specification*, giving constraints (lexical, featural, or structural) on statements expressing the option. Options serve as entry conditions for more specific subsystems.

[3] This may be related to the fact that *Quat. Res.* contains more articles than *Palaios* using chemical and radiochemical assaying, since such techniques are only applicable to younger remains from the Quaternary Period; such tools in fact are similar to the experimental techniques seen in physical chemistry. Indeed this is corroborated by the fact that the error rate between Quat Res and the PC journals was higher than Palaios and the same PC journals. More detailed study of the specific articles will be needed to test and refine this hypothesis.

By viewing language as a complex of choices between mutually exclusive options, the systemic approach is particularly appropriate to examining variation in language use. A systemic specification allows us to ask the following type of question: In places where a meaning of general type A is to be expressed in a text (e.g., "a message about action"), what sorts of more specific meanings (e.g., "standalone action" or "action on a thing") are most likely to be expressed by different types of people or in different contexts? A general preference for one or another option, when not dictated by specific content, is indicative of individual or social/contextual factors. Such preferences can be measured by evaluating the relative probabilities of different options by tagging their realizations in a corpus of texts (Halliday 1991).

As features, then, in the absence of a reliable systemic parser, we use keywords and phrases as proxy *indicators* for various systems. For example, an occurrence of the word "certainly" usually indicates that the author is making a high-probability modal assessment of an assertion. The drawback of this approach is lexical ambiguity, since the meaning of such keywords can depend on context. We reduce the effect of ambiguity, however, by using as complete a set of such *systemic indicator* keywords/phrases as possible for each system we represent, and also by using only measures of *comparative* frequency between the aggregated features. In addition, since we use very large sets of indicators for each system, it is unlikely that such ambiguity would introduce a systematic bias, and so such noise is more likely to just reduce the significance of our results instead of biasing them. Preprocessed articles in our corpus were each converted into a vector of 101 feature values (relative frequencies of system options) and the same learning protocol (using SMO) was used as in Study 1.

Features

The systemic features we used are based on options within three main systems, following Matthiessen's (1995) grammar of English, a standard SFL reference. Indicator lists were constructed by starting with the lists of typical words and phrases given by Matthiessen, and expanding them to related words and phrases taken from Roget's Interactive Thesaurus[4] (manually filtered for relevance). Keyword lists were constructed entirely independently of the target corpus. We used systems and subsystems within: CONJUNCTION, linking clauses together (either within or across sentences); MODALITY, giving judgments regarding probability, usuality, inclination, and the like; and COMMENT, expressing modal assessments of attitude or applicability. MODALITY and COMMENT relate directly to how propositions are assessed in evidential reasoning (e.g., for likelihood, typicality, consistency with predictions, etc.), while CONJUNCTION is a primary system by which texts are constructed out of smaller pieces, and so may be expected

[4] http://www.thesaurus.com

Table 3. Average error rates for linear SMO using systemic features for pairs of journals using 20-fold cross-validation.

	Historical		Experimental	
	P	***QR***	***PCA***	***PCB***
Palaios	--	26%	**9%**	**9%**
Quat Res	26%	--	**17%**	**14%**
Ph Ch A	**9%**	**17%**	--	32%
Ph Ch B	**9%**	**14%**	32%	--

Table 4. Strong features (see text) for Paleontology or Physical Chemistry, using SMO.

System	**Hist.**	**Exper.**
CONJUNCTION	Extension	Enhancement
COMMENT	Validative	Predictive
MODALITY/**Type**	Modalization	Modulation
Modalization: Manifestation	Implicit	Explicit
Modulation: Manifestation	Explicit	Implicit

to reflect possible differences in overall rhetorical structure[5]. These systems and the indicators we used are described more fully in the Appendix.

Results and Discussion

We first check inter-class discriminability (*H1*), testing the results of Study 1 above. Table 3 presents classification error rates averaged over 20-fold cross-validation. In all four **cross**-disciplinary cases, error rates are 17% or less, while in the two **intra**-disciplinary cases, accuracy is noticeably lower; *Palaios* and *Quat. Res.* are significantly less distinguishable at 26% error, while *J. Phys. Chem. A* and *J. Phys. Chem. B* are entirely undistinguishable[6]. This further supports hypothesis *H1*, as above. Moreover, consistency with Study 1 results helps to validate the approach taken in this study.

We now consider what consistent picture, if any, emerges of the rhetorical difference between the two classes of scientific articles (paleontology and physical chemistry) from the patterns of feature weights in the learned models. To do this, we ran SMO on the entire corpus (without reserving test data) for each of the four pairs of a paleontology with a physical chemistry journal, and ranked the features according to their weight for one or the other journal in the weight vector. We call a feature *strong*, if it was among the 30 with the highest absolute weights out of 101 features for the same class in models learned for all journal pairs. Among strong features, some striking patterns emerge, shown in Table 4.

[5] Other textual/cohesive systems, such as PROJECTION, TAXIS, THEME, and INFORMATION cannot be easily addressed, if at all, using a key-word-based approach.

[6] Error rates are higher for this feature set than for the function words due to the smaller number of features—clearly there are some stylistic differences that our systemic features do not capture.

First, in COMMENT, we see a preference for Validative comments by paleontologists and one for Predictive comments by physical chemists. This linguistic opposition directly supports both hypotheses *H2a* and *H2b*, related to methodological differences between historical and experimental sciences. As noted, the historically-oriented paleontologist has a rhetorical need to explicitly delineate the scope of validity of different assertions, as part of synthetic thinking (Baker 1996) about complex and ambiguous webs of past causation (Cleland 2002). This is not a primary concern, however, of the experimentally-oriented physical chemist; her main focus is prediction: the predictive strength of a theory and its predictive consistency with the evidence.

Next, we consider the (complicated) system of MODALITY. At the coarse level represented by the simple features, we see a primary opposition in Type. The preference of the (experimental) physical chemist for Modulation (assessing what 'ought' or 'is able' to happen) is consistent with a focus on prediction and manipulation of nature, and supportive of hypothesis *H2b*. The (historical) paleontologist's preference for Modalization (assessing 'likelihood' or 'usuality') is consistent with the outlook of a "neutral observer" who cannot directly manipulate or replicate outcomes, and is thus supportive of hypothesis *H2a*.

This same pattern is also seen within the complex paired features combining values for modality **Type** and **Manifestation**. Implicit variants are more likely to be used for options that are well-integrated into the expected rhetoric, while Explicit realizations are more likely to be used for less characteristic types of modal assessment, as more attention is drawn to them in the text. Keeping this in mind, note that Modalization is preferably Implicit in paleontology but Explicit in physical chemistry; just the reverse holds for Modulation. This shows that Modalization is integrated smoothly into the overall environment of paleontological rhetoric, and similarly Modulation is a part of the rhetorical environment of physical chemistry.

Finally, in the textual system of CONJUNCTION, we see a clear opposition between Extension, indicating paleontology, and Enhancement, indicating physical chemistry. This implies that paleontological text has a higher density of discrete informational items, linked together by extensive conjunctions, whereas in physical chemistry, while there may be fewer information items, each is more likely to have its meaning deepened or qualified by related clauses. This may be indicative that paleontological articles are more likely to be primarily descriptive in nature, requiring a higher information density, while physical chemists focus their attention more deeply on a single phenomenon at a time. At the same time, this linguistic opposition may also reflect differing principles of rhetorical organization: perhaps physical chemists prefer a single coherent 'story line' focused on enhancements of a small number of focal propositions, whereas paleontologists may prefer a multifocal 'landscape' of connected propositions. Future work will

include interviews and surveys of the two types of scientists to investigate these hypotheses.

Related Work

Previous work has investigated the relationship between choice probabilities and contextual factors. For example, Plum & Cowling (1987) demonstrate a relation between speaker social class and choice of verb tense (past/present) in face-to-face interviews. Similarly, Hasan (1998) has shown, in mother-child interactions, that the sex of the child and the family's social class together have a strong influence on several kinds of semantic choice in speech. These previous studies involved hand-coding a corpus for systemic-functional and contextual variables and then comparing how systemic choice probabilities vary with contextual factors via multivariate analysis. By contrast, this study uses large numbers of neutral features and machine learning to automatically build accurate classification models.

Further, by examining differences between systemic preferences across scientific genres, we are quantitatively analyzing differences in register. *Register* denotes functional distinctions in language use related to the context of language use (Eggins & Martin 1997), and may be considered to comprise: *mode*, the communication channel of the discourse; *tenor*, the effect of the social relation between the producer and the audience; and *field*, the domain of discourse. We focus in this paper on the field-related distinction between historical and experimental science, with mode and tenor held relatively constant, by using articles written by working scientists drawn from peer-reviewed journals. Our results indicate that the difference in the types of reasoning needed by historical and experimental sciences leads to correlated differences in rhetorical preferences (perhaps best understood as 'functional tenor' (Gregory 1967)).

Conclusions

We have shown how machine learning techniques together with linguistically-motivated features can be used to provide empirical evidence for rhetorical differences between writing in different scientific fields. Further, by analyzing the models output by the learning procedure, we can see what features realize the differences in register that are correlated with different fields. This provides indirect evidence for methodological variation between the sciences, insofar as rhetorical preferences can be identified which are linked with particular modes of reasoning. This study thus provides empirical evidence for those philosophers of science who argue against a monolithic "scientific method".

Future work will include validating these results against a larger corpus of articles including more scientific fields, as well as incorporating more involved linguistic processing—the rhetorical parsing methods developed by Marcu (2000) are an important step in this direction. Methods for discovering rhetorically important features such as the subjectivity collocations of Wiebe et al. (2001) may also be helpful. Further, the current study treats each article as an indivisible whole. However, as noted by Lewin et al. (2001) in their analysis of social science texts, the rhetorical organization of an article varies in different sections of the text—future work will include studying rhetorical variation across different sections of individual texts, by incorporating techniques such as those of Teufel and Moens (1998).

References

Baker, V.R. (1996). The pragmatic routes of American Quaternary geology and geomorphology. *Geomorphology* **16**, pp. 197-215.

Cleland, C.E. (2002). Methodological and epistemic differences between historical science and experimental science. *Philosophy of Science*.

Diamond, J. (1999). *Guns, Germs, & Steel.* (New York: W. W. Norton and Company).

Dodick, J. T., & N. Orion. (2003). Geology as an Historical Science: Its Perception within Science and the Education System. *Science and Education*, **12**(2).

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R.J. Sternberg, & J. Davidson (Eds.). *Mechanisms of Insight*. (Cambridge MA: MIT Press). pp. 365-395.

Eggins, S. & J. R. Martin, (1997). Genres and registers of discourse. In T. A. van Dijk, *Discourse as structure and process. A multidisciplinary introduction*. Discourse studies 1 (London: Sage), pp. 230–256.

Goodwin, C. (1994). Professional Vision. *American Anthropologist*, **96**(3), pp. 606-633.

Gould, S. J. (1986). Evolution and the Triumph of Homology, or, Why History Matters, *American Scientist* (Jan.-Feb. 1986): pp. 60-69.

Goutte, C. (1997) Note on free lunches and cross-validation, *Neural Computation*, 9(6):1246-9.

Gregory M., (1967). Aspects of varieties differentiation, *Journal of Linguistics* **3**, pp. 177-198.

Grossman, D. and O. Frieder (1998). *Information Retrieval: Algorithms and Heuristics* , Kluwer Academic Publishers.

Halliday, M.A.K. (1991). Corpus linguistics and probabilistic grammar. In Karin Aijmer & Bengt Altenberg (ed.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. (London: Longman), pp. 30-44.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. (London: Edward Arnold).

Hasan, R. (1988). Language in the process of socialisation: Home and school. In J. Oldenburg, Th. v Leeuwen, & L. Gerot (ed.), *Language and socialisation: Home and school*. North Ryde, N.S.W.: Macquarie University.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*, pp. 137-142.

Latour, B. & S. Woolgar, (1986). *Laboratory Life: The Construction of Scientific Facts* (Princeton: Princeton Univ. Press).

Lewin, B.A., J. Fine, & L. Young (2001). *Expository Discourse: A Genre-Based Approach to Social Science Research Texts* (Continuum).

Marcu, D. (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Comp. Ling.*, **26**(3), pp. 395-448.

Martin, J. R. (1992). *English Text: System and Structure*. (Amsterdam: Benjamins).
Matthiessen, C. (1995). *Lexicogrammatical Cartography: English Systems*. (Tokyo, Taipei & Dallas: International Language Sciences Publishers).
Mayr, E. (1976). *Evolution and the Diversity of Life.* (Cambridge: Harvard University Press).
Mitchell, T. (1997) *Machine Learning.* (McGraw Hill).
Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist Papers,* (Reading, MA: Addison Wesley).
Ochs, E., S. Jacoby, & P. Gonzales, (1994). Interpretive journeys: How physicists talk and travel through graphic space, *Configurations* **1**:151-171.
Platt, J. (1998), *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,* Microsoft Research Technical Report MSR-TR-98-14.
Plum, G. A. & A. Cowling. (1987). Social constraints on grammatical variables: Tense choice in English. In Ross Steele & Terry Threadgold (ed.), *Language topics. Essays in honour of Michael Halliday*. (Amsterdam: Benjamins).
Teufel, S., and Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. In *Proc. AAAI Spring Symposium on Intelligent Text Summarization*.
Wiebe, J., T. Wilson and M. Bell. (2001). Identifying Collocations for Recognizing Opinions. In *Proc. ACL/EACL '01 Workshop on Collocation, Toulouse, France, July 200.*
Witten, I.H. and Frank E. (1999). *Weka 3: Machine Learning Software in Java:* http://www.cs.waikato.ac.nz/~ml/weka.

Appendix: Systems and Features

CONJUNCTION

On the discourse level, the system of Conjunction serves to link a clause with its textual context, by denoting how the given clause *expands* on some aspect of its preceding context. Similar systems also operate at the lower levels of noun and verbal groups, while denoting similar logico-semantic relationships, e.g., "*and*" usually denotes "additive extension". Options within Conjunction are as follows:

- Elaboration: Deepening the content of the context
  - Appositive: Restatement or exemplification
  - Clarifying: Correcting, summarizing, or refocusing
- Extension: Adding new related information
  - Additive: Adding new content to the context
  - Adversative: Contrasting new information with old
  - Verifying: Adjusting content by new information
- Enhancement: Qualifying the context
  - Matter: What are we talking about
  - Spatiotemporal: Relating context to space/time
    - Simple: Direct spatiotemporal sequencing
    - Complex: More complex relations
  - Manner: How did something occur
  - Causal/Conditional:
    - Causal: Relations of cause and effect
    - Conditional: Logical conditional relations

Note that the actual features by which we represent an article are the frequencies of each subsystem's indicator features, each measured relative to its siblings. So, for example, one feature is Elaboration/*Appositive*, whose value is the total number of occurrences of Appositive indicators divided by the total number of occurrences of Elaboration indicators (Appositive + Clarifying). The relative frequencies of Elaboration, Extension, and Enhancement within Conjunction are also used as features.

COMMENT

The system of Comment is one of modal assessment, comprising a variety of types of "comment" on a message, assessing the writer's attitude towards it, or its validity or evidentiality. Comments are generally realized as adjuncts in a clause (and may appear initially, medially, or finally). Matthiessen (1995), following Halliday (1994), lists eight types of Comment, which we give here along with representative indicators for each such subsystem.

- Admissive: Message is assessed as an admission
- Assertive: Emphasizing the reliability of the message
- Presumptive: Dependence on other assumptions
- Desiderative: Desirability of some content
- Tentative: Assessing the message as tentative
- Validative: Assessing scope of validity
- Evaluative: Judgment of actors behind the content
- Predictive: Coherence with predictions

MODALITY

The features for interpersonal modal assessment that we consider here are based on Halliday's (1994) analysis of the Modality system, as formulated by Matthiessen (1995). In this scheme, modal assessment is realized by a simultaneous choice of options within four systems[7]:

- Type: What kind of modality?
  - Modalization: How 'typical' is it?
    - Probability: How likely is it?
    - Usuality: How frequent/common is it?
  - Modulation: Will someone do it?
    - Readiness: How ready are they (am I)?
    - Obligation: Must I (they)?
- Value: What degree of the relevant modality scale?
  - Median: In the middle of the normal range.
  - High: More than normal
  - Low: Less than normal
- Orientation: Is the modality expressed as an Objective attribute of the clause or as Subjective to the writer?
- Manifestation: Is the assessment Implicitly realized by an adjunct or finite verb, or Explicitly by a projective clause?

The cross-product of these subsystems gives many modality assessment types, each realized through a subset of indicators. *Simple* features are each option in each system above (e.g., Modalization/*Probability* opposed to Modalization/*Usuality*), while *complex* features are pairwise combinations of such simple features. The indicator set for each such feature is the intersection of the indicator sets for the two component features. Frequencies were normalized by the total set of occurrences of both primary systems (Modalization and Value in the previous example).

[7] Note that we did not consider the system of POLARITY, since it cannot be properly addressed without more sophisticated parsing.

Domain-Specificity in Shape Categorization and Perception

Benjamin J. Balas (bjbalas@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139-4307 USA

Abstract

We examine the influence of domain-specific knowledge on the process of learning and representing simple visual categories. Depending on whether subjects' construe simple objects as living kinds or machines, they show differential sensitivity to the importance of basic shape features. In particular, subjects who treated the objects as machines placed less importance on coarse shape differences unrelated to the described function of the objects in both categorization and productive (drawing) tasks. These findings suggest that domain-specific functional understanding of objects may influence the formation of shape categories, and perhaps the perception of these shapes.

Introduction

It is typical in the study of pattern recognition to discuss the classification of objects as though they are a relatively homogeneous class of stimuli. Many sets of 2D and 3D shape primitives have been proposed, all with the goal of describing how a wide variety of forms can be catalogued by a small set of simple building blocks (Hoffman & Richards, 1986; Biederman, 1987). Special-purpose mechanisms are seen as add-ons that are only used to process particularly interesting and ecologically significant stimuli. The nature of these specialized processes has also been explored, and there is much debate over what defines a "special" class of stimuli (Yin, 1969; Diamond & Carey, 1986; Gauthier & Tarr, 1997).

The suggestion that object recognition only employs unique processes for a small minority of object classes stands in stark contrast to work concerned with the formation and application of domain-specific theories (Carey, 1985). It has been suggested that knowledge of particular domains (like biology, physics, or psychology) may substantially affect the reasoning employed in a range of tasks. Given that theory-based reasoning may guide performance in complex scenarios, it may also be possible that human observers possess theories concerning the visual properties of objects in various domains that affect the way they recognize and represent object categories. Should this prove to be the case, it may suggest that general accounts of object recognition are too coarse, in that they fail to consider the richness of subjects' visual knowledge of a particular object category.

It is important to establish exactly what we mean when we suggest that subjects may possess theories about visual properties of objects that affect perception. We envision a simple hierarchy of object knowledge (Figure 1) ranging from high-level theories to low-level perception of shapes. At the top is amodal knowledge of particular domains. At this level (L1), abstract facts about objects in a broad domain (e.g., "artifacts") are stored propositionally (as in, "Machines are often built in factories to precise specifications"). At the second level (L2), object categories are defined in terms of the visual properties shared by members of common groups. This knowledge could be symbolic ("Tigers have stripes") or encoded in terms of visual measurements ("Tigers have lots of contrast energy at a particular spatial frequency"). Finally, at the bottom level (L3) lies the representation constructed in perceiving an object as a category instance, expressed in terms of its visual features such as size, shape, color, etc. Our work set out to explore the possibility that abstract knowledge at the highest level of this hierarchy could affect representations underlying both shape categories (L3 → L2) and shape perception (L3 → L1).

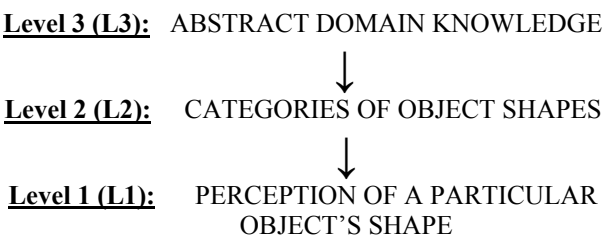


Figure 1: A schematic view of the relationships between domain theories and perception.

Previous experiments by Goldstone (1994) have demonstrated the influence of categorization on perception (L2 → L1). After forming categories of simple objects based on perceptual attributes (like luminance), subjects' ability to discriminate between luminance levels was altered to support the newly learned perceptual groups. No abstract domain knowledge (L3) was implicated in these studies. Kelemen and Bloom (1994) demonstrated an influence of domain knowledge on the representation of shape categories (L3 → L2), but did not look at influences down to L1. Their stimuli were uniform circles that could vary both in size and color. Subjects who were told the circles were microscopic animals preferred to categorizethem according to color, while those who construed them as machines preferred to categorize based on size.

Visual categorization relies on the ability to learn object attributes that vary more across categories than within them, and to accurately measure those attributes in new images.

Knowing that an object is a member of a particular domain (say "living things") may bias the observer to expect particular patterns of variability. The Kelemen and Bloom study, as well as Keil's work (1998), have both shown that color is expected to be useful for categorizing living things. We suggest that different aspects of shape may also be differentially weighted to categorize objects in different domains. The abstract (Level 3) knowledge observers possess about animals' growth and movement as well as knowledge about artifact construction may lead to influences on representations of shape categories (Level 2), and the ability to make fine-grained perceptual distinctions at Level 1.

Another possibility relevant to the categorization of living and non-living things concerns the influence of functional information on shape perception. It has been proposed that function helps set the core meaning of artifact concepts, and perceived shape may be relevant to artifact categorization primarily to the extent that it supports a functional interpretation (Bloom, 2000). A theory of non-living things may not induce any *a priori* preferences for particular visual features, but rather, flexibly bias resources towards functionally relevant information. Function may also influence which shape features are perceived to be important for biological categories, but in different ways. Landau et al. (1998) have shown that people will be more tolerant of certain non-rigid shape variation when classifying objects construed as animals (relative to those construed as artifacts).

In sum, our studies here ask two main questions not addressed in previous work. First, how far down in the hierarchy of Figure 1 do domain-specific conceptual influences extend? In particular, do they extend down to perceptual representations of individual objects? Second, what is the range of conceptual influences? Specifically, is there a role for functional understanding in forming representations of shape categories and individual shapes? In the spirit of Kelemen and Bloom, we have created novel 2-D shapes that we have named "Marcons" and "Draxels." These two populations of objects both consist of ellipses whose perimeters have been modified to contain sinusoidal bumps (Figure 2). The result is a set of stimuli that can be distinguished both by coarse shape information (ellipse eccentricity) and fine details (the frequency and/or amplitude of the bumps). This allows us to explore subtler aspects of object appearance than previous studies have done, and provides us with the ability to easily attribute function to shape properties.

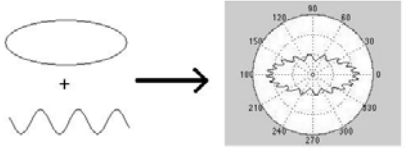


Figure 2: Constructing a Draxel.

In the experiments we present, subjects were asked to learn the distinction between Draxels and Marcons. While both groups are given identical information about the behavior of these objects and the function of particular features, some subjects are told that the stimuli are nanomachines and others are told they are amoeba-like animals. In Experiment 1, we use productive data (subjects' drawings) to determine whether domain knowledge influences perceptual representations. Drawings are a particularly useful tool in that they require subjects to make their perceptual representations explicit. In Experiment 2, we look for the effects of domain knowledge on category representations through a more controlled categorization task.

Experiment 1

We begin by presenting the results of a learning task in which subjects learn to discriminate between Marcons and Draxels. We assess the nature of their post-learning representations of the two kinds of object by collecting drawings from all subjects, and examining the extent to which differences in elongation and 'bumpiness' are expressed.

Methods

Stimuli 32 Marcons and 32 Draxels were hand-drawn on 2 ⅛" x 2 ¾" cards for this task using Crayola™ magic markers. A stencil was used to enforce a major/minor axis ratio of 2.9:1 for Draxels compared to a 2:1 ratio for Marcons. Bumps were applied to the perimeter of each ellipse such that Draxels contained 2.9 bumps/cm and Marcons contained 2.0 bumps/cm. The length of the major axis could take on one of 8 values for both Draxels and Marcons, and a dual-color contour was applied to the perimeter of the finished figures. The figures were also depicted at four different orientations (0, 90, +20, -20 degrees from vertical). No conjunction of color, size, and orientation was diagnostic of object identity, leaving only ellipse eccentricity and bump frequency as useful criterion for discrimination.

Subjects 24 naïve subjects (9 men, 15 women) were recruited from the MIT community to participate in this task. Subject age ranged from 18-40 years of age.

Procedure Subjects were initially presented with a brief introductory paragraph explaining that Marcons and Draxels were either unicellular organisms (Animals condition) or nanomachines (Machines condition). In both cases, the two kinds of object were said to participate in "agricultural revitalization" by grabbing onto various chemical compounds with their bumps, and redistributing them across depleted soil. Subjects were told that Marcons and Draxels were quite similar, but that experts could identify them very accurately despite the range of individual sizes, shapes, and colors in which the objects appeared.

After reading this paragraph, subjects were shown one example each of a Marcon and Draxel on an 8 ½" x 11" placemat. These two items were the same dimension as the largest exemplars in the stimulus set mentioned previously, but were depicted with a novel dual-color contour. Subjects were presented with the shuffled deck of 64 cards, and asked to sort the pile into two stacks such that Draxels were on one side and Marcons on the other. Subjects were permitted to take as much time as they liked to sort the cards.

After each round of sorting, the experimenter determined what false classifications were made, and presented these items to the subject for further study (grouped underneath their proper place on the sorting mat) before they were shuffled back into the deck for the next round. Subjects sorted the cards until they made fewer than 8 errors, or until they had sorted through the entire deck 4 times.

When the card-sorting task had been completed, subjects were then presented with new instructions asking them to produce drawings of Draxels and Marcons. 8 examples of each object were requested, with the additional instruction that their drawings should depict what they believed "typical" members of each category looked like, and that their set of 8 drawings should attempt to cover the range of variations that existed within each category. After completing their drawings, subjects were given a brief questionnaire asking them to rate on a 1-10 scale how important were various visual features in their concepts of these two classes, and to enumerate in free-response style the differences they perceived between the two kinds of objects.

Results

Learning Rates To determine if either group showed a particular affinity for learning to discriminate between Marcons and Draxels, we examine the number of errors made by each group after their first and second rounds of sorting. These two rounds are of particular importance in that they indicate to what extent the task is difficult with only one example of each object type and how much improvement each group undergoes by viewing a population of labeled examples.

A two-way ANOVA, with sorting round and domain as factors, revealed only a main effect of sorting round ($p < 0.05$). Subjects improve from across rounds, but neither group was particularly better at performing the discrimination between Marcons and Draxels, nor benefited more than their counterparts from receiving multiple labeled examples after their first round of sorting. We note that of the 24 participants who performed this learning task, 6 subjects (3 Animals, 3 Machines) were unable to reach our performance criterion of fewer than 8 mistakes after 4 rounds of sorting. We take this to mean that the difficulty of this initial task was intermediate, and unrelated to the domain of the objects.

Post-Learning Questionnaire Subjects' responses to the post-experiment questionnaire were analyzed to determine if there were differences between the Animal and Machine groups' explicit feature preferences. A two-factor ANOVA was run on subjects' ratings of the importance of shape differences between the two object categories, with feature type as one factor (elongation v. bumps) and subject group (animal v. machine) as the other factor. No main effects were found in the analysis, but a significant interaction ($p < 0.05$) was found between feature type and subject group. Subjects in the Animal group rated eccentricity as a more important feature, compared to subjects in the Machine group who preferred to use the bumps. (Table 1)

Table 1:
Mean ± SD ratings of feature importance (1-10 scale)

	Animals	Machines
Elongation	8.3 ± 4.1	6.5 ± 2.6
Bumps	6.8 ± 2.0	8.8 ± 3.0

Subjects' Drawings Using these responses as a guide, we turn next to the drawings of Marcons and Draxels produced by subjects after learning. (Figure 3) The eccentricity and number of bumps/cm for each figure was measured by first inscribing the largest ellipse possible inside the bumpy contour. If a particular drawing was sufficiently irregular that this proved impossible, that figure was excluded from the analysis. Only 4 out of 384 drawings (all from different subjects) were excluded in this fashion. Eccentricity was determined for each figure by measuring the major and minor axes of the inscribed ellipse, and the number of bumps/cm was determined by counting the number of bumps on the drawing and dividing by its perimeter. We use the YNOT approximation of the perimeter of an ellipse (Maertens & Rousseau, 2000) here, which was also used in the creation of the stimuli.

For each subject, we then compute the mean values of both eccentricity and bumps/cm for Marcons and Draxels across all eight drawings. The difference between these means is taken for each feature type, and divided by the maximum standard deviation of that feature within an object population. In this way, we express the differences in Marcon and Draxel shapes as a function of the separation and spread of the populations produced by each individual subject(Table 2).

Table 2:
Normalized differences between Marcon and Draxel features (Mean ± SD)

	Animals	Machines
Elongation	1.93 ± 0.98	0.79 ± 1.04
Bumps	2.42 ± 1.55	2.31 ± 1.27

A two-way ANOVA was performed on these measurements, using feature type and subject group as factors. A main effect of feature type was found, (bumps > eccentricity, $p < 0.05$) with a marginally significant effect of

subject group (Animals > Machines, $p < 0.08$). No significant interaction was found. However, to tease apart the contributions of each feature type to the weak effect of perceived domain, we conducted a further analysis for simple main effects across the Animal and Machine groups. We find in this analysis that subjects in the Animal condition expressed differences in eccentricity significantly more than subjects in the Machine condition ($p < 0.05$), while no such difference exists for the expression of bump density ($p > 0.8$).

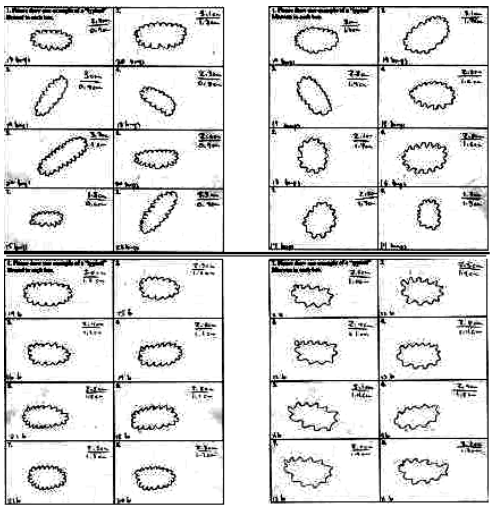


Figure 3: Examples of Draxels (left) and Marcons (right) created by subjects in the "Animals" condition (top) and the "Machines" condition (below). Note the lack of elongation differences in the lower drawings.

Discussion

Subjects' ratings of feature importance indicate an intriguing interaction between domain and the visual features perceived to be important. The drawings produced after completing the learning task show significantly more exaggeration of bumps across both domains, as well as significantly more expression of elongation differences in the Animal condition compared to the Machine condition. Bumpiness was not significantly more exaggerated in the Machine group compared to the Animal group, as one might expect from the interaction in subjects' ratings, yet there are some interesting qualitative differences in the way differences in bumpiness are expressed. Subjects in the Machine condition often pointed out subtle differences between the two categories that were not expressed or described by subjects in the Animal condition. Subjects in the Machine condition often pointed out aspects of bump symmetries and asymmetries that they felt were important to the task, and often included these details in their drawings (Figure 4). Though these features are not directly related to the difference in bump frequency, and therefore did not contribute to our quantitative analysis, it is interesting to see how these features appear in the drawings of Machine subjects while remaining almost wholly absent from the drawings of Animal subjects. It may be that more aspects of bump shape need to be explicitly considered when constructing stimuli and examining subjects' drawings.

We note that both of these effects may be a consequence of the introductory scenario given to subjects at the beginning of the task. By indicating that the bumps had a particular functional importance, the representation of bump differences may have been weighted more heavily in both groups. It is interesting to note that in the Machines condition, this seems to have resulted in an overall tendency to ignore additional shape differences. The significance of elongation differences in this group may have been compromised by the direction of functional information away from these features.

Experiment 1 provides us with direct evidence that domain knowledge can influence shape categorization (L3 → L2, in Figure 1). Evidence for effects of L3 on L1 (perceptual shape representations) is only indirect, insofar as subjects are thought to produce drawings by translating a Level 2 representation into perceptual primitives. This is an important distinction, since Level 2 representations may be more symbolic ("Draxels are skinnier than Marcons") rather than truly visual ("Draxel elongation is about 3:1"). A valuable next step would be to measure perceptual abilities directly using psychophysical methods (e.g., Goldstone, 1994), thereby establishing a more direct L3 → L1 link.

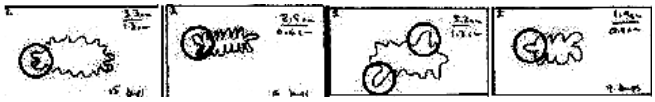


Figure 4: Over-expression of symmetry and asymmetry in "Machines" subjects. Draxels are at left, Marcons right.

Experiment 2

In our second task, we pursue possible domain-specific differences in shape processing using a more controlled task. Rather than relying on subjects' drawings, we use a triad task in which subjects are asked to classify a new object given labeled examples of our two categories. This second task allows us to more closely align our findings with the color/size asymmetry noted by Kelemen and Bloom, and allows us to balance the freedom given to subjects in Experiment 1 with a more constrained environment.

Subjects in this task are asked to make classifications given first a single example of each class, and then multiple examples of each class. Use of a single example allows for a relatively pure measure of subjects' prior beliefs concerning what visual features are important to the task, while multiple examples gives us some sense of what (if anything) changes when subjects are given evidence of what features vary across a population of objects. We look for evidence of the interaction between domain and perceived relevance of features suggested by our previous experiment, while also investigating whether or not the domain-general preference for expressing differences in bumpiness persists in a task in

which the two feature types are pitted directly against one another. In terms of our proposed hierarchy, we are looking for influences of Level 3 on Level 2, without examination of Level 1 representations in any form.

Methods

Stimuli For this task, a subset of the original stimuli were used. 4 Draxel/Marcon pairs of different color schemes were selected, with 2 pairs taken from the largest items in the original set, and the remaining 2 pairs being of intermediate size. All of these stimuli appeared at horizontal orientation to ensure that subjects would only consider shape information when making their decisions.

Additionally, two novel stimuli were created for each pair to serve as 'unclassified' items. Each novel item matched both their parents' color and size, but would match their Draxel parent for one feature type (say, elongation), and their Marcon parent for the other (bump frequency, in this example). The two stimuli in each "hybrid" pair were complementary, such that each "Draxel elongation/Marcon bumps" item had a partner with the opposite pattern of feature inheritance.

Subjects 64 subjects participated in this task, drawn from the MIT community.

Procedure Each subject read the same short description of Draxels and Marcons presented in our first task. Subjects were then told that they were being asked to help classify a new object that was either a Draxel or a Marcon, but currently unlabeled. Subjects were told that their initial answer would be based on the observation of only one example each of a Draxel and a Marcon, and that after their first response they would be given multiple examples to look at regardless of their first answer. To ensure that subjects did not feel undue pressure to change their answer given new information, all stimuli to be presented to the subject were laid face-down on the table before any responses were solicited. In this way, we minimize the possibility that subjects' might consider the new stimuli as additional information selected by the experimenter to guide them to a particular answer.

Subjects were first shown one example each of a Draxel and a Marcon, (Figure 5) matched for all attributes except eccentricity and bump frequency. A new item was then shown, drawn from one of the two hybrid stimuli created for that initial pair of stimuli. This third item matched the color and overall scale of both original examples. Subjects were then asked to classify the new item as a Draxel or a Marcon.

After their response was recorded, they were shown the three remaining examples of Draxels and Marcons. Subjects were asked a second time for a response, which was then recorded prior to rewarding the subject with M&M's for volunteering. The initial pair of examples displayed, as well as the particular probe used as the third stimulus were balanced across subjects, as was the left-right arrangement of Draxels and Marcons.

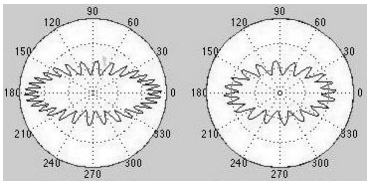


Figure 5: Digital versions of Draxels (left) and Marcons (right). Actual stimuli were hand-drawn, but we present these more regular, schematic images for clarity.

Results

On subjects' first judgments, 21 subjects in the Animals condition choose to categorize the novel stimulus by referring to the bump frequency of the examples, with 11 favoring eccentricity. In the Machines condition, 25 subjects used the bumps as a diagnostic feature, with 7 participants using the elongation of the ellipse. A chi-squared test reveals that these two distributions do not differ from each other. However, an additional goodness-of-fit test shows that the distribution of responses in the Machines condition differs from chance ($p < 0.05$), while the responses of subjects in the Animals condition do not.

Moving on to consider the responses given by subjects after viewing multiple examples, we see that 19 subjects in the Animals condition favor bump frequency for classification, compared to 13 individuals who prefer to use eccentricity. In the Machines condition, the corresponding numbers are 27 and 5 subjects respectively. Unlike the single-example data, these two distributions do differ from one another by a chi-squared test ($p<0.05$). Also, the responses of the Machines subjects differ from chance while the other responses do not.

In each group of subjects, only two subjects changed their mind when presented with multiple examples of the object classes. This indicates that our efforts to minimize undue pressure on the subjects to change answers suceeded, and that subjects were confident in their classifications.

Discussion

In our triad task, we continue to see domain-specific differences in the preference for different aspects of object shape. In correspondence with our results from Experiment 1, we see that subjects in the Animals condition do not significantly prefer one feature type to another. Moreover, as a whole, subjects who perceive the objects as machines have a greater tendency to ignore differences in elongation in favor of differences in bump frequency.

Additionally, in both conditions we see the same overall bias towards using bumps that characterized subjects' drawings in Experiment 1. The distribution of responses in the Animals condition is never different from chance in this study, making it difficult to draw a firm conclusion on this matter. However, the qualitative agreement between the pattern of results obtained here and those obtained in Experiment 1 suggests that the effects we observe genuinely

reflect people's representations of these categories, rather than task artifacts.

General Discussion

Two studies demonstrated that perceptual representations of simple visual shape categories differ depending on whether the stimuli are conceived of as living or non-living things. Our results show a difference in how various shape features are weighted across domains, but they do not yet speak to the origins of those differences. The difference could be simply one of spatial scale, or it could reflect deeper differences in how function or shape variation is conceptualized in different domains. We conjecture that the functional description of the objects plays a driving role in the effects we observed. In a pilot study, we have presented subjects with the same stimuli and task of Experiment 2, eliminating only the instruction sheet's explicit description of bump function. Preliminary results indicate that this manipulation eliminates the domain effects we see here, in that both "animal" and "machine" subjects appear equally to favor the bumps for categorization, and at a level equal to "machine" subjects in Experiment 2. Hence, for these shape stimuli, functional information appears to exert a greater influence when the objects are construed as living kinds, and its primary role appears in *decreasing* the distinctiveness of the bumps for classifying these two kinds of microorganisms. The bumps may be seen as pseudo-pods, non-rigid appendages of the microorganisms that are used to grab compounds but are not essential shape features that are stable across time in an individual, let alone as a distinctive feature for classifying objects into kinds. This interpretation is consistent with our data, although subjects' post-experiment surveys did not mention any explicit reasoning of this sort. Further research is needed to determine precisely how functional knowledge and shape representations interact here, but it is intriguing to speculate that intuitive domain theories are guiding implicit inferences concerning the possible dynamic aspects of simple shapes.

The potential effects of exposing subjects to multiple examples of an object class are also worth exploring further. In this task, we see very little change in subjects' behavior from one response to the other, but the number of observations they are allowed to make is still quite small. If presented with an extremely large population of Marcons and Draxels, subjects might undergo a more profound evolution of shape processing strategies. We note that in the Experiment 1, subjects gained far more experience with Marcons and Draxels than those that participated in our triad task. Subjects in the Animals condition also expressed an explicit preference for elongation rather than bumps in that task, which we did not see in Experiment 2. The difference between these two patterns of response may be related to the size of the observed population of each object class. The possible interaction of statistical reasoning given a population of novel objects and prior beliefs about feature relevance may prove to be a rich area for further research.

Conclusions

Taken together, these studies demonstrate that the categorization and perception of simple visual stimuli may be affected by the domain in which these objects are construed, functional reasoning about object properties, and by the amount of experience one has with a particular stimulus set. Understanding all of these influences, both as separate mechanisms and as a coherent whole, may lead to a richer understanding of human object categorization and the perception-cognition interface.

Acknowledgments

The authors would like to thank the members of the Sinha lab, as well as Erin Conwell, who put up with talking about fictitious unicellular organisms for far too long. An anonymous reviewer also provided many helpful comments. BJB was supported by an NIH pre-doctoral training grant. JBT was supported by the Paul E. Newton chair.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psych. Review, 94,* 115-147.

Bloom, P. (2000). *How Children Learn the Meanings of Words.* Cambridge, MA: MIT Press

Carey, S. (1985). *Conceptual Change in Childhood.* Cambridge, MA: MIT Press.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology, 115(2),* 107-117.

Gauthier, I. & Tarr, M.J. (1997). Becoming a 'greeble' expert: Exploring the face recognition mechanism. *Vision Research, 37(12),* 1673-82.

Goldstone, R.L. (1994) Influence of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123,* 178-200.

Hoffman, D. & Richards, W. (1986). Parts of Recognition. In A.P. Pentland (Ed.), *From Pixels to Predicates,* Norwood, N.J: Ablex Publishing Corporation.

Keil, F.C. (1998). Cognitive science and the origins of theoretical knowledge. In W. Damon & R. Lerner (Eds.), *Theoretical Models of Human Development.* New York, NY: John Wiley & Sons.

Kelemen, D. & Bloom, P. (1994). Domain-specific knowledge in simple categorization tasks. *Psychonomic Bulletin & Review, 1(3),* 390-395.

Landau, B., Smith, L, & Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Science, 2(1),* 19-24.

Maertens, R. & Rousseau, R. (2000). Een nieuwe benaderde formule voor de omtrek van een ellips, *Wiskunde & Onderwijs, 26,* 249-258.

Yin, R.K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81,* 141-145.

Modeling Attachment Decisions with a Probabilistic Parser: The Case of Head Final Structures

Ulrike Baldewein (ulrike@coli.uni-sb.de)
Computational Psycholinguistics, Saarland University
D-66041 Saarbrücken, Germany

Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK

Abstract

We describe an incremental, two-stage probabilistic model of human parsing for German. The model is broad coverage, i.e., it assigns sentence structure to previously unseen text with high accuracy. It also makes incremental predictions of the attachment decisions for PP attachment ambiguities. We test the model against reading time data from the literature and find that it makes correct predictions for verb second sentences; however, the model is not able to account for reading times data for verb final structures because attachment preferences in our training data do not match those determined experimentally. We argue that this points to more general limitations with our type of probabilistic model when it comes to realizing processing strategies that are independent of the data the parsing model is trained on.

Introduction

Experimental results show that human sentence processing is sensitive to different types of frequency information, including verb frame frequencies (e.g., Garnsey et al. 1997), frequencies of morphological forms (e.g., Trueswell 1996), and structural frequencies (e.g., Brysbaert & Mitchell 1996). Probabilistic parsing models are an attractive way of accounting for this fact, as they provide a theoretically sound way of combining different sources of frequency information into a coherent model. Typically, these models are hybrid models, combining symbolic knowledge (e.g., phrase structure rules) with frequency information (e.g., rule probabilities gleaned from a corpus).

In particular, probabilistic parsers have been used successfully to model attachment decisions in human sentence processing. Early models demonstrated the viability of the probabilistic approach by focusing on a small selection of relevant syntactic constructions (Jurafsky 1996; Hale 2001). More recently, *broad coverage* models have been proposed (Crocker & Brants 2000; Sturt et al. 2003) that can deal with unrestricted text. These models are able to account for the ease with which humans understand the vast majority of sentences, while at at the same time making predictions for sentences that trigger processing difficulties.

However, existing probabilistic models deal exclusively with English data, and thus fail to address the challenges posed by the processing of head final constructions in languages such as Japanese (e.g., Kamide & Mitchell 1999) or German (e.g., Konieczny et al. 1997). In this paper, we address this problem by presenting a probabilistic model of human sentence processing in German. The model is broad coverage, i.e., it generates accurate syntactic analyses for unrestricted text. Furthermore, it makes predictions for PP attachment ambiguities for both head initial and head final sentences. The model consists of two probabilistic modules: a syntactic module that proposes an initial attachment, and a semantic module that evaluates the plausibility of the proposed attachment, and corrects it if necessary.

We evaluate our model on reading time data for PP attachment, i.e., for structures in which a prepositional phrase can be attached either to a noun phrase or a verb. In German, PP attachment ambiguities can occur in two syntactic configurations: in verb second sentences, the verb precedes the NP and the PP as it does in English (see (1)).

(1) Iris tröstete den Jungen mit dem Lied.
Iris comforted the boy with the song.
'Iris comforted the boy with the song.'

(2) (daß) Iris den Jungen mit dem Lied tröstete.
(that) Iris the boy with the song comforted.
'(that) Iris comforted the boy with the song.'

In verb final sentences (which occur as subordinate clauses), the NP and the PP precede the verb (see (2)). As sentence processing is incremental, this means that an attachment decision has to be made before parser reaches the verb (and the frequency information associated with it). These structures therefore provide an interesting challenge for probabilistic models of sentence processing.

Reading studies (e.g., Konieczny et al. 1997, whose materials we use) have shown that in verb second sentences, the PP is preferredly attached according to the subcategorization bias of the verb (as in English). In verb final sentences, where verb frame information cannot be accessed until the end of the sentence, the PP is preferentially attached to the NP site.

The Model

Our parsing model consists of two modules: one is a syntactic module based on a probabilistic parser, which also has access to a probabilistic verb frame lexicon. This module guarantees broad coverage of language data and a high accuracy in parsing unseen text. The other module is a semantic module that uses probabilistic information to estimate the plausibility of the analyses proposed by the syntactic module.

The model uses a syntax-first processing strategy: The syntactic module proposes a set of analyses for the input and ranks them by probability. The semantic module then computes the semantic plausibility of the analyses and ranks them by plausibility score. If there is a conflict between the decisions made by the two modules (i.e., the top-ranked analyses differ), this is interpreted as a conflict between syntactic preference and semantic plausibility and increased processing effort is predicted.

Syntactic Module

Modeling Syntactic Preferences The syntactic module consists of a probabilistic left-corner parser which relies on a probabilistic context free grammar (PCFG) as its backbone. A PCFG consists of a set of context-free rules, where each rule $LHS \rightarrow RHS$ is annotated with a probability $P(RHS|LHS)$. This probability represents the likelihood of expanding the category *LHS* to the categories *RHS*. In order to obtain a mathematically sound model, the probabilities for all rules with the same left hand side have to sum to one. The probability of a parse tree T is defined as the product of the probabilities of all rules applied in generating T.

S	→	NE VVFIN.n.p NP PP	.3
S	→	NE VVFIN.n NP	.7
NP	→	ART NN	.4
NP	→	ART NN PP	.6
PP	→	APPR ART NP	1.0
VVFIN.n	→	tröstete	.8
VVFIN.n.p	→	tröstete	.2
ART	→	den	.5
ART	→	dem	.5
NE	→	Iris	1.0
NN	→	Jungen	.6
NN	→	Lied	.4
APPR	→	mit	1.0

Figure 1: Example of a PCFG

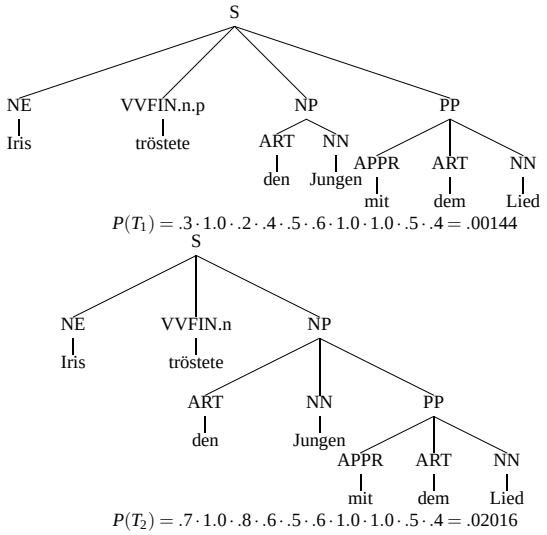


Figure 2: Example of trees generated by a PCFG

An example for a PCFG is given in Figure 1. This grammar contains the rules required to generate the two readings of (1). The readings are displayed in Figure 2, which also lists the parse probabilities, obtained by multiplying the probabilities of the rules used to generated a given tree.

This examples illustrates how PCFGs can be used for disambiguation: the two readings involve different rules (and rule probabilities), and therefore differ in their overall probabilities. In this example, reading T_2 is predicted to be preferred over T_1. Note that the grammar in Figure 1 incorporates verb frame probabilities: *tröstete* 'consoled' can either be a VVFIN.n (finite verb with an NP complement) or VVFIN.n.p (finite verb with an NP and a PP complement). The probabilities attached to these lexical items correspond to the psycholinguistic notion of *verb frame bias*, i.e., the probability of the verb occurring with a given subcategorization frame. The overall probability of an analysis is determined not only by verb frame bias, but also by structural probabilities attached to the phrase structure rules. This is a way of modeling structural disambiguation preferences (in this example, there is a bias for attachment to the NP). A PCFG therefore provides a principled way of integrating lexical preferences and structural preference, as argued by Jurafsky (1996).

Training and Test Data A PCFG is typically trained on a syntactically annotated corpus. For German, a suitable corpus is available in the form of Negra (Skut et al. 1997), a 350,000 word corpus of newspaper text. The Negra annotation scheme assumes flat syntactic structures in order to account for free word order in German. For example, there is no VP node dominating the main verb. Instead, subject, objects and modifiers of the main verb are its sisters, and all are direct daughters of the S node (see Figure 2). This means that scrambling phenomena simply alter the sequence of sisters in the tree, and do not involve movement and traces.

We checked the PP attachment preferences in Negra and found that in 60% of all sentences containing a verb and an NP object followed by a PP, the PP is attached to the verb. The corpus therefore reflects a general attachment preference for verb attachment. Additionally, we found that the subcategorization preferences for the verbs in our materials were reversed with regard to the preferences obtained by Konieczny et al. (1997) in a sentence completion task: the verbs that had a bias towards the NP-PP frame in the corpus exhibited an NP frame bias in the completion study, and vice versa.

For all subsequent experiments, Negra was split into three subsets: the first 90% of the corpus were used as training set, the remainder was divided into a 5% test set and a 5% development set (used during model development). Sentences with more than 40 words were removed from the data sets (to increase parsing efficiency).

The syntactic module was realized based on Lopar (Schmid 2000), a probabilistic parser using a left-corner parsing strategy. A grammar and a lexicon were read off the Negra training set, after empty categories and function labels had been removed from the trees. Then the parameters for the model were estimated using maximum likelihood estimation. This means that the probability of a rule $LHS \rightarrow RHS$ is estimated as $P(LHS \rightarrow RHS) = f(LHS \rightarrow RHS)/N$, which is the number of times the rule occurs in the training data over the total number of rules in the training data. Various smoothing schemes are implemented in Lopar to address data sparseness, see Schmid (2000) for details. We also complemented the Negra verb frame counts with frame probabilities from an existing subcategorization lexicon (Schulte im Walde 2002), as the Negra counts were sparse.

Semantic Module

Modeling Semantic Plausibility The semantic module determines whether an attachment decision proposed by the syntactic module is semantically plausible by deciding whether the PP is more likely to be semantically related to the preceding verb or to the preceding noun.

Our semantic model rests on the assumption that "semantic plausibility" or "semantic relatedness" can be approximated by probabilistic measures estimated from corpus frequencies. Previous work provided evidence for this assumption by demonstrating that co-occurrence frequencies obtained from various corpora (including the web) are reliability correlated with human plausibility judgments (Keller & Lapata 2003).

Training and Test Data Ideally, the same training data should be used for the syntactic and the semantic module; however, this was not possible, as the semantic module requires vastly more training data. We therefore used the web to estimate the parameters of the frequency based measures (see Keller & Lapata 2003 for a detailed evaluation of the reliability of web counts). For the selectional preference method, we used one year's worth of text from the ECI Frankfurter Rundschau corpus as training data. This unannotated corpus is the basis for the Negra corpus, but it is much larger (34 million words). The corpus was parsed using a parser very similar to the syntactic module. Tuples of verbs and head nouns of modifying PPs were then extracted according to the structures assigned by the parser.

The development and test set for the semantic module were taken from the set of 156 items from Experiments 1 and 2 of Konieczny et al. (1997). The development set consists of 68 randomly chosen sentences, the remaining 88 sentences are used as a test set. The items from Experiment 1 vary word order (verb second and verb final), verb subcategorization preference (bias for NP frame or for NP-PP frame), and attachment (to the NP or verb), which is disambiguated by the semantic implausibility of one alternative. In Experiment 2, verb subcategorization preference was not varied. The development set was used to compare the performance of different semantic and syntactic models and to set the parameters for one semantic models. The final performance will be reported on the unseen test set.[1]

Plausibility Measures In computational linguistics, a standard approach to PP attachment disambiguation is the use of configuration counts from corpora (e.g., Hindle & Rooth 1991). To decide the attachment of n_{PP}, the head noun of the PP, to one of the attachment sites (the verb v or n_{NP}, the noun phrase), we compare how probable each attachment is based on previously seen configurations involving n_{PP} and the attachment sites. In many approaches, the the preposition p is also taken into account.

As outlined above, we used web counts to mitigate the data sparseness that such a model is faced with. In this approach, corpus queries are replaced by queries to a search engine, based on the assumption that the number of hits that the search engine returns is an approximation of the web frequency of the word in question (Keller & Lapata 2003). Of course text on the web is not parsed, which makes it difficult to identify the correct syntactic configurations. We follow Volk (2001) in assuming that string adjacency approximates syntactic attachment reasonably well, and simply use queries of the form `"V PP"` and `"NP PP"`. The search engines used were `www.altavista.de` and `www.google.com` (restricted to German data). Google generally outperformed AltaVista (presumably because it indexes more pages); the results reported below were obtained using Google counts.

We experimented with a variety of plausibility measures (*site* ranges over the two attachment sites, v and n_{NP}):

(a) $\frac{f(site,p)}{f(site)}$, the Lexical Association Score (LA), computes how likely the attachment site is to be modified by a PP with the preposition p.

(b) $f(site,p,n_{\text{PP}})$, Model 1 of Volk (2001), relies on the raw trigram co-occurrence frequencies to decide attachment.

(c) $\frac{f(site,p,n_{\text{PP}})}{f(site)}$, Model 2 of Volk (2001), takes into account that high-frequency attachment sites are more likely to co-occur with PPs.

(d) $log_2\left(\frac{f(site,n_{\text{PP}})}{f(site)f(n_{\text{PP}})}\right)$, Pointwise Mutual Information (MI) measures how much information about one of the items is gained when the other is seen. This measure has previously been used for the related problem of identifying collocations (words that appear together more often than chance, Church & Hanks 1990).

(e) $\frac{f(site,n_{\text{PP}})}{f(site)} \cdot \frac{f(site,n_{\text{PP}})}{f(n_{\text{PP}})}$, Combined Conditional Probabilities (CCP) is similar to MI. It squares the joint probability term to give it more weight.

[1]Since for all models except one no parameters were set on the development set, we had to maintain a fixed development-test split to ensure the test set remained truly unseen.

As will be explained below, we experimented with these measures in isolation, but we also combined them with Clark & Weir's (2002) approach for computing selectional preference from corpora. This approach relies on a lexical data base to compute the semantic similarity between lexical items.

Results

Syntactic Module

As mentioned in the introduction, the present modeling effort was guided by the idea of building a broad coverage model, i.e., a model that explains why human sentence processing is effortless and highly accurate for the vast majority of sentences; at the same time, the model should account for psycholinguistically interesting phenomena such as processing difficulties arising from attachment ambiguities. Incrementality is crucial for predictions of this type. In its original form, the Lopar parser used for the syntactic module is not incremental and was therefore modified to achieve partial incrementality. It now outputs its ranking of the attachment alternatives in two stages: after processing the PP and at the end of the sentence. This provides a record of incremental changes in the attachment preferences of the model when processing the critical region for which Konieczny et al. (1997) report eye-movement data (the noun of the PP in Experiment 1 and the PP object in Experiment 2).

To evaluate the broad coverage of the model, we ran the syntactic module on our unseen Negra corpus test set. The model was able to assign an analysis to 98% of the sentences. As is standard in computational linguistics, we tested the accuracy of the model by measuring labeled bracketing: to score a hit, the parser has to predict both the bracket (the beginning or end of a phrase) and the category label correctly. We report labeled recall, the number of correctly labeled brackets found by the parser divided by the total number of labeled brackets in the test corpus, and labeled precision, the number of correctly labeled brackets found by the parser divided by the total number of labeled brackets found by the parser.

The model achieved a labeled recall of 66.65% and a labeled precision of 63.92%. It is similar to the baseline model of Dubey & Keller (2003), who report a maximum labeled recall and precision of 71.32% and 70.93%.

To further evaluate the syntactic model, we tested it on the test set generated from Experiments 1 and 2 of Konieczny et al. (1997). This allows us to determine whether the syntactic module is able to correctly resolve the PP attachment ambiguities even without access to any semantic information.

Table 1 shows the parser's decisions at the PP for verb final and verb second sentences. We report the number and the percentage of correct attachments per condition. In the verb final condition of Experiment 1, the parser always attached the PP to the verb. No verb frame information is available to guide the decision when the PP is processed, so the baseline is random guessing (50%). In verb second sentences, the parser can use the subcategorization preference of the verb, which leads to the correct attachment in 50% of all cases. The parser indeed reaches this baseline. In Experiment 2, the parser again always attaches the PP to the unseen verb in verb final sentences. In the verb second condition, there is a marked preference to attach according to verb bias, but only 42% of attachments are correct over both conditions.

	Verb final	Verb second
Experiment 1		
NP frame, V bias	7 (100%)	2 (29%)
NP frame, NP bias	0	5 (83%)
NP-PP frame, V bias	5 (100%)	3 (60%)
NP-PP frame, NP bias	0	2 (33%)
% correct	50%	50%
Experiment 2		
NP frame, V bias	9 (100%)	1 (11%)
NP frame, NP bias	0	5 (56%)
% correct	50%	33%
Baseline	50%	50%

Table 1: Syntactic module: correct attachment decisions at the PP for the test set from Experiments 1 and 2

Measure	CCP	MI	LA	Volk 1	Volk 2
Development Set					
# correct	23	22	17	22	21
% correct	67.6%	64.7%	50%	64.7%	61.7%
Test Set					
# correct	21	23	–	23	27
% correct	50%	54.8%	–	54.8%	64.3%
Baseline	50%	50%	50%	50%	50%

Table 2: Semantic module, verb second: results of the plausibility measures on the development and test set

Semantic Module

Verb Second Sentences As a next step, we evaluated the semantic module, again on the data derived from Experiments 1 and 2 of Konieczny et al. (1997). We again used the chance baseline (50%) that the syntactic module was unable to outperform.

The verb second sentences arguably constitute the standard case for PP attachment: Both possible attachment sites have been seen before the attachment has to be decided. In a first attempt, the five plausibility measures introduced above were tested on the development set. Table 2 shows that the CCP measure performed best, while the Lexical Association measure failed to beat the baseline. The CCP measure should therefore be chosen to model semantic attachment in verb second sentences. However, on the test set (see Table 2), the best and worst measure changed places. This time, the Volk 2 measure performed best. No measure significantly outperformed the others or the baseline.

As the performance of the CCP measure on the test set was disappointing, we experimented with a second approach that combines the Volk 2 model of PP attachment with a model of selectional restrictions. We used Clark & Weir's (2002) approach, which was extended to German by Brockmann & Lapata (2003), whose implementation we used. Relying on a semantic hierarchy (in our case: GermaNet, Hamp & Feldweg (1997)), the Clark & Weir algorithm finds the statistically optimal superclass (concept) for input nouns given a verb and the relation between noun and verb. The probability of a concept *c* given a verb *v* and relation *rel* is computed as:

$$(3) \quad P(c|v,rel) = P(v|c,rel)\frac{P(c|rel)}{P(v|rel)}$$

To find the best concept for a $\langle n,v,rel\rangle$ triple, at each step up the hierarchy, the probability estimate for the new concept is compared to that of the original concept. When the estimates differ significantly, the lower concept is assumed for the noun. The parameters of this algorithm are the statistical test used (χ^2 or G^2) and the α value which determines the level of significance required for the test. The G^2 test proved more suitable for our task, while a variation of α value had no noticeable effect.

	Prior	Average			
Measure	CCP	CCP	MI	Volk1	Volk2
Development Set					
# correct	12	11	8	9	11
% correct	60%	55%	40%	45%	55%
Test Set					
# correct	24	20	22	23	25
% correct	57.1%	47.6%	52.4%	54.8%	59.5%
Baseline	50%	50%	50%	50%	50%

Table 3: Semantic Module, verb final: Results on the development (Exp. 1) and test set (Exp. 1 and 2)

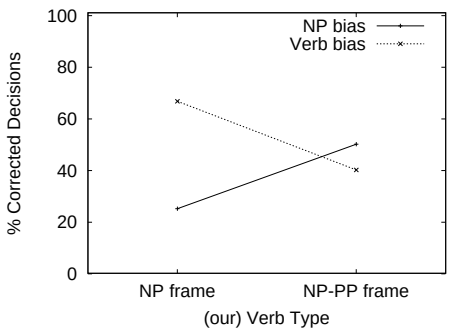
We used the development set to estimate a threshold value for the attachment decision. Coverage on the test set was only 48% due to sparse data. Whenever the Clark & Weir method did not return a value, we backed off to the decision made by the Volk 2 model (which is the most consistently performing model). Recall that this model has a 64% precision on the test data while the chance baseline is 50%. The combined model reaches 67% precision on the same data (precision for the selectional preference model alone is 70% for 48% of the data). This model performs best numerically (though not significantly so) and was used in the final model.

Verb Final Sentences A particularly interesting case arises with respect to verb final sentences (see (2)): at the critical region (once the PP has been processed), the verb is not available yet, which means that the plausibility of the combination of the verb with the head noun of the PP cannot be computed at this point. Konieczny et al. (1997) found processing difficulty in these cases when the PP was an implausible modifier of the noun, so apparently immediate semantic evaluation sets in and has to be accounted for.

In the verb final case, we therefore have to estimate the plausibility of the PP head noun modifying the NP as opposed to an unseen verb. One way of doing this is to average over the results for the PP head noun and every possible verb to obtain a generic value for verb attachment. We restrict ourselves to just the verbs in the test and development set. This backoff approach was realized for four models.

An alternative is to use the prior probability of the PP head noun as an estimate of its conditional probability with every possible verb. The prior probability of the PP head noun is its frequency divided by the size of the corpus, $f(n_{\mathrm{PP}}/N)$. In the case of web counts, N is the number of all documents searched. This figure was empirically estimated as proposed by Keller & Lapata (2003). Note that this method of backoff is possible only for the CCP measure, because the probabilities to be estimated for the other methods are too complex.

Table 3 gives the results on the development set for the items from Experiment 1. The items from Experiment 2 could not be tested as the averaging procedure is extremely costly in terms of web queries. The CCP model with simple backoff to the prior shows the best results at 60% correct attachments. We therefore used it for the final evaluation to predict attachments for verb final sentences. Table 3 also lists the results on the test set for items from Experiments 1 and 2. CCP with backoff to the prior performed better than most models that use averaging, and substantially outperforms CCP with averaging. The best model is Volk 2 with averaging. Again, no measure outperforms the baseline of 50% correct attachments or the other models.



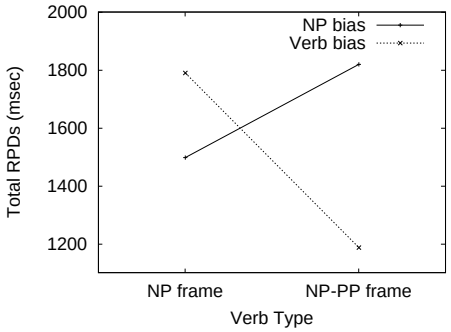
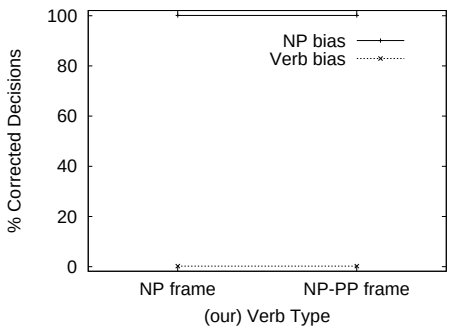


Figure 3: Exp. 1, verb second: Predictions of the combined model (left) compared to the Konieczny et al. (1997) data (right)



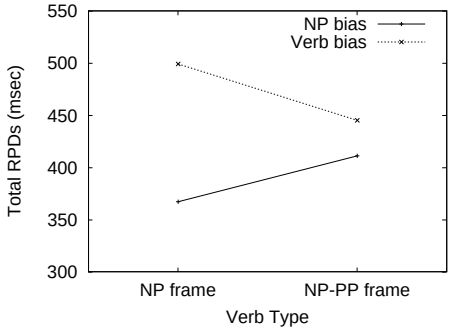


Figure 4: Exp. 1, verb final: Predictions of the CCP/Prior model (left) compared to the Konieczny et al. (1997) data (right)

Combined Model

In the previous sections, we evaluated the syntactic and semantic module separately. We found that the syntactic module performs at the level of the chance baseline of 50%, while the semantic module achieves an accuracy of up to 67% for verb initial sentences and 60% for the verb final sentences. A more interesting question is how well the model accounts for the processing difficulties that are evident in the eye-movement data reported by Konieczny et al. (1997). As mentioned at the beginning of the Results Section, our model makes predictions for the critical region used by Konieczny et al. (1997) (the PP). Recall also that we assume that a conflict between syntactic preference and semantic plausibility predicts increased processing effort.

As explained in the section on Training and Test Data above, the subcategorization variable was reversed for our data: where Konieczny et al. (1997) assume an NP-PP frame bias, we found a preference for the NP frame in our corpus (and vice versa). Below, our model's predictions are labeled with the preferences found in our data, while data from Konieczny et al. (1997) are labeled with the preferences they found. Figure 3 compares the predictions of our model with Konieczny et al.'s results in Experiment 1 for verb second sentences.[2] The graph for our model gives the percentage of correct decisions by the semantic module that are in conflict with the the decisions of the syntactic module. Such conflicts predict longer reading times, and the more conflicts in a condition, the higher we expect the average reading times to be. The figure shows that our model predicts the data pattern found by Konieczny et al. (1997) (who report regression path durations, RPDs).

[2]Note that our results are on the unseen subset of the items only, while the reading times are on all items.

In verb final sentences (Figure 4), the syntactic module always predicts verb attachment, so correct decisions for NP attachment by the semantic module always lead to a conflict. This pattern does not correspond to the Konieczny et al. (1997) reading data, which show a general preference to attach to the NP. Figure 5 shows a replication in principle of the reading time data in the verb second case. In Konieczny et al.'s (1997) pretests, all the verbs subcategorized for an NP and a PP, while in our data, they preferredly subcategorize for just an NP. Our model predicts longer reading times for the NP frame when subcategorization preference and semantic disambiguation are mismatched, which is what Konieczny et al.'s (1997) show for the NP-PP frame. The verb final case again fails: Instead of predicting preferred attachment to the NP (Matched bias for our data, Mismatched bias for Konieczny et al.'s data), the model predicts verb attachment.

Discussion

While our model replicates Konieczny et al.'s (1997) reading time results for PP attachment in the verb second case, it fails to account for reading times of verb final sentences. This failure is caused by the syntactic module which always predicts verb attachment in verb final sentences, while there is a human preference for NP attachment in these cases.

The behavior of the syntactic module is influenced by two factors. One is the probability of phrasal rules such as S → NE VVFIN.n.p NP PP. The second factor is a verb-specific frame bias, which manifests itself as probabilities for lexical rules such as VVFIN.n.p → *tröstete*. In verb second sentences, the verb's frame probability together with the phrasal rule probability determines the analysis proposed by the syntactic module. In verb final sentences, however, only the phrasal probabilities are used (as the verb is not yet avail-

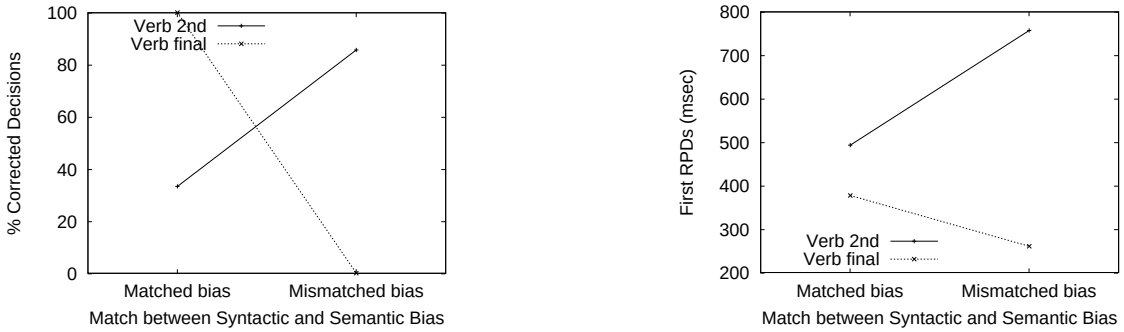


Figure 5: Exp. 2: Predictions of the combined model (left) compared to the Konieczny et al. (1997) data (right). Verbs subcategorize for an NP frame in our data and for an NP-PP frame in the Konieczny et al. data.

able), so the syntactic module makes the same prediction for all verb final sentences. This prediction is incorrect because the general PP attachment bias in the corpus is to the verb, rather than to the NP as in the reading time data.

This points to a more general problem with probabilistic models: They can only be as good as the training data. It is therefore vital to check relevant properties of the training corpus in comparison to experimental data when developing probabilistic models. Balanced corpora that consist of language data from different sources are more reliable in this respect than newspaper corpora such as the Negra corpus.

This means that the failure to model the verb final data is not a failure of probabilistic models per se; our approach would be in principle capable of modeling the general attachment preference to the NP in verb final sentences, if the attachment preference in the training data corresponded to that in the experimental results. Thus, our results strengthen the case for probabilistic models by showing that they can be applied even to head final constructions.

It is important to note, however, that our explanation of the German PP attachment data in terms of biases in the training corpus is at variance with explanations in the literature. For instance, Konieczny et al. (1997) proposes a strategy of *Parameterized Head Attachment* to explain why the parser prefers to attach incoming material (such as the PP) to existing sites (such as the verb). This strategy, which aims at the immediate semantic evaluation of the input, is designed to cope with head final structures in general, not only in the case of PP attachment. A basic PCFG model such as the one used in this paper is not able to implement such a general strategy.

Conclusions

We have presented a two-stage model parsing model that accounts for PP attachment in German. The model is able to assign correct sentence structures to unseen text and predicts average reading times in verb second sentences. For verb final sentences, the model fails to correctly predict the reading time data. The reason is that our training corpus exhibits a general bias for attaching PPs to the wrong attachment site (to the verb instead of the NP). In principle, however, our model would be able to account for the data in the verb final case if the training data were consistent with experimental findings. Our findings therefore strengthen the case for probabilistic models of language processing by showing their applicability to head final structures. At the same time, they demonstrate that probabilistic models can be highly sensitive to idiosyncrasies in the training data.

References

Brockmann, C., & Lapata, M. (2003). Evaluating and combining approaches to selectional preference acquisition. In *Proc. EACL*, (pp. 27–34), Budapest.

Brysbaert, M., & Mitchell, D. C. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *Quaterly J. of Experimental Psychology, 49A*, 664–695.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Clark, S., & Weir, D. (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, *28*, 187–206.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *J. of Psycholinguistic Research*, *29*, 647–669.

Dubey, A., & Keller, F. (2003). Probabilistic parsing for German using sister-head dependencies. In *Proc. ACL*, (pp. 96–103), Sapporo.

Garnsey, S. M., Pearlmutter, N. J., Myers, E. M., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. of Memory and Language, 37*, 58–93.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proc. NAACL*, Pittsburgh, PA.

Hamp, B., & Feldweg, H. (1997). GermaNet: A lexical-semantic net for German. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, & Y. Wilks (eds.), *Proc. ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, (pp. 9–15), Madrid.

Hindle, D., & Rooth, M. (1991). Structural ambiguity and lexical relations. In *Proc. ACL*, (pp. 229–236), Berkeley, CA.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes, 14*, 631–662.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics, 29*, 459–484.

Konieczny, L., Hemforth, B., Scheepers, C., & Strube, G. (1997). The role of lexical heads in parsing: Evidence from German. *Language and Cognitive Processes, 12*, 307–348.

Schmid, H. (2000). LoPar: Design and implementation. Unpubl. ms., IMS, University of Stuttgart.

Schulte im Walde, S. (2002). A subcategorisation lexicon for German verbs induced from a Lexicalised PCFG. In *Proc. LREC*, vol. IV, (pp. 1351–1357), Las Palmas, Gran Canaria.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proc. ANLP*, Washington, DC.

Sturt, P., Costa, F., Lombardo, V., & Frasconi, P. (2003). Learning first-pass structural attachment preferences with dynamic grammars and recursive neural nets. *Cognition, 88*, 133–169.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *J. of Memory and Language, 35*, 566–585.

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proc. Corpus Linguistics*, Lancaster.

Mapping individuation to mass-count syntax in language acquisition

David Barner (barner@fas.harvard.edu)
Department of Psychology, William James Hall, 33 Kirkland Street
Cambridge, MA 02138 USA

Jesse Snedeker (snedeker@wjh.harvard.edu)
Department of Psychology, William James Hall, 33 Kirkland Street
Cambridge, MA 02138 USA

Abstract

Various theories propose that count nouns are distinguished from mass nouns by their specification of individuation. We present evidence that, while 3-year-old children acquiring language extend words differentially on the basis of mass-count syntax, they quantify over individuals for both novel mass and count nouns. We suggest that children may begin acquisition with an underspecified representation of mass noun semantics, permitting quantification over both individuals and continuous quantities. Also, children may rely on ontologically based biases to guide quantification.

Introduction

In English, the specification of number sends a ripple through the language, determining a word's status as mass or count, licensing the use of plural morphology, and selecting among measure terms such as *many* and *much*. Children begin to show signs of such knowledge early in language acquisition, with plural production and comprehension emerging as early as 2 years of age (Brown, 1973; Cazden, 1968; Ferenz & Prasada, 2002; Gordon, 1988), sensitivity to the mass-count distinction evidenced by around 2;6 years (see Soja, 1992), and the use of measure terms like *more* and *less* emerging between 2;6 and 3 years of age (Donaldson & Balfour, 1968; Gathercole, 1985; Palermo, 1973).

The mass-count distinction provides a particularly interesting case of number specification, because it entrains important consequences for both the syntax and semantics of noun phrases. For example, count terms like *cat* can take the plural morpheme (e.g., *cats*) follow cardinal numbers (e.g., *one cat, two cats, three cats*), and be modified by quantifiers like *these, those, few* and *many* (e.g., *many cats*). Mass terms, on the other hand, can occur in none of these environments and can be distinguished by their use with terms like *much* and *little* (e.g., *I don't eat much porridge*). According to most accounts, this distributional difference corresponds to a semantic distinction whereby count nouns quantify over individuals and mass nouns quantify over non-individuals (e.g., Bloom, 1994, 1999; Gordon, 1985; Link, 1998; Quine, 1960; Wisniewski, Casey, & Imai, 1996). For example, Bloom proposed that children might begin acquisition with bidirectional mappings between syntax and semantics, as in (1), resulting in a categorical effect of number specification in noun phrases:

(1) a. count noun ↔ individual
b. mass noun ↔ non-individual

However, such mappings may not provide the full story of what children know about the mass-count distinction. As noted by Barner and Snedeker (2004), mass syntax may not have a strong interpretation like count syntax, but may be semantically unspecified and allow reference to either individuals or non-individuals. In their study, Barner and Snedeker observed that 4-year-old children and adults interpreted many mass and count nouns in the way predicted by the mappings in (1). For example, participants judged three tiny shoes to be *more shoes* than one giant shoe and one giant portion of butter to be *more butter* than three tiny portions. However, they also based judgments on number for "object-mass" terms like *furniture, jewelry, mail,* and *clothing*, which children begin to produce by around 4 years of age. In each case, participants judged six tiny objects to be more than two giant ones. Thus, count syntax led to quantification on the basis of number, while mass syntax quantified over both continuous quantities and certain individual objects.[1] As a result, it was concluded that object-mass terms like *furniture* must allow quantification over individuals due to a lexical specification of number, which is normally found in count syntax. Expressions that house this feature quantify over individuals, while those that do not assume a default interpretation of quantifying over a continuous extent (see Borer, 2004, for a similar proposal). This view is schematized in (2):

(2) a. count syntax → individual
b. mass syntax ↔ no number specification

If the dimension of measurement of mass terms is specified in part by lexical semantics as these results suggest, children may need some amount of experience with

[1] But not all objects. Terms that can be used as either mass or count (e.g., *string/s, chocolate/s, paper/s, stone/s*) quantified by number as count nouns and by continuous extent as mass nouns.

particular mass terms before using them to specify a measuring dimension. This raises the question of what mass syntax contributes to the early interpretation of noun phrases, and whether children's interpretation of mass nouns differs from that of adults at any stage of development. For example, before acquiring exceptions like *mail* and *furniture*, do children respect mappings as in (1), or is mass syntax semantically unspecified throughout development?

Two sources of evidence regarding this question suggest conflicting conclusions. First, studies of word extension indicate that children are biased to map novel count terms to physical objects and mass terms to non-solid substances. In a study by Soja (1992), English children aged 2;6 extended novel words on the basis of shape for solid objects 90% of the time when presented with count syntax, but only 76% of the time with mass syntax. For non-solid substances, children extended novel words on the basis of substance 91% of the time when presented with mass syntax, and 51% of the time when given count syntax (see also Soja, Carey, & Spelke, 1991, and Imai & Gentner, 1997). Thus, children in her study shifted their extension of novel terms according to their use in mass or count syntax.

However, results from Gathercole (1985) suggest that young children may not distinguish the referential consequences of mass and count syntax. In her study, Gathercole found that children aged between 2;6 and 5;6 quantified mostly by number for both count nouns and mass nouns. In fact, even children as old as 5;6 failed to reliably quantify by continuous extent for mass nouns, unlike the 4-year-olds in Barner and Snedeker (2004), who quantified by continuous extent for both solid and non-solid stimuli named with mass syntax (e.g., *some string*; *some mustard*). These differences are difficult to interpret, since only Gathercole tested both mass and count terms within subjects, and because stimuli used in the studies were common household objects and may have varied in familiarity and lexico-semantic properties.

In any case, the results from Gathercole's study are difficult to reconcile with those from Soja's study of word extension, and thus it remains unclear how children represent the semantics of the mass-count distinction early in acquisition. Also unclear is how the different types of knowledge used in each task are related in young children's linguistic representations. While for adults it seems necessary that word extension should predict quantification (e.g., only words that refer to discrete things quantify over individuals) such links between content and quantification may not yet be established in the minds of 2 or 3-year-olds.

No previous study has explicitly tested the quantification of mass-count syntax for novel terms (i.e. where prior lexical knowledge does not play a role). As a result, previous studies have also not examined the relationship between quantification judgments and word extension for the same objects and substances. However, both of these measures are needed in order to properly assess children's early interpretation of mass-count syntax, before lexical exceptions such as object-mass terms (e.g., *mail, silverware*) arise. For this reason, the present study assessed children's and adult's interpretation of novel mass and count terms in both word extension and quantity judgment paradigms.

Also, it remains an open question how object-mass terms come to quantify over individuals, and what the precise nature of lexical information is that distinguishes mass nouns like *mail* from mass nouns like *string*. Various researchers have suggested that factors such as complexity of structure, occurrence of multiple individuals in spatio-temporal contiguity (Wisniewski, Casey, & Imai, 1997), and shared function (Prasada, 1999) might characterize object-mass terms. The present study examined this question via the manipulation of stimulus complexity and solidity in each of the testing paradigms.

Experiment 1

The first experiment examined three main questions. First, does mass syntax have a strong interpretation early in acquisition, as measured by both word extension and quantity judgment? To examine this question, participants were tested with both methods for the same novel objects, with either mass or count syntax. Second, using this method we explored how the word extension and quantity judgment tasks are related, and whether they make use of the same underlying logical resources. Third and finally, we explored whether an object's relative complexity predicts quantification over individuals when used in mass syntax. What lexical properties, if any, might characterize object-mass nouns like *mail* and *furniture*? This question was tested by varying the shapes and substances of novel referents in the word extension and quantity judgment tasks, to include simple and complex solid objects.

Method

Subjects

Participants were 24 Harvard undergraduates and 32 children ranging in age from 3;0 to 3;6 (mean = 3;3).

Procedures and Stimuli

Each testing session comprised four trials. In each trial the participant was introduced to a novel object and heard the object named with a novel term at least four times using either unambiguous mass syntax or unambiguous count syntax on all four trials. Half of the participants were shown four simple objects and half were shown four complex objects, all of which we will call "standard" objects. The four simple standard objects were: (1) a half egg shape made of red sculpy; (2) a kidney bean shape made from painted-green das; (3) a cork shape made from black Crayola-Magic; (4) an arrow shape made from terracotta. The four complex standard objects were: (1) a gear made from orange playdo; (2) a brass t-shaped plumbing fixture; (3) a suede texture-painted reamer; (4) a clay milk pump stand. The names for these objects, always presented with either mass or count syntax (between subjects), were *fem, tannin, dak,* and *tulver*. Thus, for each trial, the

experimenter introduced the novel object by saying, for example "Oh look, this is some/a fem. Have you ever seen any fem(s) before? Do you think you have some fem(s) at home? That is some/a nice fem isn't it".

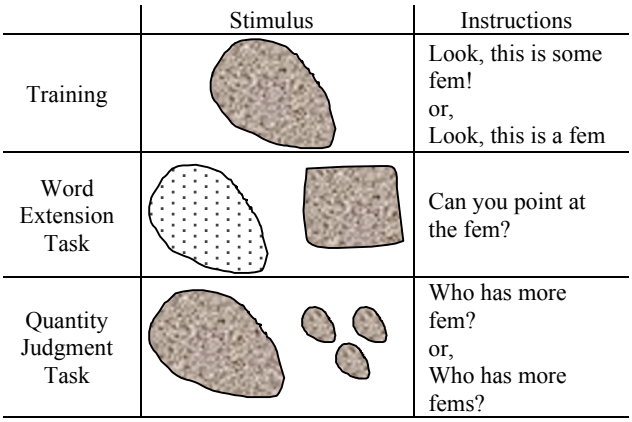


Figure 1. An example of a training trial, word extension trial and quantity judgment trial (shaded oval represents a red sculpy half-egg, white oval represents a styrofoam shape alternative, and shaded square represents a red sculpy substance alternative; small ovals represent mini half-eggs).

Following the naming of each object, participants were asked two questions, the order of which was varied systematically (see Figure 1). First, in the word extension task (see Soja, Carey, & Spelke, 1991), participants were shown two additional objects, one which matched the standard in shape, the other in substance, and were asked to choose which of the two the novel word named: "Show me some/a fem". Second, in the quantity judgment task (see Barner & Snedeker, 2004), participants were shown two characters (Farmer Brown and Captain Blue), one who was shown with the standard object and the other who was shown with three miniature versions of the object. The standard objects had a greater overall mass and volume than the three miniature objects, but were otherwise identical in shape and substance. The side on which the standard or miniatures were shown was varied systematically. Participants were told, "Farmer Brown has some/a fem(s) and Captain Blue has some/a fem(s) too. Who do you think has more fem(s)?" For both tasks, participants pointed to indicate their response. Procedures were identical for adults and children.

Results and Discussion

Word extension trials

Responses for the word extension task were coded in terms of how many times (out of four) each participant extended a word on the basis of shape.

For adults, two main results were obtained (see Figure 2). First, adults used mass-count syntax to guide word extension, extending count terms by shape 3.5 times on average (87.5% overall), compared to only 1.08 times on average (27% overall) for mass syntax, F (1, 16) = 31.1, p < .001. Second, adults also showed a main effect of stimulus type, extending novel words for complex objects 2.83 times on average (71.8%), compared to 1.75 times on average (43.8% overall) for simple objects, F (1, 16) = 6.3, p < .05. There were no interactions.

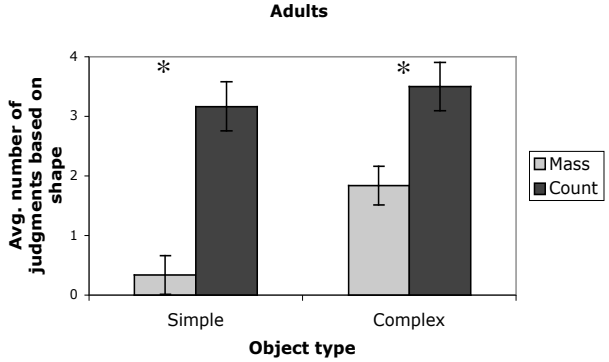


Figure 2. Word extension by adults for novel mass and count nouns.

These results indicate that adults use syntax to guide word extension, but that extension is also influenced to a large extent by the relative complexity of stimuli. Thus, mass-count syntax does not appear to be the sole factor determining word extension behavior for adults. More general properties of referents appear to play a role.

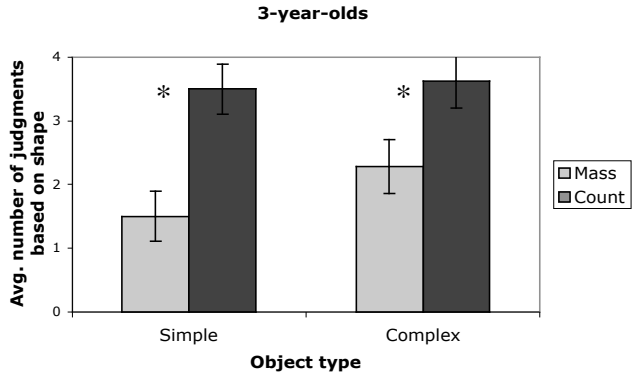


Figure 3. Word extension by 3-year-old children for novel mass and count nouns.

The children also used mass-count syntax to guide word extension (see Figure 3). They extended novel count terms by shape 3.56 times on average (89% overall), compared to 2 times on average (50% overall) for novel mass nouns, F (1, 24) = 16.9, p < .001. However, children showed no main effect of stimulus type, and no significant interactions.

Quantity judgment trials

Responses for the quantity judgment task were coded in terms of how many times (out of four) each participant chose the array with the greater number of objects. For adults, two main results were obtained (see Figure 4). First, as was the case with word extension, adults used mass-count

syntax to guide quantity judgment, using count terms to quantify by number 3.75 times on average (93.8% overall), compared to .67 times on average (16.8%) for mass nouns, F (1, 16) = 85.6, $p < .001$. Second, adults quantified by number 2.67 times on average (66.8% overall) for complex items, which was significantly more than the 1.75 times on average (43.8%) for simple items, F (1,16) = 7.6, $p < .05$. This difference appeared to be due in part to adult's interpretation of mass syntax. While adults quantified by number for some mass nouns that named complex objects, they never did so for mass nouns that named simple objects. This is interesting, since it suggests that for adults object complexity may be sufficient to permit individuation using mass syntax (i.e. the use of object-mass terms).

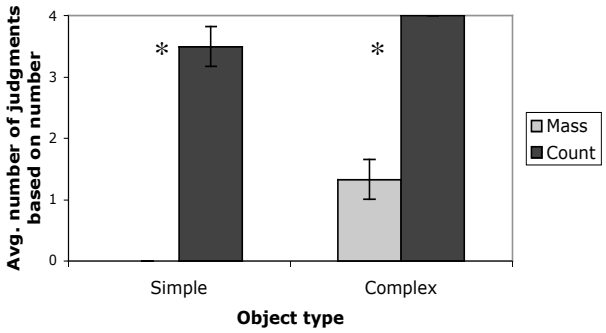


Figure 4. Quantity judgment by adults for novel mass and count nouns.

In contrast to adults, children quantified mostly by number for both simple and complex objects, for both mass and count syntax (see Figure 5). As a result, children showed no significant effect of syntax, F (1, 24) = 2.2, $p > .1$, nor of stimulus type, F (1, 24) = .09, $p > .8$.

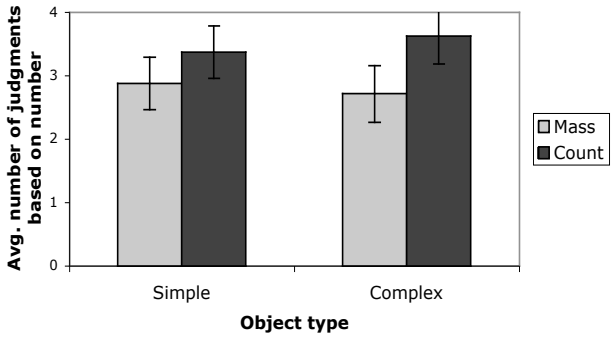


Figure 5. Quantity judgment by 3-year-old children for novel mass and count nouns.

The results of Experiment 1 indicate that while both children and adults used mass-count syntax to guide word extension, only adults appeared to use this information to guide quantity judgment. This suggests that word extension and quantity judgment may tap discrete logical resources. Among the possible explanations for this are: (1) an incomplete understanding of mass-count syntax, (2) a problem interpreting the term "more", or (3) a bias to quantify by number for discrete physical objects. These possibilities are examined in Experiment 2.

Experiment 2

The results of Experiment 1 are consistent with the idea that the interpretation of mass syntax is unspecified from early in acquisition, and allows quantification over individuals or non-individuals. However, the results are also consistent with several other possibilities, including a response bias to quantify by number regardless of syntax, or an interpretation of "more" as quantifying only by number. To rule out these possibilities, we tested participants with non-solid substances. If responses in Experiment 1 were due to either a response bias or a strong interpretation of "more", then responding on the basis of number should persist for both mass and count nouns. However, a change in children's quantification for only mass syntax would represent evidence that children modulate judgments based on mass-count information, but allow quantification over individuals when referents are construed as such. This, in turn, would support the claim that mass syntax, but not count syntax, has an unspecified interpretation regarding individuation.

Method

Subjects

Participants were 16 Harvard undergraduates and 23 children ranging in age from 3;0 to 3;6 (mean = 3;3).

Procedures and Stimuli

Procedures for Experiment 2 were identical to those used in Experiment 1. However, solid stimuli were replaced with non-solid substances. The standard substances were: red media mixer, green butter, orange paint, and brown hair gel.

Results and Discussion

Results suggest that only the adults used mass-count syntax to guide word extension (see Figure 6). Adults extended count nouns by shape 2.57 times on average (64% overall) and mass nouns 0.5 times on average (13% overall), a difference that was marginally significant, F (1, 12) = 4.8, $p < .06$. Children extended count nouns by shape 1.82 times on average (46% overall) and mass nouns 1.42 times on average (36% overall), which was not significant.

However, both the 3-year-olds and adults appeared to use mass-count syntax to guide quantity judgment (see Figure 7). Adults based judgments on number 3.83 times on average for count syntax (96% overall) and 0.5 times on average for mass syntax (13% overall). This difference was significant, F (1, 12) = 40, $p < .001$. Children based judgments on number 1.82 times on average for count syntax (46% overall) and 0.5 times on average for mass syntax (13% overall), a difference that was marginally significant, F (1, 23) = 4.1, $p < .06$. There was also a large effect of task order, F (1, 23) = 11.7, $p < .05$, that reflected a much greater distinction of mass-count syntax when

quantity judgment was tested after word extension. Children who were given the tasks in this order quantified by number 2.67 times on average for count syntax (compared to .8 times in the alternative order) and 0 times on average for mass syntax (compared to 1 time on average in the alternative order).

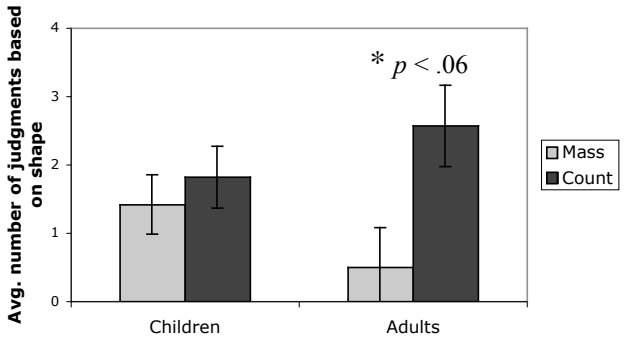


Figure 6. Word extension by 3-year-olds and adults for mass and count nouns that label non-solid substances.

A review of the word extension data revealed a similar trend, where performance on word extension more closely resembled adult performance when it followed quantity judgment, suggesting an overall effect of accumulated input on mass-count sensitivity over the course of the experiment.

Overall, results suggest that children have more than one interpretation for the word "more", but that its meaning is shifted primarily by the ontological category of referents, rather than mass-count syntax.

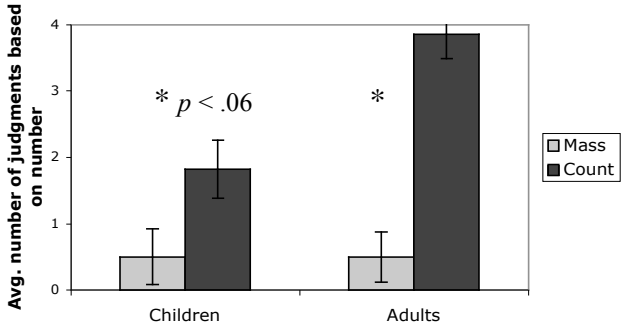


Figure 7. Quantity judgment by 3-year-olds and adults for mass and count nouns that label non-solid substances.

General Discussion

Experiment 1 revealed two main results. First, for the word extension task, both children and adults used mass-count syntax to guide their judgments, extending count terms on the basis of shape and mass terms on the basis of substance. This replicates results found for English-speaking 2.5 year olds (see Soja, 1992). Second, for solid objects, adults used mass-count syntax to guide quantity judgment and used mass nouns to quantify over individuals only for complex objects. Children, however, interpreted both novel mass and count terms as quantifying by number regardless of referent type, suggesting a lack of strong quantificational interpretation for mass syntax.

Results from Experiment 2 indicated that children could use mass-count syntax to guide quantification for non-solid substances. Across conditions, children's interpretation of mass syntax appeared more affected by referent type than the interpretation of count syntax. The number bias found in Experiment 1 disappeared mainly for mass nouns when referents were changed from discrete physical objects to non-solid substances.

Table 1
Summary of results for Experiments 1 and 2
(✓ = effect of syntax; ✘ = no effect of syntax)

	Solids		Non-solids	
	W.E.[1]	Q.J.[2]	W.E.	Q.J.
3-year olds	✓	✘	✘	✓
Adults	✓	✓	✓	✓

[1] Word extension
[2] Quantity judgment

These results (summarized in Table 1) have three important consequences. First, children appear to employ distinct mechanisms for performing quantity judgments and word extensions. The main source of this difference, and of the difference between children and adults in this study, was children's treatment of simple solid objects. While both adults and children quantified by number for mass nouns that referred to complex objects, only children quantified by number for mass nouns that referred to simple objects and that were extended by substance. Interestingly, these results appear to be consistent with observations that young children are biased to enumerate spatio-temporally defined individuals when counting segmented objects (e.g., counting a fork cut in half as two forks; see Shippley & Shepperson, 1990; Wagner & Carey, 2003). In these tasks, children are unable to use criteria of individuation for specific words, despite being able to correctly extend them. By analogy, when performing quantity judgments, children seem to ignore criteria of individuation for newly learned words and instead employ the more primitive criteria of the "object" sortal (Xu, 1997). Children appear to base quantification on spatio-temporal individuals whenever arrays are composed of physical objects, and only later in acquisition use specific sortal information to guide judgments.

It should be noted that although a spatio-temporal bias may account for the content of individuals that children quantified (i.e. rather than specific sortal knowledge) it cannot explain the failure of mass syntax to determine the dimension of measurement. Our evidence suggests that 3-year-olds are only beginning to understand the effect of mass-count syntax on quantification, and that they do not have strong syntax-semantics mappings of the type proposed by Bloom (1999). Thus, this study provides further evidence children never use one-to-one mappings

between syntax and semantics of the type proposed by Bloom (1999). Instead, it seems that only count syntax is ever truly specified for a uniform interpretation (Barner & Snedeker, 2004).

Given children's difficulty using syntax to guide quantity judgment, how do they succeed with word extension? Minimally, the task requires an ability to distinguish mass and count distributional frames and to determine a novel word's criteria of application, or content. One possibility is that these two types of knowledge are sufficient for solving word extension, and may operate somewhat independently of actual mass-count semantics. For example, children may observe that a novel term like *blicket* has been used in the same syntax as a word like *plastic*, for which shape is an irrelevant dimension, and therefore extend the word on the basis of substance. Novel word extensions may be made on the basis of correlations between mass-count syntax and previous extensions, and may be the product of simple mappings between content and syntax, bypassing noun phrase quantification altogether.

Based on this, our results suggest an ability to use mass-count syntax to guide word extension by 2;6 (e.g., Soja, 1992) may not reflect mature knowledge of mass-count semantics. Even by 3;6, children are only beginning to understand the full effect of mass-count syntax on quantification, which arguably defines the distinction. Early in acquisition, children may approximate adult behavior conditions by exploiting correlations between content and distributional frames, and by employing ontologically based biases for interpreting quantifiers like "more". Sometime between 3;6 and 4, children appear to recognize that count syntax specifies number as a dimension for comparison, and that in this way it differs from mass syntax (Barner & Snedeker, 2004). Mass nouns continue to be interpreted based on lexical properties, as shown by adults in this study, and never specify a uniform dimension for measurement.

Acknowledgements

We would like to thank Susan Carey and Peggy Li for their helpful comments on previous versions of this paper. This work was funded in part by a grant from the National Science Foundation to the second author (IIS-0218852, schaward 5710001440).

References

Barner, D., & Snedeker, J. (2004). Quantity judgments and individuation: Evidence that mass nouns count. Under review.

Bloom, P. (1999). The role of semantics in solving the bootstrapping problem. In: Jackendoff, R., Bloom, P., Wynn, K. (Eds.), *Language, Logic, and Concepts: Essays in Memory of John Macnamara*. Cambridge, MA: MIT Press.

Borer, H. (2004). *In Name Only*. Unpublished manuscript, University of Southern California, Los Angeles.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Cazden, C.B. (1968). The acquisition of noun and verb inflections. *Child Development, 39,* 433-448.

Donaldson, M. & Balfour, G. (1968). Less is more: A study of language comprehension in children. *British Journal of Psychology, 59,* 461 – 471.

Ferenz, K., & Prasada, S. (2002). Singular or plural? Children's knowledge of the factors that determine the appropriate form of count nouns. *Journal of Child Language, 29*, pp. 49-70.

Gathercole, V. (1985). More and more and more about more. *Journal of Experimental Child Psychology, 40,* 73-104.

Geach, P. (1962). *Reference and Generality*. Ithaca, NY: Cornell University Press.

Gordon, P. (1985). Evaluating the semantic categories hypothesis: the case of the mass/count distinction. *Cognition, 20,* 209–242.

Gordon, P. (1988). Mass/count category acquisition: Distributional distinctions in children's speech. *Journal of Child Language, 15,* 109-128.

Gupta, A. (1980). *The logic of common nouns*. New Haven, Conn.: Yale University Press.

Imai, M., & Gentner, D. (1997). A cross-linguistic study on early word meaning. Universal ontology and linguistic influence. *Cognition, 62,* 169-200.

Link, G. (1998). *Algebraic semantics in language and philosophy*. Stanford, CA: Center for the Study of Language and Information.

Macnamara, J. (1986). *A border dispute: The place of logic in psychology.* Cambridge, MA: MIT Press.

Palermo, D. (1973). More about less: A study of language comprehension. *Journal of Verbal Learning and Verbal Behavior, 12,* 211-221.

Prasada, S. (1999). Names for things and stuff: An Aristotelian perspective. In R. Jackendoff, P. Bloom, & K. Wynn (Eds.). *Language, logic, and concepts: Essays in honor of John Macnamara* (pp. 119–146). Cambridge, MA: MIT Press.

Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Soja, N.N. (1992). Inferences about the meanings of nouns: the relationship between perception and syntax. *Cognitive Development, 7,* 29-45.

Soja, N.N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meaning: object terms and substance terms. *Cognition, 38,* 179-211.

Wagner, L., Carey, S. (2003). Individuation of objects and events: a developmental study. *Cognition, 90,* 163 – 191.

Wisniewski, E.J., Imai, M., & Casey, L. (1996). On the equivalence of superordinate concepts. *Cognition, 60,* 269-298.

Xu, F., & Carey, S. (1996). Infants' metaphysics: the case of numerical identity. *Cognitive Psychology, 30,* 111 – 153.

Cultural Differences in the Cognition and Emotion of Conditional Promises and Threats – Comparing Germany and Tonga

Sieghard Beller (beller@psychologie.uni-freiburg.de)
Andrea Bender (bender@psychologie.uni-freiburg.de)
Department of Psychology, University of Freiburg
D-79085 Freiburg, Germany

Abstract

When addressing conditional inducements using a multi-level approach, several cognitive components appear to be of basic character: linguistic preferences for either a promise or threat are connected to the motivational background; the concepts themselves are unilateral and complementary; and emotional responses in subsequent interactions follow appraisal-theoretic predictions. Whether these apparently essential components really are basic was examined in a cross-cultural experiment conducted in Tonga. The results support the conceptual universality; however, in practice, the Tongan participants tended to avoid threats in favor of promises and indicated less anger following broken promises.

Introduction

In pursuing their own individual goals, people occasionally feel a need to change the behavior of others accordingly. When asking to have one's intentions considered is not of much use, a strategy often chosen is to formulate a conditional promise or threat. For instance, a mother longing for silence may tell her noisy son: "If you are quiet for the next hour, I will give you an ice-cream." Or, being already unnerved, she may say instead: "If you are not quiet for the next hour, there will be no TV this afternoon!" Being intrigued by the pleasant anticipation of an ice-cream the boy might feel motivated to fulfill her request; if she then fails to keep her promise, his positive affects will turn negative.

Conditional promises and threats operate, as these introductory examples show, with goals, expectations, incentives or penalties, and obligations (cf. von Wright, 1962). They express personal motives and are formulated in a specific linguistic manner (Fillenbaum, 1978). They demand a decision whether or not to cooperate, and subsequent actions are often followed by emotional reactions. In order to efficiently work as inducements, knowledge about these components needs to be – and usually is – shared among the partners of an interaction.

But is this knowledge also shared across cultures? In other words, is the understanding of conditional inducements culture-specific or do conditional promises and threats belong to the core concepts of human thinking and acting that are universally equal? If the latter is the case, are there certain aspects of conditional inducements that vary across languages and cultures? Will the interactions subsequent to inducements universally elicit emotional reactions, and if so, are the same emotions then elicited, or will appraisals – and thus the emotional reactions – differ?

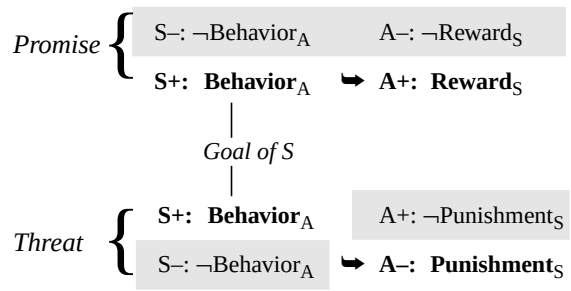
In the reasoning tradition, the analysis of people's understanding of conditional promises and threats focuses on the conditional relation (e.g., Evans & Twyman-Musgrove, 1998; Newstead, Ellis, Evans & Dennis, 1997). By evaluating the inferences that people draw from such statements with the yardstick of propositional logic, it was possible to identify several extra-logical factors: the temporal order of the actions, the promisor's control of the incentive, and the directionality. While this approach helps to detect effects of particular contents on reasoning, it is not sufficient to explain them. To overcome this limitation, a broader multi-level approach was recently proposed and empirically tested (Beller, 2002; Beller, Bender & Kuhnmünch, 2004; cf. Beller & Spada, 2003). It builds on a motivational analysis of why conditional inducements are used, and integrates linguistic, deontic, behavioral and emotional aspects.

So far, most of the experimental data has been gathered in "Western" cultures, leaving the question of whether consensus on conditional inducements may extend beyond cultural boundaries open. To tackle this question, we therefore replicated one of our German experiments in a culture that largely differs from our individualistic Western culture: the Polynesian Kingdom of Tonga. Before presenting the results, we will address the theoretical background and give a short description of relevant aspects of Tongan culture.

Components of Conditional Inducements

Conditional promises and threats are speech acts (Searle, 1969) that are motivated by personal goals (i.e. to obtain something with another person's support) and expectations of this person's behavior. The speaker (S) wants an addressee (A) to perform a certain behavior with a positive value for the speaker (S+: Behavior_A). The speaker must expect that the addressee is *not* willing to show this behavior voluntarily (S–: $\neg\text{Behavior}_A$); otherwise an inducement would not be necessary. Thus, the speaker has to induce a behavioral change, which can be motivated in two ways: by promising to reward the desired behavior (➥ A+: Reward_S) or by threatening

to punish the undesired behavior (➥ A–: Punishment_S). These consequences should be under the speaker's control and should not occur for any other reason, as otherwise they would not be able to develop their motivational effect. The different motivational background of promises and threats are summarized in the following schemas (grey boxes depict the expected actions without the inducement):



In both cases, the addressee can freely decide whether to cooperate or not; the speaker then responds to the addressee's behavior, that is, the actions are ordered temporally. Further, the speaker may use a conditional "If P, then Q" to express the inducement. A conditional points out a necessary consequence "Q" of an antecedent possibility "P", and that is exactly what the speaker intends to do on the motivational level. The canonical formulations are:

"If you do P [S+], then I will reward you with Q [A+]"
"If you do P [S–], then I will punish you by Q [A–]".

Due to the temporal order of the events, neither of the conditionals is reversible, and yet they are tightly interwoven. The speaker always announces (explicitly or implicitly) that he or she will react positively after the addressee has shown the desired behavior, and negatively otherwise. The threat may thus be interpreted as implying the complementary promise

"If you *refrain* from doing P [S+],
then I will *not* punish you by Q [A+]."

While the addressee can freely decide whether to cooperate or not, the speaker cannot. If the addressee cooperates, the speaker is *obligated* to cooperate as well, and when intentionally violating this obligation, the speaker cheats the addressee.

Because the addressee is induced to change a planned behavior – and may experience a positive or negative consequence – the respective interaction will most probably elicit an emotional reaction, as the event has considerable goal relevance (Lazarus, 1991). According to appraisal theories (cf. Ellsworth & Smith, 1988; Roseman, Antoniou & Jose, 1996; Scherer, 1997), joy should be elicited when the addressee obtains what was promised to him or her, relief in the case of avoided punishment, and anger in the case of being cheated.

These linguistic and emotional predictions were confirmed in two studies with German samples (Beller, 2002; Beller, Bender & Kuhnmünch, 2004). Participants preferred the canonical conditionals as appropriate formulations for an intended promise or threat. These are understood as not reversible, but as implying the complementary threat (or promise). The emotional reactions were chosen by the participants as predicted by appraisal theories. The question remains, though, of how these components are defined by members of a completely different culture, particularly one in which individualism is not as marked as in "Western" culture. Will their linguistic preferences and emotional reactions converge with – or diverge from – those obtained in Germany?

The Context of Social Interactions in Tonga

The Kingdom of Tonga, an island state in the Southwestern Pacific with a Polynesian culture, is inhabited by roughly 100 000 people. The island group of Ha'apai, where the data was collected, is one of its most traditional areas with approximately 10 000 people living on two main islands and fifteen outer islands.

Tongan society is hierarchically structured with older people having higher rank than younger ones, sisters higher than brothers, and nobles higher than commoners. While this rank is ascribed through birth, status can be acquired through individual efforts. However, these individual efforts need to pursue common interests, and among the most prestigious behavior is therefore engaging in social activities, keeping social ties and complying with social norms (Bernstein, 1983; Marcus, 1978). Social harmony is particularly emphasized and consequently, negative emotions such as anger (*'ita*) or envy (*meheka*) as well as open conflicts are disapproved of. Failing to restrain one's negative emotions is shameful and may diminish one's social status (Morton, 1996). Cooperation and sharing with others are core values. In exchange situations, certain (though mostly implicit) rules apply, depending for instance on the object of the deal or the participants' relative social status and rank (e.g. Bender, 2002; Evans, 2001).

Most of this contextual knowledge still does not reveal how Tongans understand and deal with conditional promises and threats. In principle, we presumed that the basic *understanding* of conditional inducements does not differ across cultures, but that their *use* may differ: depending on what is socially more accepted, either promises or threats may be preferred. With regard to appraisal processes, we further presumed a similar pattern of emotional responses. Taking into account the disapproval of socially disruptive behavior, however, we also expected a smaller proportion of both threats and indication of anger in the Tongan sample.

Experiment

In order to investigate the impact of culture on understanding and dealing with conditional inducements, we replicated a study we had conducted on linguistic preferences and emotional reactions in Germany (Beller, Bender & Kuhnmünch, 2004, Experiment 1) in Tonga.

Method

Participants. Sixty-seven students from two classes of St. Joseph's College in Pangai participated in the experiment. St. Joseph's College is the second largest of four secondary schools in the district center of the Ha'apai island group. Thirty-one students were male and thirty-five female (one did not indicate his or her gender). The mean age was M = 15.4 years (SD = 0.83; *range*: 14-17 years). Both classes received a small gift for their participation in the form of a contribution (50 pa'anga, approximately 33 US dollars) to their class funds.

Materials. As in the original study, we used two pairs of questionnaires that referred to a situation in which a boy would like to obtain something from a schoolmate. Sione wants to borrow Finau's bike, while Finau wants Sione to help him with his homework. Within each pair, the speech act was varied. In the first pair, Sione was in the role of the speaker and used either a promise to reach his goal (promise version) or a complementary threat (threat version). In the second pair, the role allocation was switched, with Finau now in the role of the speaker trying to reach his goal either by a (reversed) promise or by a complementary threat. Each questionnaire comprised four tasks.

(1) In the *formulation task*, the individual goals of both boys were given. For example, the promise version of Sione's inducement reads as follows (threat version printed in square brackets): *Usually, Finau doesn't lend his bike to his schoolmates. However, Sione wants to borrow it today. Sione tries to reach this goal by promising [threatening] Finau [with] something. Sione knows that Finau would like his help with his homework today, but usually Sione does not help him [and usually Sione helps him]*. The instructions then required the participants to choose the most appropriate conditional for the speaker's promise [threat] from four possible formulations: the canonical, the complementary, and the two reversed statements.

Each of the other three tasks first presented the speech act in canonical form, for example: *Sione promised: "Finau, if you lend me your bike, then I will help you with your homework" [Sione threatened: "Finau, if you do not lend me your bike, then I will not help you with your homework."]* (2) The *sequence task* asked participants to identify the typical sequence of actions once the conditional promise or threat has been uttered: Will Finau or Sione be the first to decide to act? (3) The *inference task* asked participants to draw the most probable conclusion from the given canonical conditional: What follows – the complementary, the reversed, or the reversed-complementary conditional? (4) The *emotion task* stated that the addressee has fulfilled the speaker's goal. The instruction then required participants to indicate first the speaker's action when "keeping the rule" (vs. "not keeping the rule") and second, the addressee's feeling towards the speaker afterwards. Four critical emotions were given (Tongan translation in brackets): joy (*fiefia*), relief (*fiemālie*), sadness (*loto-mamahi*), and anger (*'ita*). For exploratory purposes, we also included shame (*mā*), un-amusement (*ta'eoli*) and anxiety (*manavahē*), as these emotions seem to be particularly salient in the Tongan context. Participants were instructed to choose the most appropriate emotion.

All materials were presented in the participants' native language, that is Tongan in this study and German in the original study. Both languages have close equivalents to the English terms "promise" and "threat", and the situations described are familiar in both cultures.

Design and procedure. A between-subject design was used. Participants were randomly assigned to one of the four experimental conditions (n = 16 or 17 for each) corresponding to the four questionnaires. The data collection took place in the classrooms. Each participant received a booklet with a general instruction and the four tasks in the following order: formulation, sequence, inference, and emotion task. Participants were instructed to answer all questions in the given order, and they were granted as much time as they needed.

Results

As we found only marginal differences between the two promise versions and between the two threat versions, the data is reported in aggregated form. The comparative data of the German study is taken from Beller, Bender and Kuhnmünch (2004, Experiment 1).

Formulation task: We expected participants to choose the *canonical* promise or threat. The formulation preferences are shown in Table 1. In the German sample, on average 89.2 % of the participants chose the canonical form equally for promises and threats ($\chi^2(1, N = 65) = 1.34$; $p = .247$). The Tongan results differed from this in two respects. First, the canonical conditional was chosen less frequently (52.2 % on average), and second, this difference mainly resulted from a specific formulation preference: while most Tongan participants also preferred the canonical promise, many of them indicated that, when a *threat* was to be made, the speaker should rather use the complementary *promise* instead ($\chi^2(1, n = 52) = 6.10$; $p = .014$).

Sequence task: We expected both conditional promises and threats to be understood as unilateral speech acts that imply a typical action sequence: the addressee

Table 1: Proportions of choosing each conditional as the speaker's adequate promise or threat.

	Germany		Tonga	
Conditional	Promise (n = 32)	Threat (n = 33)	Promise (n = 33)	Threat (n = 34)
Canonical	**.94**	**.85**	**.64**	**.41**
Complementary	.06	.15	.12	.38
Reversed	–	–	.18	–
Rev.-complementary	–	–	.06	.21

Table 2: Proportions of choosing the most adequate implication of a given promise or threat.

	Germany		Tonga	
Conditional	Promise (n = 33)	Threat (n = 31)	Promise (n = 33)	Threat (n = 34)
Canonical	*given*	*given*	*given*	*given*
Complementary	**.88**	**.84**	**.48**	**.65**
Reversed	.09	.16	.45	.21
Rev.-complementary	.03	–	.06	.15

first decides whether or not to cooperate. The predicted sequence was chosen by 87.8% of the participants in the German sample (N = 66) and by 74.6% in the Tongan sample (N = 67) aggregated across all four task versions ($p < 0.001$ based on the binomial distribution with $r = 1/2$ being the probability of guessing).

Inference task: We expected participants to choose the *complementary* conditional "*If not*-P *then not*-Q" as the most appropriate inference from a given canonical promise or threat "*If* P *then* Q". The results are shown in Table 2. In the German sample, 85.9 % of the participants inferred the predicted conditional on average. There was no significant difference between promises and threats ($\chi^2(1, n = 63) = .648$; $p = .421$). Again, the Tongan data differed from the German data – the complementary conditional was chosen less frequently (56.7 % on average) – and this difference again resulted from a specific inferential preference. While most Tongan participants also preferred the complementary conditional as the most reasonable inference from a threat, many of them – consistent with their formulation preference – avoided inferring the complementary *threat* from a promise in favor of the reversed *promise* ($\chi^2(1, n = 60) = 3.79$; $p = .051$).

Table 3: Proportions of choosing emotional reactions of the addressee (+ positive; – negative) when the speaker did vs. did not keep the rule.

		Germany		Tonga	
Emotion		Promise (n = 33)[a]	Threat (n = 33)[a]	Promise (n = 33)	Threat (n = 34)
		The Speaker Kept the Rule			
Relief	(+)	.27	.52	.21	.29
Joy	(+)	.67	.18	.70	.62
Others	(–)	.09	.30	.09	.09
		The Speaker Did Not Keep the Rule			
Sadness	(–)	.09	.06	.48	.53
Anger	(–)	.94	.91	.27	.26
Others	(+/–)	–	–	.24	.21

[a] Due to missing/double answers, the proportions do not always add up to 1.00 in these columns.

Emotion task: Participants were first required to decide which action the speaker has to take in order to keep "the rule" (i.e., to cooperate given that the addressee cooperated before) or not to keep "the rule" (i.e., to react defectively even though the addressee fulfilled the speaker's goal). Almost all participants indicated the appropriate action (all participants in the German sample and 96.3% in the Tongan sample, aggregated over both questions; n = 134 answers).

How does the addressee respond emotionally after the speaker did or did not keep "the rule"? The results of the emotion tasks are shown in Table 3. In line with appraisal theories of emotion, the addressee was said to feel a positive emotion if the speaker kept the rule (85.8 % positive vs. 14.2 % negative), and a negative emotion was said to result if the speaker did not keep the rule (3.0 % positive vs. 97.0 % negative; $\chi^2(1, n = 267$ answers) = 185.3; $p < 0.001$).

Separate analyses were performed for the cooperative and the non-cooperative situation in both samples. The dependent variable "emotion" was classified into three categories in both cases: relief, joy, and all other emotions in the cooperative situation, and sadness, anger, and all other emotions in the uncooperative situation.

In the German sample, keeping the promise resulted in joy (joy: 66.7 %; relief: 27.3 %), while in the case of an avoided threat, relief predominated (joy: 18.2 %; relief: 51.5 %; $\chi^2(1, n = 54) = 10.65$; $p = 0.001$). Interestingly, 24.2 % of the participants indicated that the addressee feels angry, apparently because the addressee was "forced" to cooperate by a threat, as described by Heilmann and Garner (1975). In the Tongan sample, the addressee was quite uniformly said to feel joy (joy: 65.7 %; relief: 25.4 % on average), independent of the speech act ($\chi^2(1, n = 61) = 0.604$; $p = 0.437$).

In cases where the speaker did not keep the rule, we expected a difference between the two cultures with regard to the attribution of anger. The data supports this prediction. While the German participants predominantly ascribed anger (anger: 92.4 %; sadness: 7.6 %), half of the Tongan participants indicated that the addressee will feel sad in this situation (anger: 26.9 %; sadness: 50.7 %; $\chi^2(1, n = 118) = 43.9$; $p < 0.001$).

Discussion

The cultural comparison revealed many commonalities in the understanding of conditional promises and threats, but also some characteristic differences.

First, both German and Tongan participants chose formulations and inferences with the addressee's action in the antecedent proposition in accordance with the typical action sequence "addressee first". The Tongan participants, however, tended to avoid explicit threats and preferred to use promises instead.

Second, conditional inducements have a potential for emotional interactions in both cultures. Positive and

negative emotions were generally attributed in line with appraisal-theoretic predictions. Consistent with findings from another domain (Bender, 2001), indication of angry reactions is largely reduced in Tonga: in Germany, anger clearly predominated over sadness after the speaker's defection, while the Tongan sample showed the reversed pattern, with sadness twice as frequent as anger. While we had expected a difference for the negative emotions, we had not expected any difference for the positive ones, but found that the German students distinguished between joy and relief depending on the speech act whereas the Tongan participants did not.

In general, the differences found in the Tongan sample can be explained by the cultural background.

The tendency of roughly half of the participants to avoid threats in favor of promises is in accordance with respective social rules. In addition, as cooperation and particularly sharing with others are core values in Tongan society, threats may simply be not appropriate as a means of initiating an exchange.

A conclusion from the affective differences is not as easy to draw. With regard to the emotional reaction upon the speaker's non-cooperation, two explanations are conceivable: the lack of anger indication may either be due to the fact that anger is socially not acceptable or that it arises less often in such situations. The latter could be the case if not fulfilling one's obligation to reciprocate is regarded as so unusual that this failure is appraised differently, thus giving rise to different emotions.

Taking together findings from other domains allows us to presume that our Tongan participants may indeed have appraised the described event differently from the German sample. Several studies have indicated that – despite the culture independence of most appraisal dimensions – cultural differences do occur with regard to goals and values and with regard to certain appraisals that involve rather complex concepts such as causation or responsibility (e.g., Mauro, Sato & Tucker, 1992; Norenzayan, Choi & Nisbett, 1999). The latter concepts are directly relevant for anger as this emotion is elicited if another person is held personally responsible for a negative event. Personal responsibility, however, is not as easily ascribed in some cultures as it is in Western ones. Particularly in cultures with an interdependent self-concept (cf. Markus & Kitayama, 1991), causal attributions are more often also made in view of the circumstances (cf. Morris, Nisbett & Peng, 1995). The ambivalent or rather open description of the experimental scenario may thus lead to two different interpretations of the speaker's non-cooperation: in Germany, participants may interpret the scenario as involving a rather high level of personal responsibility, while the Tongan participants may also take circumstances into account. This difference would then result in diverging emotional responses: a dominance of angry reactions in Germany, but one of sad reactions in Tonga. This interpretation is supported by findings from a subsequent experiment in Germany (Beller, Bender & Kuhnmünch, 2004, Experiment 2) in which we varied personal responsibility: Anger dominated in situations with high personal responsibility and decreased in favor of sad reactions in situations with low responsibility.

The different indications of joy versus relief upon mutual cooperation after a threat also reflect cultural differences. On the one hand, this may again be due to different appraisal patterns. Focusing on the positive outcome of cooperation elicits joy (as in the Tongan sample), while focusing on the transition from an expected negative outcome to the final positive one elicits relief (as in the German sample). On the other hand, the difference may also have lexical origins. There is no Tongan word that precisely translates as "relief". The word we chose (after discussing the content scenarios with our Tongan partners) was *fiemālie*. This comes closest to "relief", but also encompasses "to be easy in mind, contented, satisfied, free from pain, discomfort or sorrow". These connotations could also be the reason why some of our Tongan participants did not choose *fiemālie* in the particular situation.

This last point highlights a particular difficulty of language-based cross-cultural research: even if terms exist that have the same core meaning across languages, they often contain a spectrum of connotations that do not map. Although well known (e.g., Tihanyi, 2002), this problem is not easy to resolve. While a linguistic tool has been developed in recent years for identifying and describing such differences in meaning (Wierzbicka, 1993, 1999), there is no way to circumvent them. Data interpretation thus requires the differences in meaning to be noted, and may indeed even profit from this insight into different semantic fields of a concept.

As we have seen, conditional inducements are fairly complex concepts as they involve personal values and goals, a social consensus on reciprocity and obligations, the consideration of the situational context, and an estimation of the addressee's goals and the appropriateness of the chosen inducement. Complex concepts by no means *generally* converge across cultures, as has been shown to be the case, for instance, with concepts of causality (Boyer, 1996; Morris, Menon & Ames, 2001), concepts of the self (Kanagawa, Cross & Markus, 2001), or even spatial concepts (Levinson, 1996). Our cross-cultural comparisons suggest, though, that concepts of deontic social norms (Bender & Beller, 2003) and of promises and threats comparable to ours do exist in a completely different culture. When used to induce an exchange, however, they are used somewhat differently, and subsequent interactions may result in different emotional responses.

In a certain sense, the results of the study produced more questions than they answered. For this reason, a replication of the second experiment by Beller, Bender and Kuhnmünch (2004), which addresses the ascription of responsibility more thoroughly and includes deontic aspects, is being conducted in Tonga. As it is part of an anthropological inquiry into the cultural context of promises and threats, participant observation and in-depth interviewing will provide further data that may bolster the ecological validity of our present results.

Acknowledgments. The data collection took place during field research of the second author in Tonga, which was funded by a grant from the Deutsche Forschungsgemeinschaft (DFG). We would like to thank Josef Nerb, Hannah Swoboda, and Stefan Wahl as well as Dan Sperber for discussion and valuable comments on earlier versions of this paper. We are particularly indebted to Moana and Sione Faka'osi for the Tongan translation of the material, for the opportunity to conduct the research at St. Joseph's College in Pangai, and for their unfailing support.

References

Beller, S. (2002). Conditional promises and threats – Cognition and emotion. In W. D. Gray, & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 113-118). Mahwah, NJ: Erlbaum.

Beller, S., Bender, A., & Kuhnmünch, G. (2004). *Understanding conditional promises and threats.* Manuscript submitted for publication.

Beller, S., & Spada, H. (2003). The logic of content effects in propositional reasoning: The case of conditional reasoning with a point of view. *Thinking and Reasoning, 9,* 335-378.

Bender, A. (2001). "God will send us the fish" – Perception and evaluation of an environmental risk in Ha'apai, Tonga. *Research in Social Problems and Public Policy, 9,* 165-190.

Bender, A. (2002). Environmental models, cultural values, and emotions: Implications for marine resource use in Tonga. *International Research in Geographical and Environmental Education, 11,* 58-62.

Bender, A., & Beller, S. (2003). Polynesian *tapu* in the 'deontic square': A cognitive concept, its linguistic expression and cultural context. In R. Alterman, & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.* CD-ROM; Cognitive Science Society (USA).

Bernstein, L. M. (1983). *Ko e lau pe (It's just talk): Ambiguity and informal social control in a Tongan village.* Ann Arbor: University Microfilms International.

Boyer, P. (1996). Causal understandings in cultural representations: Cognitive constraints on inferences from cultural input. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 615-649). Oxford: Clarendon Press.

Ellsworth, P. C., & Smith, C. A. (1988). From appraisal to emotion: Differences among unpleasant feelings. *Motivation and Emotion, 12,* 271-302.

Evans, M. (2001). *Persistence of the gift: Tongan tradition in transnational context*. Waterloo, Canada: Wilfrid Laurier University Press.

Evans, J. St. B. T., & Twyman-Musgrove, J. (1998). Conditional reasoning with inducements and advice. *Cognition, 69,* B11-B16.

Fillenbaum, S. (1978). How to do some things with IF. In J. W. Cotton, & R. L. Klatzky (Eds.), *Semantic factors in cognition* (pp. 169-231). Hillsdale, NJ: Erlbaum.

Heilman, M. E., & Garner, K. A. (1975). Counteracting the boomerang: The effects of choice on compliance to threats and promises. *Journal of Personality and Social Psychology, 31,* 911-917.

Kanagawa, C., Cross, S. E., & Markus, H. R. (2001). "Who am I?" The cultural psychology of the conceptual self. *Personality and Social Psychology Bulletin, 27,* 90-103.

Lazarus, R. S. (1991). *Emotion and adaptation.* New York: Oxford University Press.

Levinson, S. C. (1996). Relativity in spatial conception and description. In J. J. Gumperz, & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 177-202). Cambridge: Cambridge University Press.

Marcus, G. E. (1978). Status rivalry in a Polynesian steady-state society. *Ethos, 6,* 242-269.

Markus, H. R. & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224-253.

Mauro, R., Sato, K., & Tucker, J. (1992). The role of appraisal in human emotions: A cross-cultural study. *Journal of Personality and Social Psychology, 62,* 301-317.

Morris, M. W., Menon, T., & Ames, D. R. (2001). Culturally conferred conceptions of agency: A key to social perception of persons, groups, and other actors. *Personality and Social Psychology Review, 5,* 169-182.

Morris, M. W., Nisbett, R. E. & Peng, K. (1995). Causal attribution across domains and cultures. In D. Sperber, D. Premack & A. J. Premack (Eds), *Causal cognition* (pp. 577-612). Oxford: Clarendon Press.

Morton, H. (1996). *Becoming Tongan: An ethnography of childhood.* Honolulu: University of Hawai'i Press.

Newstead, S. E., Ellis, M. C., Evans, J. St. B. T., & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning, 3,* 49-76.

Norenzayan, A., Choi, I., & Nisbett, R. E. (1999). Eastern and Western perceptions of causality for social behavior: Lay theories about personalities and situations. In D. A. Prentice & D. T. Miller (Eds.), *Cultural divides: Understanding and overcoming group conflict* (pp. 239-272). New York: Russell Sage.

Roseman, I. J., Antoniou, A. A., & Jose, P. E. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion, 10,* 241-277.

Scherer, K. R. (1997). Profiles of emotion-antecedent appraisal: Testing theoretical predictions across cultures. *Cognition and Emotion, 11,* 113-150.

Searle, J. R. (1969). *Speech acts*. Cambridge: Cambridge University Press.

Tihanyi, C. (2002). Ethnographic and translation practices. *Anthropology News, 43,* 5-6.

von Wright, G. H. (1962). On promises. *Theoria, 28,* 276-297.

Wierzbicka, A. (1993). A conceptual basis for cultural psychology. *Ethos, 21,* 205-231.

Wierzbicka, A. (1999). *Emotions across languages and cultures*. Cambridge: Cambridge University Press.

Retrieval Structure Construction During Reading: Experimentation and Simulation

Cédrick Bellissens (cedrick.bellissens@univ-paris8.fr)
Laboratoire Cognition et Usages
CNRS & Université de Paris VIII, 2 rue de la liberte
93526 St Denis, France

Guy Denhière (denhiere@up.univ-mrs.fr)
Laboratoire de Psychologie Cognitive
CNRS & Université de Provence, 3 place Victor Hugo
13331 Marseille, France

Abstract

The aim of this study was to investigate the construction of a retrieval structure during reading, according to the hypothesis that text macrostructure is used in Long-term working memory (Ericsson & Kintsch, 1995) to maintain encoded information in an accessible format. We first designed an experiment for testing the hypothesis that retrieval structure is a macrostructure of the text. Then, we conceived and run a model inspired by CI-LSA Framework (Kintsch, Patel & Ericsson, 1999) in which a generalization process create macropropositions. Results are that simulation data were found to be highly correlated with participants' data.

Macrostructure as Retrieval structure

Classical view of working memory assumes that during reading relevant information is stored in a Short Term Memory Buffer (STMB; Baddeley, 2000; Kintsch, 1988). In contrast, Ericsson & Kintsch (1995) argued that during reading, readers could encode information in an accessible format in a retrieval structure in long-term working memory (LTWM) consisting of retrieval cues associated with encoded information in Long Term Memory (LTM). Ericsson and Kintsch (1995) showed that clearing STMB by using a reading interruption procedure (Glanzer, Fisher & Dorfman, 1984) does not lead to comprehension impairment. We have also shown that the more a text is familiar, the more readers can construct an efficient LTM retrieval structure based on the content of the text (Bellissens & Denhière, 1998; Denhière & Bellissens, 1999).

A CI-LSA Framework has been proposed (Kintsch, 1998 ; Kintsch, Patel & Ericsson, 1999) to explain LT-WM intervention in the comprehension process. The main characteristic of this framework is the combination of a model of semantic memory, LSA (Landauer & Dumais, 1997), and of a model of comprehension, CI (Kintsch, 1988).

LSA represents potential signification of words belonging to a textual corpus in a semantic space. In the semantic space, vectors represent the words and the cosine of the contained angle of two vectors is an estimation of their respective words similarity. The more the cosine is close to 1, the more the two words are considered as semantically similar. The cosine similarity can be used to weight links in an associative network in which each node is a word. In CI, the use of associative network is the way to represent activated knowledge associated with concepts and propositions derived from text processing. Hence, architecture of a text segment representation in CI-LSA is a network that comprehends propositions and concepts (nodes) linked with each other by relations (links) weighted by LSA cosine. This CI-LSA combination improves the previous comprehension model proposed by Kintsch (1988) by the fact that LSA can model knowledge base in CI.

While Ericsson and Kintsch (1995) assumed that the episodic structure of a text could be a retrieval structure, Kintsch, Patel and Ericsson (1999) did not explain exactly how the CI-LSA framework simulates the construction of a retrieval structure. We assume that each text segment processing results in an individual episodic trace. We propose that the set of episodic traces generated by the text processing is replaced by a macrostructure, resulting from the application of a generalization process (O'Reilly & Rudy, 2000). This generalization process generates macropropositions that are associated with the encoded information for further use as semantic retrieval cues. We argue that if two relevant segments of a text, associated with such retrieval cues, are given after a reading interruption, readers should easily reinstate these retrieval cues and counteract the interruption effect. For example, imagine an individual reads a text about the invention of a

machine. He/she reads that the machine has a particular function. Then, the text says that the machine includes one component with a specific function, then a second component with an other specific function. At this moment, the reading is interrupted and then resumed after the presentation of a probe sentence. After Ericsson and Kintsch (1995) and Bellissens and Denhière (2003), the probe sentence should help reader for reinstated semantic cues in STM associated with information encoded in LTM when the previous text has been processed.

The present paper tries to explain how this reinstatement usually occurs and how the reinstatement depends on the organization of the retrieval structure. As we assume that the construction of the retrieval structure is a generalization process, we predict that a sentence mentioning the two components of the machine is a better probe sentence than a sentence comprising two distinct facts: the machine function and the function of a machine component; Indeed, the first type of probe sentence relies on a particular part of the retrieval structure of the text, a macroproposition: machine usually possesses components. The second relies on two separated parts of the macrostructure: the functions of the machine and its component.

What we have just described is a formal view of the conditions we constructed for the following experiment. We predict that if a categorical probe sentence (CAT) is inserted after a reading interruption, readers should resume the reading faster than with a functional probe sentence (FUN).

Table 1: Texts facts

Nb	Sentence
Title	The /Machine/
1	The /Machine/ was invented in /Date/ by /Inventor/
2	The /Machine/ possesses a /Component 1/ to /Component 1 Function /.
3	The /Machine/ is mainly used to /Main Function of the Machine/.
4	The /Machine/ possesses a /Component 2/ to /Component 2 Function /.
5	This component thus contributes to the function of the /Machine/.
6	Since its invention, the /Machine/ is an equipment that was modernized.
7	The modernization of the /Machine/ was useful for the development of /Human Domain/.
CAT	The /Machine/ possesses a /Component 1/ and a /component 2/ to /Component 2 Function /.
FUN	The /Machine/ is used to /Main Function of the Machine/ and possesses a /Component 2/ to /Component 2 Function /.
ORI	The /Machine/ possesses a /Component 2/ to /Component 2 Function/.

Experiment

Participants

Participants were 64 students from Université de Provence, Aix-en-Provence, France.

Materials and Procedure

Twelve pairs of experimental texts and eight pairs of filler texts were generated. A text pair consisted of a main text, an interrupting text, and five comprehension questions. The main experimental texts and the main filler texts described a machine (e.g., the automobile, the elevator, the phonograph, etc.). The main experimental texts all had the same structure (see, Table 1) but the main filler texts did not. The interrupting texts were stories totally unrelated to the main texts.

An experimental trial included the presentation of the main text, the interrupting text, and five text comprehension questions. It began with the self-paced display of the main text, sentence by sentence. An experimental trial included the presentation of a main text, an interrupting text, and five text comprehension questions. It began with the self-paced display of the main text, sentence by sentence. To go on to the next sentence, the reader had to press the space bar on the keyboard. Sentence 2 of the main texts mentioned a machine component and its particular function. Sentence 3 stated the general function of the machine. Sentence 4 gave a second machine component and its particular function. Then a message saying "Attention! Reading time is limited" ("Attention ! Lecture en temps fixé") appeared on the screen. This message stayed on the screen for 1500 ms and meant that the interrupting text would start on the next page. The interrupting text was displayed sentence by sentence for a fixed amount of time (4500 ms per sentence). After the presentation of the interrupting text, a message saying "Attention! Reading time no longer limited" ("Attention ! Lecture en temps libre") appeared on the screen.

Then, one of 4 Probe sentences was displayed: either (i) the original Probe Sentence that was the sentence 4 of the main text (ORI), or (ii) categorical Probe Sentence that contained the two machine components and the second component specific function (CAT), or (iii) a functional Probe Sentence that contained the general machine function, the second component and its specific function (FUN); or (iv) a without Relation Sentence that was a new sentence (WRE). Following the display of the probe sentence, a critical sentence 5 was then displayed. Sentence 5 contained an anaphoric device; the referent was the machine component mentioned in sentence 4 (see, table 2).

In each list, 50% of the main texts were interrupted: Filler texts were never interrupted.

At the end of reading, participants were asked to answer 4 comprehension questions about the main text and 2 comprehension questions about the interrupting text. Second question was the critical question for assessing understanding because its correct answer depends on the encoding of the correct antecedent to anaphora in the fifth sentence.

In the control condition, the first part of the main texts was not displayed; and in the experimental condition, all materials were presented.

Results

Control condition. In the control condition, the first part of the text was not presented. Results are that the critical sentence was read faster in the cued conditions than in the without relation condition, 4221 ms vs. 4597 ms, $F(1,31) = 3,6$ $p<.05$. The critical sentence was read at a same rate in the ORI, CAT and FUN conditions. Note that the sixth sentence was read faster in the probe sentence conditions than in the WRE condition, 3488 ms vs. 3823 ms, $F(1,31) = 4,86$ $p<.05$. Percentage of correct answers to the critical question was greater in the FUN condition than in the CAT condition, 85% vs. 55%, $F(1,31) = 6,49$ $p<.05$ and was greater in the probe sentence conditions than in the WRE condition, 76% vs 42 %, $F(1,35) = 7,8$ $p<.01$.

Table 2: example of text (translated from French)

Nb	Sentence
Title	The elevator
1	The elevator was invented in 1859 by the American Otis.
2	The elevator possesses a sliding door to protect the users from the outside.
3	The elevator is mainly used to reach the floors of a building.
4	The elevator comprises a solid winch that controls the rise of the cabin.
5	This component thus contributes to the utility of the elevator.
6	Since its invention, the elevator is equipment that was modernized.
7	The modernization of the elevator was useful for the development of the residences.
CAT	The elevator possesses a sliding door and a solid winch that controls the rise of the cabin.
FUN	The elevator is used to reach the floors of a building and comprises a solid winch that controls the rise of the cabin.
ORI	The elevator comprises a solid winch that controls the rise of the cabin.

Experimental condition. The critical sentence was read faster in the CAT condition than in the FUN condition, 3869 ms vs. 4437 ms, $F(1,31) = 9.33$ $p<.01$.

The percentage of correct answers to the critical question was greater in the probe sentence condition than in the WRE condition, 55.6% vs. 36.0%, $F(1,31) = 3.9$ $p<.05$., but there was no difference between the CAT and the FUN conditions.

Simulation with CI-LSA+Generalisation Framework

Procedure

Construction Phase. We assumed that a text segment could be represented as an associative network consisting in an explicit and an implicit representation. Explicit representation is the result of a predication analysis of the segment. Implicit representation contained the assumed activated knowledge associated with the concepts and the propositions embedded in the text segments. In the model, the implicit representation was made of all of the nearest neighbors (n=2), taken in an LSA space (General reading up to 1st year college), of each of the concepts and propositions represented in the explicit network. Explicit network was weighted by using the rules provided by Kintsch (1988): link weights between propositions and argument, embedded propositions, and proposition that shared an argument were equal to 1 while others were equal to zero. Implicit network was weighted by using cosine similarity from LSA space. If between two vectors, expressing elements of the net, cosine was .20, the weight of their link in the network was .20.

Integration Phase. This phase is described in Kintsch (1988). Here, the integration phase simulated activation spreading in the net. When the network was settled, the more connected information got the greatest final activation value and for that reason remained in a short-term memory buffer, for the next sentence processing.

Short term working memory. Short-term working memory was involved in the construction and integration phases and in the accessibility of the most activated information. Short-term memory in the model is a temporary memory buffer. It contained the most activated information at the end of a processing cycle. We transformed all final activation values in z-notes. Let a be the average of the activation values, e the standard deviation of the distribution v_i the final activation value of a node I, the z-note of I is:

$$z_i = (v_i - a)/e.$$

Only the nodes with z-notes above a given constant were kept in the Short-Term Memory buffer. Hence, the number of elements in the buffer varied as a function of a limited amount of activation and of the network weighting.

In order to simulate a reading interruption as in the experiment, we emptied the memory buffer: We did not keep any nodes of the last processed sentence.

Long-term working memory. Long-term working memory was a retrieval structure consisting in encoded information and macropropositions associated with it. To simulate the construction of a retrieval structure, we build up a matrix containing in row, all activated information during the different processing cycles and, in column, the different processing cycles. Each cell of the matrix contained the final activation value of an element in a given cycle. If the element was not activated in the given cycle, the cell contained a zero. Hence, this matrix contained all episodic memory vectors. To this matrix was applied a singular value decomposition to simulate a generalization process. The decomposition resulted in three matrices (see, Landauer & Dumais, 1997) but we use the matrix representing the coordinates of each element in a memory space. The elements belonging to that space were projected in a two-dimensional space by reducing its dimensionality. The procedure resulted in a two-column matrix. The macrostructure matrix was the product of the two-column matrix and its transposed matrix. The macrostructure matrix was considered as a generalization of the episodic memory traces encoded from the text.

Table 3 : reading times (RT) and activation forces (AF) of the critical sentence and the probe sentence in Original (ORI), categorical (CAT) and functional (FUN) conditions.

	Critical sentence			Probe sentence		
	ORI	CAT	FUN	ORI	CAT	FUN
RT	14,29	13,16	15,09	8,96	6,94	6,04
AF	11,36	13,76	11,07	11,80	16,66	17,87

Macropropositions reinstatement and reading resumption. After a reading interruption, we assume that readers might be able to readily retrieve mental representation of the text from long-term working memory by reinstating cues associated with encoded information. In the model, macropropositions were reinstated in the short-term memory buffer and were used as context to process the next sentence. As in the experiment above, we used three kind of probe sentences. In a first step, we constructed for each, an episodic trace. As they resume some parts of the previous text, their representation shared information with previous sentences that now belonged to the macrostructure matrix. But the shared information was not as such important for each probe sentence. In the categorical condition, information given by the probe sentence referred to a categorical macroproposition "machine has component", while in the functional condition, the probe sentence referred to "machine has a function" and "component has a function". In a second step, we multiplied the episodic vector trace of the probe sentence by the macrostructure matrix. The result of the product was an echo vector. We z-transformed the coordinates of the echo vector and only kept the z-notes above 1.5. The respective nodes were taken to make part of the critical sentence network. In a final step, activation spread in the critical sentence net.

For each probe sentence, and for the critical sentence, in each condition, we calculated an activation force. The final activation values of each sentence were z-transformed and we summed all the z-values above 0.

Results

We applied this procedure to three experimental texts. For each sentence, the mean activation force and the mean reading times per character and proposition are presented in the table 3. First, for each text, and in each of the probe conditions, the anaphora antecedent was retrieved from the macroproposition structure. Second, the activation force obtained by the model can predict the reading times obtained by participants. The correlation between reading times and activation forces were negative and significantly different to zero, $r=-.83$, $z(6) = -2.0$, $p < .05$.

Discussion

As predicted, the CAT sentence led to a faster critical sentence reading time than does the FUN sentence. This difference is not due to the fact that the CAT sentence contained two potential antecedents for the anaphora in critical sentence. First, the difference appeared only when the first part of the main text was presented. Second, although in control condition, the percentage of correct answers was greater in the FUN condition than in the CAT condition, this difference was not found in the experimental condition. This result indicates that without the first part of the text, readers did not discriminate the right antecedent from the categorical cue sentence. Moreover, in control condition, the presentation of the WRE sentence exerted an effect on the critical sentence 5 and on the sentence 6 reading times. This indicated that, in the control condition, subject could not rely on a retrieval structure built in LTM to counteract irrelevant information effect on text processing. It results that when the readers have the opportunity to construct a retrieval structure, a categorization of two pieces of information (the two machine components) had occurred during text processing, based on meaning overlap between the two learning episodes: sentence 2 and sentence 4.

The CI-LSA+generalisation framework could explain how a retrieval structure was build up: During reading, each sentence of the text was processed and stored in episodic memory. Then, as a function of meaning overlap between them, sentences representations were associated to the same semantic macroproposition in

LTM or were not associated. The remaining question is when does generalization process occur during reading? Does it occur during the encoding process or during the retrieval process ? Research has to answer these question in the future. Nevertheless, generalization process and more generally integration process appear to be necessary processes for building the retrieval structure that permits texts comprehension.

References

Baddeley, A.D. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, *4*, 417-423

Bellissens, C., & Denhière, G. (2002). Word order or environment sharing: A comparison of two semantic memory models. *Current Psychology Letters*, *8*, 47-60.

Denhière, G., & Bellissens, C. (1998). *Retrieval from long-term working memory.* Spoken paper presented at the 8th annual meeting of the Society for Text and Discourse. Madison.

Denhière, G., & Bellissens, C. (1999). *Long-Term Working Memory: A frequency effect on retrieval time during reading*. Spoken paper presented at 40th annual meeting of the Psychonomics society. Los Angeles.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211-245.

Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior*, *23*, 467-486.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge. University Press.

Kintsch, W., Patel, V.L., & Ericsson, K.A. (1999). The role of long-term working memory in text comprehension. *Psychologia*, *42*, 186-198.

Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

O'Reilly, R. C., & Rudy J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*, 389-397.

The Effect of Cue Predictability on Long-Range Dependencies in Response Times versus Response Durations

Brandon C. Beltz (bbeltz@gmu.edu)
George Mason University, 4400 University Drive
Fairfax, VA, 22030-4444, USA

Christopher T. Kello (ckello@gmu.edu)
George Mason University, 4400 University Drive
Fairfax, VA, 22030-4444, USA

Abstract

Two experiments were conducted in order to test the effects of cue predictability on serial dependencies in response times and response durations. Predictability in the timing (Experiment 1) and identity (Experiment 2) of response cues was manipulated. Results of both experiments showed that long-range dependencies in response times were stronger when cues were predictable versus unpredictable. By contrast, long-range dependencies in response durations were unaffected by cue predictability. Results are discussed in light of five hypotheses about the source of long-range dependencies in human behavior.

Introduction

In most psychological experiments, the variability in human behavior is divided into two categories: some variations in measurement are explained by the experimental factors, and other variations are not. The latter category is often termed *error variance*, and it usually does not play a role in theorizing about the psychological processes under examination. One reason why researchers ignore error variance is because they often assume that it is effectively random, or possibly the product of mundane factors such as practice, fatigue, or perseveration. These assumptions lead one to think of error variance as uninformative or, at best, irrelevant.

A growing body of experimental results has recently prompted some researchers to pay closer attention to the ostensibly random fluctuations in human behavior. It appears that, contrary to popular belief, these fluctuations tend to exhibit patterns that persist over time. A transparent way to think about these patterns is through the *autocorrelation* function. Suppose that X_t is a time series of measurements taken from a participant in an experiment. The autocorrelation of this time series is defined as (Wagenmakers, Farrell, & Ratcliff, in press),

$$C(k) = \frac{E[\{X_t - \mu\}\{X_{t+k} - \mu\}]}{E[\{X_t - \mu\}^2]},$$

where $E[]$ is expected value, μ is the mean of X_t, and k is some number of measurements between the time series and an offset copy of itself.

If measurements are strictly independent of each other, then $C(k)$ is zero for all $k > 0$. The time series is not correlated with itself at any offset, and hence, there are no persisting patterns in the fluctuations. This condition is often referred to as *white noise* (see top series in Figure 1), and it is common to assume that error variance is some type of white noise (e.g., Gaussian). However, it turns out that measurements of human behavior are often not characterized by white noise. Instead, they exhibit serial dependencies such that $C(k)$ is positive for some $k > 0$.

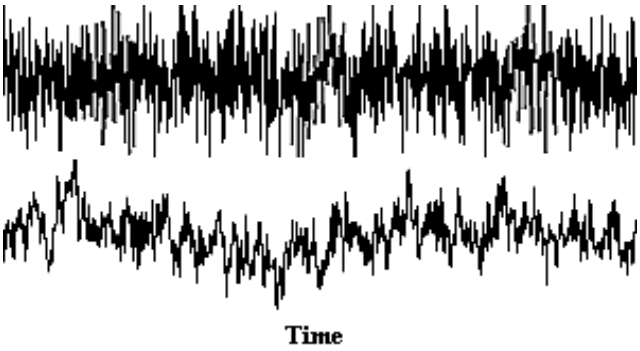


Figure 1: Illustrations of white noise (top) and pink noise (bottom; from Gilden, 2001).

Serial dependencies have been found in a wide variety of human behaviors (for a review, see Van Orden, Holden, & Turvey, 2003). With respect to the study of perception and cognition, serial dependencies have been found in experiments on mental rotation (Gilden, 1997), lexical decision (Gilden, 1997), perceptual learning (Wagman, Dahle, & Schmidt, 2002), simple reaction time (Ward & Richard, 2001), and visual search (Aks, Zelinsky, & Sprott, 2002).

A major question about these findings concerns the kind of dependencies that were observed. The authors of these studies interpreted their findings as evidence for a particular kind of serial dependency often referred to as *long-range dependency,* of which *1/f noise* or *pink noise* are special cases (see bottom series in Figure 1). In a long-range dependent series, $C(k)$ is positive and decreases as a power of k,

$$C(k) = |k|^{-\gamma}, \qquad 0 < \gamma < 1.$$

Long-range dependency is of special interest because it appears to be ubiquitous in nature (see Van Orden et al., 2003), and it has some intriguing properties such as fractal structure, i.e., a change in the time scale of measurement does not affect the distributional properties of a long-range dependent time series. Long-range dependencies have motivated a number of general theories about the sources of fluctuations in human behavior, and these theories were the focus of the current experiments.

However, before the theories are addressed, it must be noted that long-range dependencies can be difficult to distinguish from *short-range dependencies*, in which *C(k)* decreases exponentially with *k* (Wagenmakers et al., in press),

$$C(k) = \phi_1^{|k|},$$

where $-1 < \Phi_1 < 1$. Although *C(k)* declines more quickly in short-range dependent series compared with long-range dependent series (hence their names), the difference in rates of decline can be rather small. Nonetheless, short-range dependent series have very different properties (e.g., they can be generated by simple autoregressive processes), and they lead to different kinds of theories about fluctuations in human behavior. Therefore, Wagenmakers and his colleagues argued that empirical tests of long-range dependency must treat short-range dependency, rather than white noise, as the null hypothesis. Using this more stringent criterion, Wagenmakers et al. still found long-range dependencies in measurements of human behavior under a variety of experimental conditions. Their findings and analyses confirm that long-range dependency is a real phenomenon.

Explanations of Long-Range Dependence

Why do fluctuations in human behavior exhibit long-range dependencies? Only certain kinds of processes are known to produce long-range dependencies (for a review, see Wagenmakers et al., in press). The ostensibly special status of long-range dependencies has prompted researchers to search for general properties of human behavior that might explain their source(s). Here we review five explanations that have been offered.

Three Time Scales. Any specific observation of long-range dependence can be mimicked mathematically by the combination of three sources of white noise that operate on different time scales, each scale separated by an order of magnitude. In the context of perceptual and cognitive processes, Ward (2002) has suggested that unconscious, preconscious, and conscious processes may be three such sources of white noise whose combination is observed in fluctuations of human behavior.

While the transparency of this explanation is appealing, it is somewhat brittle because any three particular scales of white noise will mimic long-range dependence *only* for a single, particular scale of measurement (see Van Orden et al., 2003). What this means is that three-scale accounts must be fit to data posthoc. By contrast, true long-range dependence exists over all scales of measurement (within the limits of the system in question) due to its fractal structure. Long-range dependence in human behavior has, in fact, been found across a range of scales of measurement (for a review, see Van Orden et al., 2003).

Many Short-Range Dependencies. Granger (1980) showed that, under certain circumstances, the summation of many short-range dependent series can produce a true long-range dependent series. Ding, Chen, and Kelso (2001) proposed that long-range correlations found in timing tasks (and, by extension, in other kinds of tasks) may be the result of such summations. Their argument was based on the premise that cognitive processes are supported by large-scale networks of neural processes. Ding et al. reasoned that, in at least some cases, such neural networks will be characterized by large sets of short-range dependent processes. If the timing of behavior is driven by the summation of these processes, then fluctuations in timing will exhibit long-range dependence.

Ding et al. (2001) made the further statement that more difficult tasks require larger numbers of short-range dependent processes. This statement leads to the prediction that long-range dependencies will be stronger in more difficult tasks. In support of this prediction, they reported two timing tasks in which participants were asked to match their rates of tapping with the beat of a metronome. In one condition, participants were asked to tap in synchrony with the metronome. In another condition, participants were asked to tap at the midpoint between each pair of beats (i.e., to syncopate). Syncopation is a more difficult tapping task (e.g., less stable; see Kelso, DelColle, & Schner, 1990) compared with synchronization, and fluctuations in syncopated tapping exhibited stronger evidence of long-range dependence compared with synchronized tapping.

Mental Set. Gilden (2001) proposed that experimental tasks whose demands are relatively consistent across trials invoke a "mental set" in the participant. Gilden's definition of mental set entailed the repeated formation of mental representations necessary to perform the task. When the task is consistent, Gilden proposed that a dynamic of memory is created by this repetition such that memory components interact on multiple time scales. Under some circumstances, interactions of this nature have been shown to generate long-range dependencies (e.g., see Jensen, 1998).

Gilden (2001) left the nature of his proposed memory components unspecified, but his hypothesis was nonetheless formulated in sufficient detail to make a testable prediction. If mental set is broken by sudden changes in task demands, then the hypothesized dynamic of memory would not have an opportunity to form, and long-range dependencies in response fluctuations should disappear. Gilden tested this prediction by measuring series of reaction times to color or shape discriminations when each of these tasks was blocked, compared with a mixed condition in which participants had to switch between tasks across trials. Gilden found evidence of long-range dependence in the blocked conditions, but not the mixed conditions. These findings were consistent with his mental set explanation of long-range dependence.

Strategy Shifts. By definition, a long-range dependent series is stationary in the sense that its distributional characteristics do not change over time. However, a long-range dependent series can be difficult to distinguish from some kinds of non-stationary series that go through changes in their distributional characteristics over time.

It is probably true that any given experimental task can be performed in a number of ways, despite any and all efforts to make the task demands as explicit and precise as possible. If each means of performing a task is termed a "strategy", then it is very possible that a participant will change his or her strategy for performing a task over the course of an experiment. If strategy shifts occurred repeatedly over the course of measurement, they would have the potential to mimic long-range dependence. Wagenmakers et al. (in press) presented a computational demonstration of how strategy shifts (shifts in response criteria, in this case) can create non-stationary fluctuations in response times that mimic long-range dependencies.

Interaction-Dominant Dynamics. Van Orden et al. (2003) proposed that, at a very general level, humans are composed of many component processes that all interact on multiple time-scales. Their proposal was based on the fundamental idea that the structure and complexity seen in human behavior is a phenomenon of self-organization, and that self-organizing systems are ones that have interaction-dominant dynamics. They argued that it is these dynamics, intrinsic to human beings (and many other types of systems), that give rise to long-range dependencies in human behavior.

As general as they are, the ideas put forth by Van Orden and his colleagues (2003) lead to a testable prediction. If long-range dependence is the intrinsic signature of self-organization in human behavior, then any perturbations to behavior caused by external factors should disrupt the intrinsic dynamics, thereby obscuring their signature. Van Orden et al. argued that the results to date on long-range dependencies in human behavior (e.g., as cited in the other explanations listed here) are consistent with this prediction.

Current Experiments

Two experiments are reported here that were designed to explore a factor that was predicted to modulate the degree of long-range dependence in RT fluctuations. The factor was motivated by the explanations just listed. In particular, we tested whether sources of variability *external* to the participant would reduce the degree of long-range dependence in fluctuations of human behavior. Key presses were the measured behaviors, and sources of external variability were manipulated by the degree of *cue predictability*.

In Experiment 1, predictability in the timing of response cues was manipulated to be either completely predictable or completely unpredictable. When cues were predictable, fluctuations in response times were driven primarily by the participant. The cues themselves had little bearing on behavior because they were entirely redundant; participants knew that the next cue would always appear one second after the previous response (see Methods section). By contrast, when the timing of cues was unpredictable, the timing of responses had to be driven primarily by the cue itself, rather than any expectancies internalized by the participants.

If long-range dependence is internal to human behavior, then external variability should mask it. This idea is consistent with some previous explanations of long-range dependence (see Discussion section). This idea also leads to a further prediction that is quite counterintuitive. Participants were asked to press a key as soon as they perceived a cue. Thus, the task demands were satisfied when the finger moved down and the key made contact with its sensor. The task made no demands on when participants should lift their finger off the key. Therefore, fluctuations in the *durations* of key presses should be free to reflect internal variability, provided that the timing of the downward motion can be dissociated from timing of the upward motion. If so, we should observe *no* effect of predictability on the degree of long-range dependency in response durations.

In Experiment 2, sources of external variability were introduced by a different means. The *identity* of cues, instead of the timing of cues, was manipulated to be predictable or unpredictable. Two different cues signaled two different responses. Cue identity was made predictable or not by giving a preview or not of each upcoming cue. Analogous to the manipulation of predictability in Experiment 1, the preview manipulated the degree to which behavior was driven by the cues themselves, versus expectancies about the cues.

Experiment 1

Participants. Eighteen participants were recruited for the experiment. Sixteen were undergraduates who participated for course credit, and two were graduate students who were compensated for their participation.

Procedure. Each participant saw one block of predictable cues and one block of unpredictable cues, with block order counterbalanced across participants. Participants were instructed to press the space bar with their dominant hand as quickly as possible every time they saw an "X" flash on the screen. Demonstrations and practice blocks were given before each experimental block. Participants were instructed to wait till they saw an "X" before responding; if they pressed the space bar before a cue appeared, they heard a warning tone. Each block consisted of 1100 cues and took about 25 minutes to complete. The experimenter stayed in the room with the participant throughout the experiment. Participants took a short break between blocks.

Participants were seated about two feet away from a CRT monitor, and each cue appeared for about 50 ms in the center of the screen in Times New Roman font. A pair for visual flankers appeared immediately following each cue, and remained on the screen until the participant pressed the space bar. The flankers provided a redundant cue that the computer was awaiting a response (in case the participant missed a cue by accident).

Each subsequent cue was timed relative to the previous response. In the predictable condition, the next cue always appeared 1 s after the previous response was given. In the unpredictable condition, the timing of the next cue was sampled randomly from an exponential distribution with a mean of 1 s, a minimum of 1 ms and a maximum of 12 s. The exponential distribution was used because it has a flat hazard function, which means that the probability of receiving a cue was constant as a function of wait time (Simpson et al., 2000). The time from cue to key press was recorded (response time), as well as the length of time that participants pressed each key (response duration).

Results

To illustrate the time series structures that were typically observed, the series of response times for one participant in the predictable and unpredictable conditions are shown in Figure 2. The series of response durations for this participant are shown in Figure 3.

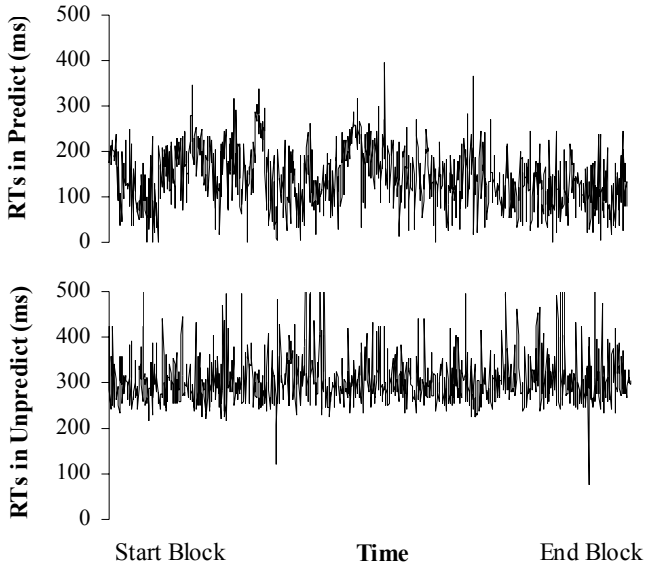


Figure 2. Response times for one participant in the predictable (top) and unpredictable (bottom) conditions of Expt 1 (responses above 500 ms have been truncated).

Averaged across participants, the percentage of anticipatory responses was 1.99% in the unpredictable condition, and 3.9% in the predictable condition. All anticipatory responses were removed from the analyses. The mean correlation of response times with response durations was r = .02 in the unpredictable condition, and r = -.21 in the predictable condition.

Spectral analyses are standardly used to measure the degree of long-range dependence in a time series, and we adopted the method of spectral analysis described by Holden (unpublished; also see Gilden, 1997). In particular, outliers were first removed from each time series (values > 1000 ms or outside 3 SDs of each participant's mean for each measure in each condition). Then, linear and quadratic trends were removed to avoid dependencies caused by practice or fatigue. A power spectrum was then computed over 1024 of the remaining data points, and log frequency was regressed against log power. The slope of this regression line in log-log coordinates was used as a measure of serial dependence: more negative slopes correspond to stronger degrees of serial dependence.

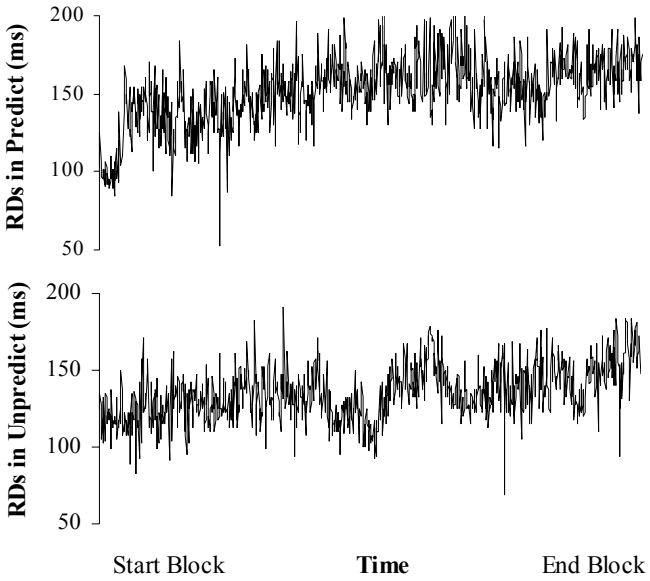


Figure 3. Response durations for one participant in the predictable and unpredictable conditions of Expt 1.

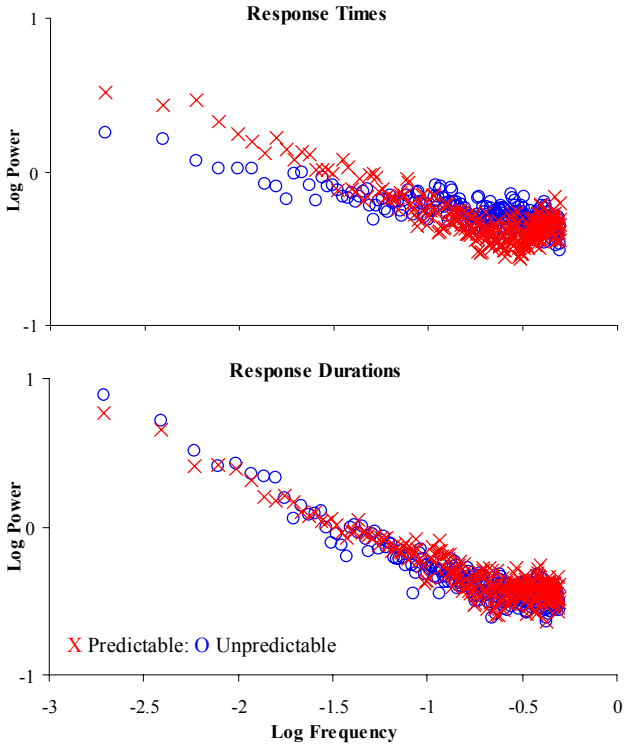


Figure 4. Aggregate spectral plots for Expt 1.

The aggregate power spectra, averaged across participants for each condition, are plotted in Figure 4. For response times, slopes in the predictable condition were reliably more negative than slopes in the unpredictable condition, $t(17)$= 4.26, $p < .001$. For response durations, there was no reliable difference in slopes, $t(17) < 1$. Moreover, slopes for

response durations were reliably more negative than slopes for response times, $t(35) = 7.21$, $p < .001$.

Experiment 2

Participants. Eighteen undergraduates participated in the experiment in exchange for course credit.

Procedure. The procedure was identical to that used in Experiment 1, except for the following changes. The response cue was either '>' or '<', and participants were instructed to press the right arrow key for the former, and the left arrow key for the latter. Flankers appeared on either side of the response cues as signals to respond, and the flankers always appeared 1 s after the previous response was given. In the preview condition, the next response cue always appeared immediately following the previous response; thus, participants had 1 s to process the cue and prepare their response. In the no-preview condition, each cue appeared in conjunction with its signal to respond; thus, participants had to process the cue and choose their response as quickly as possible.

Results

Averaged across participants, the percentage of anticipatory responses was .03% in the unpredictable condition, and .45% in the predictable condition. The percentage of errors was .80% and .27%, respectively. All anticipatory responses were removed from the analyses, but the few errors were retained. The mean correlation of response times with response durations was $r = .05$ in the unpredictable condition, and $r = -.08$ in the predictable condition.

The aggregate power spectra are plotted in Figure 5. For response times, slopes in the predictable preview condition were reliably more negative than slopes in the unpredictable no-preview condition, $t(17) = 2.31$, $p < .05$. For response durations, there was a small but unreliable difference in slopes, $t(17) = 1.80$, $p < .09$. Moreover, slopes for response durations were reliably more negative than slopes for response times, $t(35) = 3.55$, $p < .001$.

Discussion

Two experiments were reported in which long-range dependencies were measured as a function of cue predictability. Results showed greater degrees of dependency in series of response times when the cues were predictable, both in terms of timing and identity. By contrast, results showed large and comparable degrees of dependency in all series of response durations. The observed dissociation between response times and response durations was consistent with the idea that external sources of variability mask the long-range dependence that is intrinsic to human behavior.

It also appeared that the effect of predictability in cue timing (Experiment 1) was stronger than that in cue identity (Experiment 2), albeit further experiments are necessary to bear this out. One possible explanation is unpredictable timing introduces more external variability compared with unpredictable choice responding. However, to test this explanation, one would need to develop a more explicit means of parsing internal and external sources of variability.

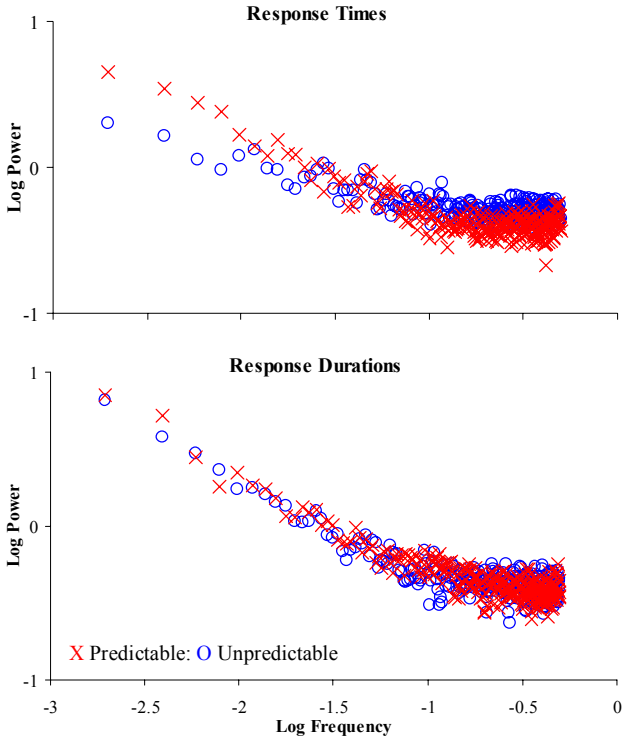


Figure 5. Aggregate spectral plots for Expt 2.

It is important to note that the hypothesis of long-range dependence was not explicitly tested against the short-range alternative in the current data. We did not conduct these tests because the IDD predictions could be tested without them. However, the long-range/short-range distinction is important, and we plan to address this issue in future work.

How do the current results bear on the five explanations of long-range dependence outlined in the Introduction section? We address this question here for each explanation in turn.

Three Time Scales. Sources of white noise on three different time scales could be used to mimic the long-range dependencies (or lack thereof) for each participant in each condition of the two reported experiments. However, these parameter fits would be posthoc, and they would offer no insight into the differences in degree of long-range dependence between experimental conditions.

Many Short-Range Dependencies. As noted earlier, this explanation leads one to predict that greater degrees of long-range dependence should be found in more demanding tasks. The unpredictable conditions were clearly more demanding because their mean RTs were much greater. However, the unpredictable conditions showed *lesser* degrees of long-range dependence compared with the predictable conditions. Moreover, summations of short-range dependencies appear to offer no insight into the observed differences in dependencies between response times and response durations.

Mental Set. Gilden (2001) proposed that long-range dependencies should be weaker when a person's mental set is repeatedly broken or interrupted. One could imagine that participants were able to maintain a more stable mental set in the predictable conditions compared with the unpredictable conditions, which would make the current results consistent with the mental set explanation. It is less clear how the mental set explanation would apply to the differences in long-range dependence between response times and response durations. One would presumably have to propose that these behaviors are governed by different mental sets, but given the close physical relationship between a button press and its release, the idea of different mental sets seems implausible. The bottom line is that the mental set explanation is not yet formulated to the point where it might offer insight into the current results.

Strategy Shifts. Wagenmakers et al. (in press) conjectured that participants might be more apt to shift strategies, and therefore exhibit long-range dependencies in their behaviors, when they are bored. Participants were almost certainly bored in all of the current experimental conditions, but one could argue that the predictable conditions were more boring than the unpredictable ones. If so, the finding that long-range dependencies in response times were stronger in the predictable conditions is consistent with the strategy shifts explanation. However, one would have to apply this explanation to response durations as well, and there was no such effect on long-range dependencies in this measure. It remains to be seen whether a strategy shift explanation could be made to account for these results.

Interaction-Dominant Dynamics. This explanation states that long-range dependencies come from the interdependent dynamics that underlie the self-organization of human behavior. These dynamics are hypothesized to be perturbed by external forces. If sources of external variability are thought of as external forces, then all the results reported herein are consistent with the interaction-dominant dynamics explanation. Predictability was a force on response times, but not response durations, because the task made demands on the former but not the latter.

In conclusion, the current results are, for the time being, most consistent with the interaction-dominant dynamics explanation. Of course, these explanations are all in their infancy; it would be an overstatement at this point to refer to them as theories. Be that as it may, the results were clear and far from trivial to explain. We believe that further empirical and theoretical investigations into the sources of long-range dependence in human behavior will prove to be valuable to studies of perception and cognition.

Acknowledgments

We thank Anthony Novak for his assistance with data collection, and Guy Van Orden and Jay Holden for helpful conversations about the theory and methodology surrounding long-range dependence. The work was funded in part by NIH Grant MH55628, and NSF Grant 0239595.

References

Aks, D. J., Zelinsky, G. L., & Sprott, J. C. (2002). Memory across eye-movements: 1/f dynamic in visual search. *Nonlinear Dynamics, Psychology and Life Sciences, 7*, 161-180.

Chen, Y., Ding, M., & Kelso, J. A. S. (2001) Origins of timing errors in human sensorimotor coordination. *Journal of Motor Behavior, 33*, 3-8.

Ding, M., Chen, Y., & Kelso, J. A. S. (2002). Statistical analysis of timing errors. *Brain & Cognition, 48*, 98-106.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science, 8*, 296-301.

Gilden, D. L. (2000). Cognitive emissions of 1/f noise. *Psychological Review, 108*, 33-56.

Granger, C. W. J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics, 14*, 227-238.

Jensen, H. J. (1998). *Self organized criticality*. Cambridge, England: Cambridge University Press.

Holden, J. G. (2003). *Gauging the fractal dimension of cognitive performance*. Unpublished manuscript, California State University, Northridge, CA.

Kelso, J. A. S., DelColle, J. D., & Schoner, G. (1990). Action-perception as a pattern formation process. In M. Jeannerod (Ed.), *Attention and performance XIII* (pp. 139-169). Hillsdale, NJ: Erlbaum.

Simpson, W. A., Braun, W. J., Bargen, C., & Newman, A. J. (2000). Identification of the eye-brain-hand system with point processes: a new approach to simple reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1675-1690.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*, 331-350.

Wagenmakers, E., Farrell, S., & Ratcliff, R. (in press). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychonomic Bulletin and Review*.

Wagman, J. B., Dahle, C., & Schmidt, R. C. (2002, May). Perceptual learning dynamics. Paper presented at the *International Conference on Brain and Behavior*, Delray Beach, FL.

Ward, L., & Richard, C. M. (2001). *$1/f^{\alpha}$ noise and decision complexity*. Unpublished manuscript, University of British Columbia, Vancouver, British Columbia, Canada.

Ward, L. M. (2002). *Dynamical cognitive science*. Cambridge, MA: MIT Press.

Linguistic Untranslatability vs. Conceptual Nesting of Frames of Reference

Giovanni Bennardo (bennardo@niu.edu)
Department of Anthropology and Cognitive Studies Initiative, Northern Illinois University
DeKalb, IL 60115 USA

Abstract

This work focuses on the concept of frame of reference. Levinson (2003) suggested that spatial information encoded in one frame of reference cannot be translated into another one. While this is partially true in language, I argue for a nesting relationship between frames of reference at the conceptual level. A set of spatial concepts suggested by Lehman & Bennardo (2003) informs this investigation. In closing a new typology of frames of reference is proposed.

Introduction

The concept of frame of reference (FOR) is widely used in the literature about the mental and linguistic representations of spatial relationships. After a review of the terminologies used in different disciplines such as philosophy, linguistics, psycholinguistics, developmental and behavioral psychology, brain sciences, and vision theory Levinson (2003) proposed a definition of the concept and a typology of frames of reference.

A FOR is defined as a system of three coordinated axes that create a 3-dimensional space within which spatial relationships are established cognitively and expressed linguistically. Levinson's typology of FOR includes three systems labeled relative, intrinsic, and absolute. When a FOR is realized linguistically, the information coded in one FOR (e.g., 'the ball is behind me') is not translatable into another (e.g., 'the ball is south of me'). While I agree with most of the discussion presented by Levinson, I find problematic the untranslatability issue. I suggest that untranslatability only holds between linguistically instantiated FORs, while at the conceptual level they are nested into each other.

This investigation uses a set of spatial concepts found in Lehman & Bennardo (2003). This conceptual apparatus is the result of analyses conducted on English spatial prepositions, and languages like Burmese, Thai, Italian, and Tongan (Polynesian). After sketching the apparatus, the three FORs are analyzed ending with a suggestion about their conceptual contents and a new typology.

The conceptual apparatus

A computational approach to the general architecture of cognition was adopted to arrive at the set of spatial concepts suggested by Lehman & Bennardo (2003). Within this approach, cognition is conceived as computational (cf. Ballim & Wilks, 1992), thus generatively 'abstract'. Only the characteristics of the computational, or, relational spaces that make up what we call 'cognition' are reiterated in each cognitive module and not the specific characteristics of the substantive content that instantiate these 'abstract' relationships.[1]

A computational approach to cognition can be proposed by accepting compositionality without embracing a Fregean (logico-positivist) point of view and by turning to the domain of mathematics (e.g., algebra and geometry). In mathematics the primitives of a system are a set of axioms. These axioms generate indefinitely many theorems and each theorem can establish a foundation for yet another theorem. Furthermore, theorems may share parts with other theorems in a redundant manner. The set of relational properties of any cognitive system could be, then, nothing but a theorem derived from a set/s of other theorems. Such a system is compositional by definition.

The linguistic analyses in Lehman & Bennardo (2003) yielded the following set of spatial concepts.[2]
State: Object; Place Or Locus; Neighborhood: Vicinity, Contact, Interiority; Motion: Time; Direction; Path: Beginning, Body*, End, (Direction)*; Verticality: Angle: Unit, Quantity (+ or -); Horizontality: Visibility, Left or Right; Center; Part*. (*conceptual content of *Vector*)

Some concepts are not primitives, but rely on other concepts of the same group to function as their axioms. This is the minimal set of axioms that is necessary to account for the theorems (e.g. prepositions, directionals, spatial nouns) that make up the representations of spatial relationships in the languages analyzed.

The concept of Object is used with the meaning of any entity existing in a possible world, either concrete or abstract, e.g. table, idea. The place of an Object is the actual amount of Space that it occupies. In other words, a Place is the set of all points within the boundary of an Object (including the boundary points). The Locus of an Object in projective geometry is defined as the collapse of a Place onto any of its interior points. Then, a Locus is a neighborhood of possible projection points, the lower limit being one point. Thus, while a Place is defined by the size, shape, and specific geometry of the Object, a Locus is not and can be arbitrarily reduced to a point.

The concept of Neighborhood includes the concept of Vicinity (more than zero distance) between two Objects, the concept of Contact (zero distance) between them and the concept of Interiority, or, one Object in the interior of another. The Neighborhood's border is pragmatically determined. These concepts make up the concept of State.

The concept of Motion is an ordered sequence (consequently, with a Direction) of Places (of an Object)

[1] Hirschfeld & Gelman (1994) draw a similar distinction between 'module' and 'domain.'

[2] From now on a concept is indicated by initial capital letter.

in Time, bounded by two Places without either left or right directionality in a disjunctive fashion and never missing both. The concept of Path, instead, is a geometrical (purely spatial) description of motion 'abstracted' from Motion. The focus is not on the moving Object, but on the ordered sequence of Places, now considered as Loci. The concept of Motion is inextricably tied to Time, but the concept of Path is partially free from it. In fact, we can indicate a Path at $Time_1$ and then indicate another Path at $Time_2$ and state that they are the same without incurring a contradiction as would happen if the two parts of the comparison were two instances of Motion. The instances of time used in the construction of a Path are not unique, but they are repeatable.

Two features that Path also shares with Motion are ordered sequence and boundedness. The interior points of a Path are an ordered sequence of Loci with a Direction, that is, they are Vectors with a finite magnitude. This magnitude we call its Body and consists of a set of Loci whose members may at a limit be one, thus, overlapping with the first constitutive Locus. The boundary of a Path consist of two Loci, a Vector that lacks left directionality (Beginning), and one that lacks right directionality (End). Object and Place (axioms of State) participate in the construction of Motion. Locus, instead, participates only in the construction of Path. Thus, the difference between Place and Locus is used to separate the temporally bound Motion from the spatially bound Path.

Verticality and Horizontality were not analyzed in as much detail as State and Motion, and only some conceptual components are indicated. First, Object, Locus, and Vector (a Beginning, a Body or magnitude, and a Direction) participate in their composition. The concepts indicated for Verticality are Angle and Quantity (Increasing or Decreasing). The instantiation of one or other type of Quantity will determine the 'up' or 'down' Direction of a Vector. Angle and Quantity are also part of the concept of Horizontality together with those of Visibility and Left or Right. Visibility contributes to the construction of a 'front-back axis.' After this, Left or Right can be constructed. Finally, the two concepts of Center and Part were added after the analyses of Tongan directionals and spatial nouns (Bennardo, 2000).

The conceptual content of the Relative FOR

Levinson (2003) defines a relative FOR in this way:

> This [a relative FOR] is roughly equivalent to the various notions of viewer-centered frame of reference mentioned above (e.g. Marr's 2.5D sketch, or the psycholinguistics' 'deictic' frame). But it is not quite the same. It presupposes a 'viewpoint' V (given by the location of a perceiver in any sensory modality), and a figure and ground distinct from V. It thus offers a triangulation of three points, and utilizes co-ordinates fixed on V to assign directions to figure and ground. (Levinson, 2003, p. 43)

He continues by pointing out that the viewpoint V does not necessarily coincide with the speaker even though deictic uses can be considered 'basic' or 'prototypical.' (Levinson, 2003, p. 43)

The axiomatic distinction between the figure F (sensory input, Object) and the viewer V (or cognizer in any sensory channel) is conceivable as a primary one, but the distinction between viewer V and ground G can be dispensed with and be regarded as constructed at a later stage in the ontogenic sequence. Both the research on the visual system (Marr, 1982) and on the developmental sequence (Cohen, 1985; Pick, 1993; Bowerman & Levinson, 2001) point towards the primacy of a stage in which viewer V and ground G are conflated. It is exactly the capacity to assign independent sets of coordinates to objects that marks one of the milestones of cognitive development (Piaget and Inhelder, 1956). I feel, then, justified in suggesting this definition for the relative FOR.

A Basic relative FOR is one centered on the speaker, viewer, cognizer (viewer). From the viewer three axes (or six vectors) are constructed, one vertically and two on the horizontal plane (front-back and left-right). In other words, the viewer can be thought of as a point and as such it implies a field (space) around it. This field will be oriented. This orientation process takes into consideration gravity and several bodily characteristics, both static (orientation of face, eyes, etc.) and ambulatory (habitual direction of movement). The viewer necessarily (ontogenically) maps these axes onto himself/herself, that is, considers himself/herself the origin of these axes.

Any Object in this field will be described in relation to the viewer, thus viewer V and ground G are considered as conflated, rather, they have not yet been cognitively constructed as separated. If the viewer moves in any direction, the axes will move accordingly, keeping their origin on the viewer. These axes are an abstraction from 3-D and conic spaces that originate on the viewer and where the angular limit is 180˚. Each axis, in fact, stands for a collection of possible axes whose limits are provided by the following (on three sides) relevant axes.

The appearance of a second Object in the field of the viewer creates the double possibility of treating this latter in direct relationship with the viewer, thus, continuing to map the axes on the viewer, or relating the second Object to the first one. This latter case entails the possibility of assigning orienting axes to the first Object, thus making it function as if it were the viewer. The orientation of the axes mapped onto this Object are the same as that of the axes the viewer had mapped on himself/herself. That is, the coordinates of the viewer's field can be kept constant. The second Object (figure) is described as in relationship with the first Object (ground). The front or 'away' Vector of the viewer is now divided in two parts by the first Object. Then, a new possibility is created. The ground front-back axis may keep the same orientation of the viewer's field, thus, we get the Translation subtype of relative FOR (e.g., 'the ball is in front of [beyond] the tree'). Or the front and back mapping can be flipped so that the front of the first Object (ground) faces the viewer, thus, yielding the Reflection subtype of relative FOR (e.g., 'the ball is in front of [facing viewer] the tree').

Both the Translation and the Reflection subtype of relative FOR are subtypes of the Basic relative FOR. In fact, their left and right assignments are congruent with those of the viewer. In other words, the first Object or ground is not yet considered as a point with an oriented field of its own, but is still tied to the field of the viewer.

It is not possible to arrive at the construction of the Translation and Reflection subtypes without using (consciously or unconsciously) a Basic subtype of relative FOR. In fact, there would be no axis to 'translate' or 'reflect' at all without having already constructed one in advance. And this can only have happened through the use of a Basic relative FOR.

This typology is suggested: a Basic, a Translation, and a Reflection relative FOR.[3] The label Basic highlights the ontogenic primacy of the construction in which viewer and ground G are conflated, that is, a set of coordinates is mapped onto the viewer by him/herself. The other two subtypes are derivative from the Basic and represent a move towards recognizing that Objects have relationships among themselves and not just with the viewer.

Let us now look into the conceptual content of the Relative FOR. We already know from the content of the conceptual apparatus that both vertical and horizontal axes are only the interaction of a subset, labeled Vector, of the concept of Path (Beginning, Body and Direction) with the two concepts of Verticality and Horizontality. The Beginning of these Vectors is a unique anchoring point (the viewer) that is a Locus because its geometric features are not relevant in the construction of the relative FOR (also, the Beginning of a Path is by definition a Locus). Another participating concept is Orthogonality that is the distinguishing factor between vertical and horizontal axes, and between front-back and left-right axes. Orthogonality contains Angle and Unit (degree) with a fixed Quantity attached to this last (90˚ degrees).

All the conceptual content so far listed brings with it other more finely grained content, and, specifically, Vector (as a subset of Path) and Locus. Moreover, this FOR assigns front and back to the Object that becomes the ground by mapping the viewer's coordinates onto it (see the Translation and Reflection subtypes). This mapping can be done by simply applying the 'repeat function' (as for the Translation subtype) or by applying the 'repeat function' and then letting Visibility determine which side is front or back (as for the Reflection subtype). The side that is not visible (beyond the Object) is the back in the same way as it is done for one's body.

The concept of Figure (any possible Object) is also a participant in this construction. Is this Object to be considered a Locus or a Place? Do its geometrical characteristics matter in constructing a FOR? Perceptually these geometrical characteristics are available, but do they play a role in the construction of the FOR? The answer is 'no.' Knowledge about the geometrical characteristics of the Figure does not seem to play any role in the construction of a FOR (see Talmy, 2000). What is relevant, instead, is the fact that a point, the Object, is being picked up in the world by means of a 'choice function' (clearly not provided by perception, but by our intentional thinking) and later put in a spatial relationship with either oneself or another Object according to a specific set of coordinates (in this case a relative FOR).

What does it mean to 'pick an object' in the world? In order to 'see'[4] an object our line of sight has to 'meet' it in the world. In order to think of this object as separated from our self, our line of sight has to be conceived as first leaving our eyes, then penetrating the outside world and finally meeting the object. In other words, the actual construction of any Object requires our use of the concept of Path, with a Beginning (self), Body (penetration of the world outside self), and an End (object in the world). It follows that Path needs to be postulated as participating in the construction of the Relative FOR.

The conceptual difference between the Basic and the Translation and Reflection subtypes of the relative FOR is the following. The two subtypes consider the viewer and two Objects (instead of only one), iteratively employing the concept of Path, a process that needs the use of the 'repeat' function. The coordinates of the oriented field of the viewer are still mapped onto the viewer. In the Translation type the front Vector is kept constant in orientation, while in the Reflection type its orientation is changed as a consequence of the salient use of the concept of Path (from viewer to Object) and Visibility.

The assignment of front and back that distinguishes between the Translation and the Reflection subtypes is left open to cultural variations. Since all the axiomatic conceptual material is already available (Locus, Path, Vector, etc.), the 'repeat' function can be arbitrarily applied to any of these concepts. However, minimally, the Translation subtype is conceptually simpler. In fact, it does not require the use of the concept of Visibility. Then, it is the salience of Visibility in specific cultures that may determine the preferred instantiation of one subtype over the other. Cases in point are Dutch (and English) speakers who habitually use the Reflection subtype (Levinson, 2003); Hausa and Tongan speakers who habitually use the Translation subtype (Hill, 1982; Bennardo, 2000); and Japanese speakers who use both (Levinson, 2003). This is a summary of the conceptual content of the Relative FOR. *Object; Locus; Path; Six Vectors, each with a Beginning (Locus of viewer, or anchor point), Body, Direction; Verticality; Horizontality; Visibility; Orthogonality, with an Angle, Unit (degree), Fixed Quantity (90˚ degrees).*

To this we need to add the 'choice function' used to construct an Object. For the Translation and Reflection subtypes, we must add also the 'repeat function' yielding two Objects and the construction of the front and back Vectors onto one of them. Both functions are axiomatic cognitive processes. For the Reflection subtype, repeated use of the concept of Visibility must be added.

The Intrinsic frame of reference

An Intrinsic FOR is one centered on an Object that is not the viewer. From the Object, three oriented axes (or six vectors) are constructed, one vertically and two on the horizontal plane. Any Object in the space defined by these coordinates is described in relation to the Object from which the space was constructed. When the Object

[3]The 180˚ rotation subtype of relative FOR in Levinson (2003) is not indicated here because within this work that subtype is considered as an instantiation of an intrinsic FOR.

[4]This discussion is limited to only visual input.

moves, the axes will move accordingly keeping their origin on the Object and the assigned orientation as well.

What differentiates the Relative and the Intrinsic FORs is that the Beginning of the Vectors is not from the viewer, but from an Object other than viewer. We have already seen that this is also the case for the Translation and Reflection subtypes of relative FOR. What is it that distinguishes these latter two FOR from the Intrinsic one?

The difference lies in the quality of the oriented field that is constructed for the Object or ground. This field is completely independent in orientation from the viewer; it is in other words a new separate field from that of the viewer. This difference has very important consequences, among which the most relevant is that the description of the spatial relationship between two Objects will be freed from references to the viewer. This, however, does not mean that the field of the viewer has not been used to construct the new field. Specifically, when we express linguistically a spatial relationship by utilizing an Intrinsic FOR, conceptually we must have used a Basic relative one in order to arrive at the construction of the first Object (Figure) and the second Object, making this latter a ground by constructing from it oriented axes.

What remains to be seen is how the axes of this new field are oriented. Typically the following three concepts have been associated with the Object that functions as ground in order to orient the axes mapped onto it: Animacy, Habitual Direction of Motion, and Habitual Use (Herskovits, 1986; Talmy, 2000). It is understood then that these three concepts participate in the construction of the Intrinsic FOR in a disjunctive fashion. That is, usually only one is necessary. It is, then, a specific characteristic of the Object picked to function as ground that determines the orientation of the axes mapped onto it. Only one axis need to be oriented, typically the frontal one, and the orientation of the others will follow.

Regarding the conceptual content of the Intrinsic FOR, all the content suggested for the Relative FOR needs to be postulated for the Intrinsic FOR as well. We also have to include the 'choice function' and the 'repeat function.' New concepts to be added are Animacy, Habitual Direction of Motion, Habitual Use (disjunctively used, even though they may overlap), and finally Part. In fact, the Object onto which the coordinates are mapped, needs to be assigned a 'front.' That is, a minimal subdivision of the Object into parts must be done.

The Absolute frame of reference

An Absolute FOR is neither centered on the viewer nor on an Object. First, the two Vectors related to the vertical axis are constructed. Second, on the horizontal plane one or more Objects (e.g., areas, points, landmarks) in the field of the viewer are chosen as orienting points. Third, either the viewer or any Object in its field is put into relationships with these Objects or fixed points.

Two examples of this system are the one that uses cardinal points, and the one that uses landward-seaward directions used by the speakers of many Oceanic languages. In many other cases the environmental features selected differ profoundly and may range from a mountain to a lake, or from a river to a building.

The process of selecting fixed orienting points in the environment requires minimally the activation of the Relative FOR. Once these fixed points have been conceptually established and agreed upon socially, these same points may function as an orienting framework between either one Object in the field of the viewer and one of the fixed points (e.g. North) or between any two Objects in the field of the viewer and one of the fixed points. We have already seen in the previous discussion of the other two types of FOR that the process of selecting/choosing an Object to function either as figure or ground implies the use of the concept of Path. For the construction of the absolute FOR, then, we need minimally either one or two Paths required for the construction of the Object or Objects to be put into relationship with any of the orienting fixed points. To these we have to add two (for the Oceanic system) or four (for the cardinal points system) Paths for the choice of the fixed points of reference.

Table 1 summarizes the conceptual content of the various types of FOR. A capital X indicates the presence of a concept in the construction of a FOR. For the concepts of Object, Path and Vector a number indicates how many times the concept is minimally used.

Table 1: The conceptual content of FORs

Concept/Axiom	Relative			Intrinsic	Absolute
	Basic				
		Transl	Reflect		
Locus	X	X	X	X	X
Object	1 + V	2 + V	2 + V	2 + V	1/2+2/4+V
Path	1	2	2	2	3/5 or 4/6
Vector	6	6	6	10	6
Verticality	X	X	X	X	X
Horizontality	X	X	X	X	X
Orthogonality	X	X	X	X	X
Visibility			X		
Part				X	
Animacy**				X**	
Hab Dir Mot**				X**	
Habitual Use**				X**	
Choice Funct*	X	X	X	X	X
Repeat Funct*		X	X	X	X

*These two are cognitive processes.
**Only one is necessary.

From Table 1 it can be seen how the conceptual axiomatic content of the Basic relative FOR is properly contained in its entirety in all the others, both subtypes (Translation and Reflection) and types (Intrinsic and Absolute). The Intrinsic and the Absolute are both derived from the Relative, although not in an ordered sequence. The Relative FOR, then, is suggested as an axiom for both the Intrinsic and the Absolute ones.

The Intrinsic and Absolute FORs are made of two different sets of concepts. The Intrinsic FOR expresses more attention to the nature of the Object functioning as ground (see the participation of the concepts of Part, Animacy, Habitual Direction of Motion, Habitual Use in

Table 1). The Absolute FOR, instead, expresses greater attention to the nature of the field (see the participation of a greater number of Objects and Paths in Table 1).

These findings are perfectly congruent with those of Baayen & Danziger (1994) and Levinson (2003) regarding a preferred use of the Intrinsic FOR by speakers of Mayan languages, where an extremely elaborate vocabulary also exists for describing parts of objects. Similar congruency can be highlighted with the findings of Levinson (2003) concerning the preferred use of the Absolute FOR by speakers of Australian Aboriginal languages where a very elaborate system of naming landmarks in one's environment has also been reported.

Finally, we look closely at the issue of 'untranslatability' among the various FORs suggested in Levinson (2003, p. 57-59). When we consider FORs as instantiated into linguistic expressions, it is true that in principle only two cases of translation are possible from one FOR to another (i.e., from either Absolute or Relative to Intrinsic). Do we deduce that there is 'untranslatability' among FORs at the conceptual level? Our discussion points towards a negative answer. In fact, the conceptual content of the Relative FOR has been suggested as an axiom for the Intrinsic and the Absolute ones. Thus, if at the linguistic level we find 'untranslatability' between FOR, at the conceptual level we find 'nesting.' Besides, the direction of the translatability from Relative and Absolute to Intrinsic correlates with one independent field in the former vs. two independent fields in the latter.

A Radial subtype of the absolute FOR

Bennardo (1996) reported the results of an investigation of the uses of FORs in Tongan language, spatial cognition, and culture. During this investigation a Radial subtype of absolute FOR was suggested as having a privileged status. This FOR consists in positing a center in one's field out of which movement is conceived either centripetally or centrifugally on any plane. This finding requires a reexamination of the typology of FORs. In particular, we need to look closer at the subtypes of the absolute FOR: Radial, Single Axis, and Cardinal Points.

For each of the three subtypes there are two cases to be considered. The first is when the ground is the viewer, e.g. "X is north of me.' The second is when the ground is an Object different from the viewer, 'e.g. X is north of Y.' Each case yields different conceptual content.

In the first case we need a Center that is the viewer and an Object (Figure). A Path from the viewer to the Object is also required as well as (minimally) a Vector made up of the Body of a Path and its End (centrifugal movement) or Beginning (centripetal movement). Either the End or the Beginning of this Path would be co-indexed with the Center. In the second case we have to add a second Object, which will function as Center, and a Path that is used to determine this Center (or second Object). The difference between the two cases is crucial. Choosing a Center different from viewer, makes possible the construction of a second field different from the one constructed around the viewer.

The conceptual content for the Single Axis subtype consists, in the first case, of one Object (Figure) plus two Objects (the two ends of the axis), the viewer, three Paths (from the viewer to the three Objects), and six Vectors (up, down, front, back, left, and right). In the second case, an Object for the new ground Object and a Path for its construction are added. A new field different from the viewer's is not constructed.

The conceptual content for a Cardinal Points subtype consists, in the first case, of one Object (Figure) plus four Objects (the cardinal points), the viewer, five Paths (from the viewer to the five Objects), and six Vectors (up, down, front, back, left, and right). An Object is added for the new ground Object and a Path for its construction. A new field different from the viewer's is not constructed.

Table 2: Conceptual content for types of Absolute FOR

	Concept		
Subtype of Absolute FOR	Object	Path	Vector
Radial 1	1 + V	1	1
Radial 2	2 + V	2	1
Single Axis 1	1 + 2 + V	3	6
Single Axis 2	2 + 2 + V	4	6
Cardinal Points 1	1+ 4 + V	5	6
Cardinal points 2	2 + 4 + V	6	6

In Table 2, the conceptual content of the Radial subtype is the simplest. The contents of the two Radial subtypes are also simpler than the content of the Basic subtype of the relative FOR in Table 1. Consequently, the axiomatic relation between the Relative, the Intrinsic, and the Absolute FORs needs some further attention.

A new typology of Frames of Reference

In discussing the relationships between the types and subtypes of FOR three parameters are considered. The first is the magnitude of the conceptual content, that is, the number of concepts necessary to derive each theorem. The second is the reference that will be made to axiomatic relationships. When the content of a FOR is completely contained in another, then the former will be considered an axiom of the latter. The third is the emergent properties that each FOR displays. Namely, it must be considered if a FOR is based on the construction of one or two fields.

The minimal conceptual content and the construction of only one field associated with the Radial 1 subtype of the absolute FOR make it the choice as the most basic. This FOR, then, is an axiom for all the other types and subtypes. Its great simplicity makes it highly context bound and hence very unlikely to be the only one that any individual/culture will have. Nonetheless, it represents a minimal stage of spatial organization assigned to the external world. Evidence from languages around the world suggests that this system is always used, i.e., in demonstrative systems.

Looking for a FOR at the second stage of complexity, or more precisely, the first theorem derived from a set of axioms whose content is at a limit only the Radial 1 axiom, we confront two options. The first is to choose the Radial 2 subtype of the absolute FOR, simple in conceptual content, though it uses two fields. The second

is the Basic relative FOR that is more complex conceptually (it needs six Vectors instead of one), but uses only one field. A decision is not strictly necessary at this juncture; both options are viable. Empirically, there are language speakers that choose to use prevalently one option only (e.g. English speakers choose a Basic relative FOR), and other that choose both (e.g. Tongan speakers).

We have stated that the Relative FOR functions as an axiom for the Absolute and the Intrinsic FOR, and for two subtypes of the Relative FOR. These latter keep the single field feature, but increase their conceptual content because of their complex treatment of the front-back axis.

The Single Axis and Cardinal Points subtypes of the absolute FOR and the Intrinsic FOR are obtained in substantially different ways. The two Absolute FORs represent an increased conceptual content from the Relative FOR and use a single field. This is confirmed by the fact that they both use the vertical axis. The Radial 1 and 2 subtypes did not have it in their conceptual content.

The intrinsic FOR is obtained by an increased conceptual complexity due to two other factors (besides the addition of the vertical axis). The first is a closer attention devoted to the Object that functions as figure. The second is the construction of two fields (the viewer's and the figure's). We have seen that the construction of two fields is part of the conceptual content of the Radial 2 subtype of the absolute FOR. Then, we suggest that the conceptual content of the intrinsic FOR is derived from the Basic subtype of the relative FOR, from the Radial 2 subtype of the absolute FOR, and from conceptual characteristics of the Object/Figure.

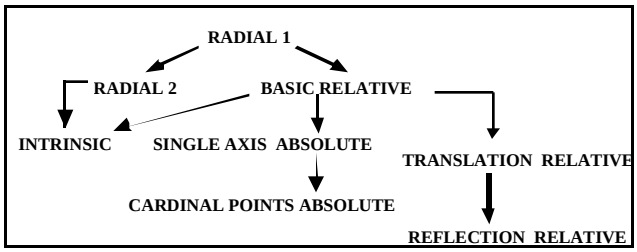


Figure 1: A typology of Frames of Reference

The arrows in Figure 1 indicates that the FOR receiving the content of another FOR treats this latter as an axiom of its conceptual content. Further conceptual material is added at each stage. Thus, the necessity of a new label for that particular type of FOR.

Conclusion

The first part of this work was devoted to the introduction of the conceptual apparatus that is the major theoretical tool employed in the analyses of the conceptual content of FORs. Each member of the typology of FORs suggested by Levinson (2003) was later analyzed. and a primary revision was suggested. Nesting of FORs was proposed at the conceptual level instead of untranslatability. Then, a Radial subtype of absolute FOR was introduced. This made clear that a further revision of the typology was needed. Finally, the revision resulted in the proposed typology of FORs in Figure 1. It is believed that this typology can be useful for further investigation of FORs.

Acknowledgement

My special thanks go to Kris Lehman and Janet Keller for reading through various drafts of this work. Mistakes, misrepresentations or fallacies are mine.

References

Baayen, H., & Danziger E. (Eds.) (1994). *Max-Planck Institute for Psycholinguistics: Annual Report 14, 1993*. Nijmegen, The Netherlands: Max-Planck Institute for Psycholinguistics.

Ballim, A., & Wilks Y. (1992). *Artificial Believers: The Ascription of Belief*. Hillsdale, NJ: Lawrence Erlbaum.

Bennardo, G. (1996). *A Computational Approach to Spatial Cognition: Representing Spatial Relationships in Tongan Language and Culture*. Doctoral dissertation, Department of Anthropology, University of Illinois at Urbana-Champaign, Urbana, Illinois.

Bennardo, G. (2000). Language and Space in Tonga: 'The Front of the House is Where the Chief Sits!' *Anthropological Linguistics*, 42,4, 499-544.

Bowerman, Melissa, & Levinson S. C. (2001). *Language Acquisition and Conceptual Development*. Cambridge: Cambridge University Press.

Cohen, R. (Ed.) (1985). *The Development of Spatial Cognition*. London: Lawrence Erlbaum Associates, Publishers.

Herskovits, A. (1986). *Language and Spatial Cognition : An Interdisciplinary Study of the Prepositions in English* . Cambridge: Cambridge University Press.

Hill, C. (1982). Up/Down, Front/Back, Left/Right. A Contrastive Study of Hausa and English. In J. Weissenborn & W. Klein (Eds.), *Here and There: Cross-Linguistic Studies on Deixis and Demonstration*. Amsterdam: John Benjamins.

Hirschfeld, L. A., & Gelman S. A. (Eds.) (1994). *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Lehman, F. K. & Bennardo G. (2003). A Computational approach to the cognition of space and its linguistic expressions. *Mathematical Anthropology and Cultural Theory*, 1, 2, 1-83. See: www.mathematicalanthropology.org/

Levinson, S. C. (2003). *Space in language and culture: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Piaget, J., & Inhelder B. (1956). *The Child's Conception of Space*. New York: The Humanities Press Inc.

Pick, H. L. (1993). Organization of Spatial Knowledge in Children. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), *Spatial representation: Problems in Philosophy and Psychology*. Oxford: Basil Blackwell Ltd.

Talmy, L. (2000). *Toward a Cognitive Semantics, Volume 1: Concept Structuring Systems*. Cambridge, MA: The MIT Press.

Simulated Action in an Embodied Construction Grammar

Benjamin Bergen (bergen@hawaii.edu)
Dept of Linguistics, 1890 East-West Hall, 569 Moore Hall
Honolulu, HI 96822

Nancy Chang (nchang@icsi.berkeley.edu)
Shweta Narayan (shweta@icsi.berkeley.edu)
International Computer Science Institute, 1947 Center St., Suite 600
Berkeley, CA 94704-1198, USA

Abstract

Various lines of research on language have converged on the premise that linguistic knowledge has as its basic unit pairings of form and meaning. The precise nature of the meanings involved, however, remains subject to the longstanding debate between proponents of arbitrary, abstract representations and those who argue for more detailed perceptuo-motor representations. We propose a model, Embodied Construction Grammar (ECG), which integrates these two positions by casting meanings as schematic representations embodied in human perceptual and motor systems. On this view, understanding everyday language entails running mental simulations of its perceptual and motor content. Linguistic meanings are parameterizations of aspects of such simulations; they thus serve as an interface between the relatively discrete properties of language and the detailed and encyclopedic knowledge needed for simulation. This paper assembles evidence from neural imaging and psycholinguistic experiments supporting this general approach to language understanding. It also introduces ECG as a model that fulfills the requisite constraints, and presents two kinds of support for the model. First, we describe two verbal matching studies that test predictions the model makes about the degree of motor detail available in lexical representations. Second, we demonstrate the viability and utility of ECG as a grammar formalism through its capacity to support computational models of language understanding and acquisition.

Introduction

Many theories of language take the basic unit of linguistic knowledge to be pairings of form and meaning, known as *symbols* or *constructions* (de Saussure 1916; Pollard & Sag 1994; Goldberg 1995; Langacker 1987). This view stems from the simple observation that language serves to convey meaning, using form. A speaker must thus know what linguistic forms are appropriate to encode the meanings s/he wishes to convey, and vice versa for an understander.

The nature of the meaning representations of linguistic units, however, remains very much at issue. Suggestions in the literature range from relatively abstract representations, including both feature structures (Pollard & Sag 1994) and logical representations (May 1985), to more concrete perceptual- or motor-based representations (Langacker 1987; Barsalou 1999; Glenberg & Robertson 2000).

Each of these approaches faces difficulties. Abstract symbol systems, whether feature-based or logical, invite the question of how (or even whether) they are ultimately linked to human perceptual, motor, affective, and other sorts of experience. There is strong evidence, seen below, that such embodied knowledge is automatically and unconsciously brought to bear during language understanding. Moreover, language users naturally make a broad range of associative and causal inferences based on language, a process not easily represented in an abstract symbol system.

Conversely, a theory of linguistic meaning cannot be based on perceptuo-motor information alone. Linguistic units can be combined in ways that are not strictly predictable from their semantic properties. Our ability to judge the grammaticality of sentences like Chomsky's (1957) classic *Colorless green ideas sleep furiously* example provides strong evidence of linguistic structure distinct from motor, perceptual, or other world knowledge. Additionally, our ability to understand sentences like *My pet chicken kissed me on the cheek* (even though chickens don't have lips, presumably a prerequisite for kissing) shows that grounded motor knowledge does not suffice to account for our ability to extract meaning from language.

One concrete solution to the drawbacks of purely abstract and purely perceptuo-motor approaches is to characterize mental representations as schematizations over modal knowledge (Fillmore 1982; Langacker 1987; Lakoff 1987; Barsalou 1999; Talmy 2000). This compromise view retains the best of both worlds: while language *use* involves the activation of perceptual and motor mechanisms, linguistic *units* themselves need only refer to schematic representations of these mechanisms. Proposals along these lines have inspired work investigating how the perceptual and motor structures underlying word meaning might be represented and schematized in computational models of human language processing (Regier 1996; Bailey 1997; Narayanan 1997). But the nature of the lexical and grammatical units that link these structures with linguistic forms has not yet been articulated precisely enough to support formal or computational implementation.

This paper synthesizes diverse evidence for an integrated view of language use and presents Embodied Construction Grammar (ECG), a formally specified instantiation of the approach. We begin by surveying evidence of the importance of perceptual and motor simulation in higher-level cognition, especially in language use. We then briefly outline the ECG formalism and show how it supports a model of human language use in which linguistic meanings serve to parameterize motor and perceptual structures. The remainder of the paper presents two kinds of support for the model. First, we describe a pair of verb matching studies

that test predictions the model make about the degree of simulative detail in lexical representations. Second, we demonstrate the viability and utility of ECG as a grammar formalism precise enough to support computational models of language understanding and acquisition.

Mental Simulation in Language Use

Evidence for simulation

Perceptual and motor systems play an important role in higher cognitive functions, like memory and categorization (Barsalou 1999; Glenberg & Robertson 2000; Wheeler, Petersen & Buckner 2000; Nyberg et al. 2001), as well as motor (Lotze et al. 1999) and perceptual (Kosslyn, Ganis & Thompson 2001) imagery. It would thus be surprising if there were no role for perceptual and motor systems in language use as well.

Some theorists have proposed that perceptual and motor systems perform a central function in language production and comprehension (Glenberg & Robertson 2000; Barsalou 1999). In particular, they have suggested that understanding a piece of language entails internally simulating, or mentally imagining, the described scenario, by activating a subset of the neural structures that would be involved in perceiving the percepts or performing the motor actions described.

Several recent studies support this notion of simulation in language understanding, based on the activation of motor and pre-motor cortex areas associated with specific body parts in response to motor language referring to those body parts. Using behavioral and neurophysiological methods, Pulvermüller, Haerle & Hummel (2001) and Hauk Johnsrude & Pulvermüller (2004) found that verbs associated with different effectors were processed at different rates and in different regions of motor cortex. For example (Pulvermüller et al. 2001), when subjects perform a lexical decision task with verbs referring to actions involving the mouth (e.g. *chew*), leg (e.g. *kick*), or hand (e.g. *grab*), areas of motor cortex responsible for mouth/leg/hand motion displayed more activation, respectively. Tettamanti et al. (ms.) have also shown through an imaging study that passive listening to sentences describing mouth/leg/hand motions activates different parts of pre-motor cortex (as well as other areas, specifically BA 6, 40, and 44).

Behavioral methodologies also offer convergent evidence for the automatic and unconscious use of perceptual and motor systems during language use. Recent work on spatial language (Richardson et al. 2003; Bergen To Appear) has found that sentences with visual semantic components can result in selective interference with visual processing. While processing sentences that encode upwards motion, like *The ant climbed*, subjects take longer to perform a visual categorization task in the upper part of their visual field; the same is true of downwards-motion sentences like *The ant fell* and the lower half of the visual field. These results imply that language, like memory, evokes visual imagery that interferes with visual perception.

A second experimental method (Glenberg & Kashak 2002), tests the extent to which motor representations are activated for language understanding. The findings from this approach have shown that when subjects are asked to perform a physical action in response to a sentence, such as moving their hand away from or toward their body, it takes them longer to perform the action if it is incompatible with the motor actions described in the sentence. This suggests that while processing language, we perform motor imagery, using neural structures dedicated to motor control.

A third method, used by Stanfield & Zwaan (2001) and Zwaan et al. (2002), investigates the nature of visual object representations during language understanding. These studies have shown that implied orientation of objects in sentences (like *The man hammered the nail into the floor* versus *The man hammered the nail into the wall)* affected how long it took subjects to decide whether an image of an object (in this case, a nail) had been mentioned in the sentence, or even to name that object. It took subjects longer to respond to an image that was incompatible with the implied orientation or shape of a mentioned object. These results imply that not just trajectory and manner of motion, but also shape and orientation of objects, are represented in mental simulations during language understanding.

Linguistic knowledge as a simulation interface

Language understanding seems to entail the activation of perceptual and motor systems, which work in a dynamic, continuous, context-dependent, and open-ended fashion. Linguistic form, by contrast, is predominantly discrete — a word either precedes another word or does not; a morpheme is either pronounced or not, and so on. How do linguistic representations pair relatively discrete linguistic forms with continuous, dynamic, modal perceptuo-motor simulations? The notion of *parameterization* offers an answer.

Grammatical knowledge governing the productive combination of linguistic units appears to draw primarily on schematic properties of entities and events (Langacker 1987; Goldberg 1995), such as whether an entity can exert force or move, or whether an action involves the exertion of force or causes motion. Thus, for the purposes of language understanding, which involves determining what linguistic units an utterance uses and how they are combined, it may be sufficient for words and morphemes to generalize over perceptual or motor detail and encode only the important, distinctive aspects of actions and percepts required to perform a simulation. These parameterized representations are not abstract, amodal symbols, since they are directly grounded in action and perception, but they are distinct from the simulative content they parameterize.

This simulation-based view of language understanding has immediate consequences for theories of language. Only meaning representations that can be usefully fed to the simulation are viable; at the same time, constructions are freed from having to capture the encyclopedic knowledge handled by simulation. This division of labor between the meaning representations of linguistic constructions and the detailed structures that support simulation provides the means for finite, discrete linguistic structures to evoke the open-ended, continuous realm of possible meanings that language users may communicate. ECG is a theory of language that conforms to these constraints.

Embodied Construction Grammar

Embodied Construction Grammar (Bergen & Chang To Appear; Chang et al. 2002) aims to be a theory of language suitable for integration in a grounded, computationally implemented, simulation-based theory of human language use. It resembles other Construction Grammars (Kay & Fillmore 1999; Goldberg 1995; Croft 2001) in counting form-meaning pairings as the basic linguistic unit, and in aiming for full coverage of linguistic behavior. But ECG also serves as precisely the interface between language and simulation described above. It thus differs from other grammatical theories in emphasizing the embodiment of the grammatical system: constructions pair schematic form representations with schematic meaning representations, which are further constrained to be abstractions over perceptual and motor representations that can be simulated, or over characteristics of simulations in general.

A detailed description of the formalism is given in Bergen & Chang (To Appear); ECG has also been applied to a wide range of linguistic phenomena, including argument structure, reference, predication, and morphology, in a variety of languages. We concentrate here on showing how the representational tools of ECG satisfy and exploit the constraints of a simulation-based approach to language understanding. We first describe the high-level interactions posited in the model between linguistic constructions and the dynamic processes of language understanding they support, and then illustrate these with a simple example.

The Language Understanding Process

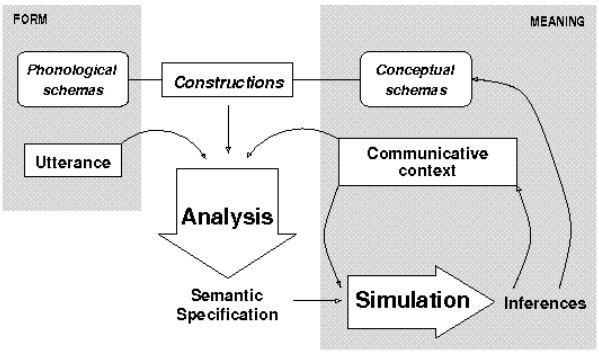


Figure 1. Language understanding in ECG.

The main source of linguistic knowledge in ECG is a large repository of constructions that express generalizations linking the domains of form (typically, phonological schemas and relations) and meaning (conceptual schemas and relations). Some constructions directly specify which perceptual and motor mechanisms to deploy, while others (especially larger phrasal and clausal constructions) specify how to combine the parameterized representations corresponding to different kinds of imagery. Still other constructions may affect the mode of simulation itself; the passive construction, for example, modulates what perspective is to be taken in the simulation of a given action.

There are also two main dynamic processes (large arrows in Figure 1) that interact with constructional knowledge during language comprehension. The first is the *analysis* process, which takes an input utterance in context and determines the set or sets of constructions most likely to be responsible for it. This process is thus roughly analogous to parsing, though it additionally incorporates contextual information, following Tanenhaus et al. (1995) and Spivey et al. (2001). The product of the analysis process is a structure called the *semantic specification* (or *semspec*), which specifies the conceptual schemas evoked by the constructions and how they are related. The second process is *simulation*, which takes the semspec as input and exploits representations underlying action and perception to simulate the specified events, actions, objects, relations, and states. The resultant inferences shape subsequent processing and provide the basis for the language user's response.

Embodied Construction Grammar in action

This section shows how the process just described would produce the appropriate simulation and resulting inferences for the sentence *Mary bit John*. The understander first tries to recognize the sequence of sounds in terms of form schemas. In speech or in sentences with novel or ambiguous word forms, this may require sophisticated categorization. Here, the forms are straightforwardly recognized as three form schemas ('Mary', 'bit', and 'John') with the appropriate temporal ordering relations among them, shown as vertical arrows on the left-hand side of Figure 2.

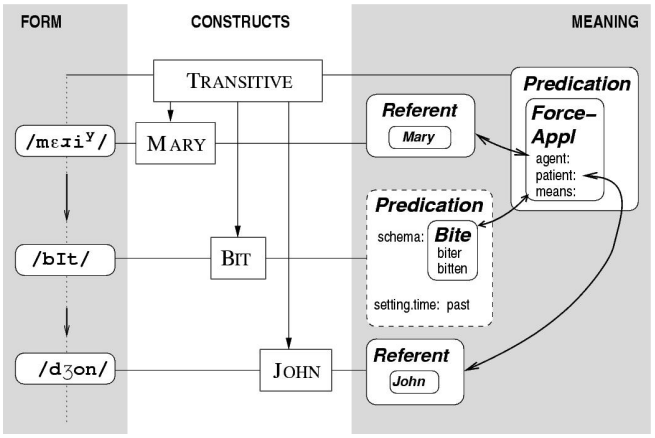


Figure 2: The (simplified) analysis of Mary bit John.

Next, the analysis process hypothesizes which constructions (instantiated as *constructs*) could account for the utterance; these are constructions whose form elements are present in the utterance. The four constructions relevant for this utterance are shown in the middle column of Figure 2. The JOHN and MARY constructions each bind some phonological form (on the left) with a particular schema for its referent (on the right). The BIT construction binds its phonological form with a predication that features a schema called Bite, which is the schematization of perceptual and motor knowledge about biting. Additionally, the predication is specified as being about the past, which will become relevant when inferences are propagated, after simulation.

The clausal TRANSITIVE construction binds together the forms and the meanings of the three lexical constructions, which serve as its constituents. On the form side, it specifies an ordering relation (Agent precedes Verb precedes Patient),

while on the meaning side, the clause is linked to a predication that encodes the application of force by some means, where that means is bound to the Bite schema, the agent is Mary, and the patient is John. The clausal construction thus contributes its own structure and schematic meanings, and effects bindings among these and those of its constituents. In our example, the analysis process succeeds in finding a set of constructions that match the utterance and whose different constraints fit together, or unify, in all three domains: form, meaning and construction.

The completed analysis process produces a semspec (consisting of the meaning schemas and bindings, shown in the right-hand column in Figure 2), which is used as input for the next step, the mental simulation of the described scene. The semspec indicates which perceptual and motor structures should be activated and how they are related. It might also specify other parameters of the simulation, such as the perspective from which to simulate. Our example is in the active voice, not passive (e.g. *John was bitten by Mary*), so would by default be simulated from Mary's perspective, resulting in the activation of a motor schema for biting (though features of the surrounding context could result in an "experiencer" simulation perspective instead).

Although our example omits many details of analysis and simulation (including how the model supports, e.g., reference resolution, construal, and sense disambiguation (for a discussion of these, see Bergen and Chang (To Appear)), it nonetheless demonstrates how ECG captures the idea of parameterization or schematization. A verb form like 'bit' centrally includes a Bite schema in its meaning; this schema is a parameterization over perceptual and motor knowledge about biting. To figure out who is biting whom — that is, to understand how the meaning of 'bit' relates to the meanings of the other words in the utterance — only very general knowledge about biting (that it is an action in which force is exerted by one participant on another) is required. The simulation process makes use of the perceptual and motor knowledge underlying this schematic representation, and provides detailed perceptual and motor content that can support inference and, on the current account, constitutes understanding.

The remainder of the paper offers support for the ECG model from two different sources: a pair of behavioral experiments testing a prediction of the model, and the implementation of the formalism within computational models of language acquisition and understanding.

Experiment: Embodied Verbal Representation

The ECG approach to language claims that verbal semantics involves the activation of detailed motor knowledge about performing or perceiving the relevant action. One reflection of this prediction might be in the representation of the effector used to perform the action described by a verb. To test this hypothesis, we performed two related behavioral experiments. In the first, subjects were shown a picture and then a verb, and asked to decide whether the word correctly described the picture. In the second, subjects were asked to decide whether two verbs had nearly the same meaning. We predicted that subjects would take longer to reject as matches an image-verb or verb-verb pair that depicted different actions using the same effector, compared to the case when the two non-matches used different effectors.

Method

Study 1. 39 native English speakers participated for course credit. They were told that they would see a picture of a person performing an action on a screen, followed by a verb, and were instructed to decide as quickly as possible whether the verb was a good description of the picture.

During each trial, subjects were presented a stick figure of a person carrying out an action for 1 sec, a visual mask for 450 msec, and a blank screen for 50 msec. Then the verb was displayed, and stayed on the screen until the subject pressed "yes" or "no". All verbs were presented in the center of the screen. All actions were predominantly performed using one of three effectors: foot, hand or mouth. More detailed discussion of the methodology can be found in Bergen, Feldman & Narayan (2003).

Study 2. 53 native English speakers participated on a volunteer basis or for course credit. They were told that they would see a word appear on a screen and were instructed to decide as quickly as possible whether a second word that appeared meant more or less the same as the first word.

During each trial, subjects were presented with a fixation cross in the center of the screen for 2 sec, followed by an English action verb for 1 sec, a visual mask for 450 msec, and a blank screen for 50 msec. Then the second verb was displayed, and stayed on the screen until the subject pressed "yes" or "no". All verbs were capitalized and presented in the center of the screen. Verb pairs in critical trials pertained to motor actions of the following categories:

Matching: Near-synonyms, e.g.
SCREAM and SHRIEK; DANCE and WALTZ

Non-matching, same effector: Verbs with clearly different meanings, using the same effector, e.g.
SCREAM and LICK; DANCE and LIMP

Non-matching, different effector: Verbs with clearly different meanings, using different effectors, e.g.
SCREAM and STEP; DANCE and YELL

More detailed discussion of the methodology can be found in Narayan, Bergen & Weinberg (2004).

Results

Study 1. Counting only replies that were correct and within 2s.d. of the mean for a given subject, mean reaction times were 751ms for different-effector mismatches, 799ms for same-effector mismatches, and 741ms for matches. Using a standard ANOVA, the difference between the mismatching conditions was found to be significant ($p<.0001$).

Study 2. Counting only replies that were correct and within 3s.d. of the mean for a given subject, mean reaction times were 930ms for different-effector mismatches, 1030ms for same-effector mismatches, and 1070ms for near-synonyms.

Following Clark (1973), we performed two ANOVAs, with subjects and items as nested random factors, and from these determined that the RT difference between the

mismatch conditions is significant (minF'(1,126)=9.0808, $p<0.005$). Post hoc tests showed that the non-matching different-effector condition is significantly different from the matching condition (minF'(1,126)=9.781), and the non-matching same-effector condition is not significantly different (minF'(1,126)=2.0002).

Discussion. Subjects took significantly longer to reject either a picture-verb pair as matches or a verb pair as near-synonyms when the two actions shared an effector than when they did not. Since this effect occurred when part of the task was non-linguistic (Study 1), this is unlikely to be a mere lexical effect. Moreover, the presence of the effect with purely linguistic stimuli (Study 2) means it is not due to strictly visual properties of the stimuli, either. Instead, these results suggest that understanding motion verbs goes beyond accessing abstract structures; modal information about bodily action, such as the effector used, is involved.

Importantly, the results imply that verb meaning does involve evoking modal motor representations: words encoding particular motor actions (kick, chew) contribute to the perceptuo-motor content of mental simulations.

ECG computational implementation

ECG is compatible in its broad outlines with a large body of linguistic and psycholinguistic research. But it is subject to the important additional constraint of being computationally precise. As we have described it, understanding even the simplest utterance involves multiple dynamic processes interacting with a variety of linguistic and embodied representations. Many of these are inspired by ideas from cognitive linguistics that have not been previously formalized, let alone used in any implemented system. It is thus crucial that we validate the framework by offering concrete implementations. In this section we briefly describe how the formalism serves as the lynchpin for computational models of linguistic use and acquisition.

Formally, the ECG construction and schema formalisms have much in common with other unification-based grammars (e.g., Pollard & Sag 1994), including notations for expressing features, inheritance, typing, and unification/coindexation; it also has additional mechanisms that increase its expressivity and flexibility.

As described earlier, the ECG formalism is designed to play a role in language understanding as the key interface between constructional analysis and the embodied simulation. Bryant (2003) describes an implemented construction analyzer that takes as input a grammar of ECG constructions and a sentence, and produces a semspec that provides the parameters for a simulation. The analyzer extends methods from syntactic parsing (particularly partial parsing and unification-based chart parsing) to accommodate and exploit the dual form-meaning nature of constructions. Specifically, it consists of a set of *construction recognizers*; each recognizer seeks the particular input form (or constraints) of its corresponding construction, and upon finding it checks the relevant semantic constraints. If multiple analyses are possible, the analyzer uses a *semantic density* metric to choose the analysis whose semspec is the most semantically coherent and complete. Thus, in constrast with typical language understanding systems in which syntactic parsing precedes semantic interpretation, the construction-based analyzer incorporates semantic constraints in parallel, reflecting the central role played by meaning in the ECG formalism.

The semspec produced by the analyzer provides parameters for simulation using active, modal structures. A broad range of embodied meanings have been modeled using *executing schemas* (x-schemas), a dynamic representation motivated in part by motor and perceptual systems (Narayanan 1997; Bailey 1997). X-schemas can model sequential, concurrent, and asynchronous events. The Bite schema, for example, parameterizes a Bite x-schema that captures the continuous mouth actions culminating in a particular forceful application of the teeth of the Biter to the Bitten. A simulation engine based on x-schemas has been implemented (Narayanan 1997) and applied to model the semantics of several domains, including verbal (Bailey 1997) and aspectual semantics (Chang, Gildea & Narayanan 1998), metaphorical inference (Narayanan 1999), and frame-based perspectival inference (Chang et al. 2002).

Although we have focused so far on language understanding, the ECG formalism is also designed to support a computational model of the acquisition of early phrasal and clausal constructions (Chang & Maia 2001; Chang 2004). This model takes ECG as the target representation to be learned from a sequence of utterances in context. Learning is usage-based in that utterances are first analyzed using the process described above; the resulting (partial) semspec is used along with context to prompt the formation of new constructions. The model has been applied to learn simple English motion constructions from a corpus of child-directed utterances, paired with situation representations. The resulting learning trends reflect cross-linguistic acquisition patterns, in particular the learning of lexically specific verb island constructions before more abstract grammatical patterns (Tomasello 1992). They also demonstrate how the ECG formalism serves as an interface between language comprehension and acquisition.

The implementations described here do not provide direct evidence of the cognitive reality of ECG. But they do demonstrate its utility and flexibility, and, by offering an integrated and precisely specified instantiation of simulation-based language understanding and use, serve as an existence proof for the overall approach.

Conclusions

If the embodied view presented above is correct, then the human capacity for language understanding relies on activating internal motor and perceptual simulations on the basis of linguistic input. These simulations can serve any of the purposes that linguistic information is conventionally put to — their content can be stored, thereby updating the internal knowledge base; their inferences can be propagated such that the understander can draw conclusions needed in discourse; or the actions they include can be performed in cases where the language involves instructions or requests.

The computationally viable and empirically supported model described above views linguistic units as pairings between schematic representations of form and schematic

representations of meaning. Those representations are not abstract and arbitrary; rather, they are tightly bound to the perceptual and motor substrates over which they generalize. This approach permits insights into how language is integrated with perceptual and motor knowledge in the cognitive system, and sheds light on what meaning means.

Acknowledgements

Our thanks to Jerome Feldman, George Lakoff, Srini Narayanan, Zachary Weinberg, and other members of the NTL research group, as well as two anonymous reviewers for their helpful comments.

References

Bailey, D. 1997. When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs. PhD thesis, UC Berkeley.

Barsalou, L. W. 1999. Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-609.

Bergen, B. & Chang, N. To Appear. Embodied Construction Grammar in Simulation-Based Language Understanding. In J-O. Östman and M. Fried (eds.) Construction Grammar(s): Cognitive and Cross-language dimensions. John Benjamins. http://www.icsi.berkeley.edu/NTL/papers/ecg-tr-02-004.pdf

Bergen, B. To appear. Experimental methods for simulation semantics. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, and M. Spivey (eds.) Methods in Cognitive Linguistics: Ithaca.

Bergen, B., Narayan S., & Feldman, J. 2003. Embodied verbal semantics: evidence from an image-verb matching task. Proceedings Cognitive Science 25. Mahwah, NJ: Erlbaum.

Bryant, J. 2003. Constructional Analysis. UC Berkeley M.S. Report.

Chang, N, Gildea, D. & Narayanan, S. 1998. A Dynamic Model of Aspectual Composition. Proceedings Cognitive Science 20.

Chang, N. & T. V. Maia. 2001. Learning grammatical constructions. Proceedings Cog. Sci. 23. Mahwah: Erlbaum.

Chang, N., Feldman, J., Porzel, R., & Sanders, K. 2002. Scaling Cognitive Linguistics: Formalisms for Language Understanding. 2002. Proceedings of SCANALU.

Chang, N. 2004. Constructing Grammar: A computational model of the acquisition of early constructions. Ph.D. thesis, UC Berkeley.

Chomsky, N. 1957. Syntactic Structures. Mouton.

Clark, H. 1973. The language-as-fixed-effect fallacy. Journal of Verbal Learning and Verbal Behavior, 12:335–359.

Croft, W. 2001. Radical Construction Grammar. Oxford: Oxford University Press.

Fillmore, C. J. 1982. Frame semantics. In Linguistic Society of Korea (eds.), Linguistics in the Morning Calm.

Glenberg, A. M. & Kaschak, M. P. 2002. Grounding language in action. Psychonomic Bulletin & Review.

Glenberg, A. M. & Robertson, D. A. 2000. Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. JML, 43, 379-401.

Goldberg, A. 1995. Constructions: A Construction Grammar Approach to Argument Structure. Chicago: U Chicago Press.

Hauk, O., Johnsrude, I. & Pulvermüller, F. 2004. . Somatotopic representation of action words in human motor and premotor cortex. Neuron. 41(2): 301-7.

Kay, P. & Fillmore, C. J. 1999. Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. Language 75/1: 1-33.

Kosslyn, S. M., Ganis, G. & Thompson, W. L. 2001. Neural foundations of imagery. Nature Reviews Neurosci., 2:635-642.

Lakoff, G. 1987. Women, fire, and dangerous things. Chicago: U Chicago Press.

Langacker, R. 1987. Foundations of Cognitive Grammar: Theoretical Perspectives 1. Stanford: Stanford University Press.

Lotze, M., Montoya, P., Erb, M., Hülsmann, E., Flor, H., Klose, U., Birbaumer, N., & Grodd., W. 1999. Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fMRI study. Jrn. Cog. Neurosci 11(5): 491-501

May, R. 1985. Logical Form. Cambridge: MIT Press

Narayan, S., Bergen, B., & Weinberg, Z. 2004. Embodied Verbal Semantics: Evidence from a lexical matching task. Proceedings of Berkeley Linguistics Society 30. Berkeley.

Narayanan, S. 1997 KARMA: Knowledge-based Action Representations for Metaphor and Aspect. Ph.D. thesis, UC Berkeley

Narayanan, S. 1999. Moving Right Along: A Computational Model of Metaphoric Reasoning about Events. Proceedings of the Nat. Conf. on Artificial Intelligence (AAAI '99): 121-128.

Nyberg, L., Petersson, K.-M., Nilsson, L.-G., Sandblom, J., Åberg, C., & Ingvar, M. 2001. Reactivation of motor brain areas during explicit memory for actions. NeuroImage, 14, 521-528.

Pollard, C. & Sag, I. 1994. Head-Driven Phrase-Structure Grammar. Chicago: University of Chicago Press.

Pulvermüller, F., Haerle, M., & Hummel, F. 2001. Walking or Talking?: Behavioral and Neurophysiological Correlates of Action Verb Processing. Brain and Language 78, 143–168.

Regier, T. 1996. The Human Semantic Potential: Spatial Language and Constrained Connectionism. Cambridge: MIT Press.

Richardson, D. C., Spivey, M. J., McRae, K., & Barsalou, L. W. 2003. Spatial representations activated during real-time comprehension of verbs. Cognitive Science.

de Saussure, F. 1916. Course de linguistique générale. Paris: Payot.

Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M. & Sedivy, J.C. 2001. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. Cognitive Psychology.

Stanfield, R. A. & Zwaan, R. A. 2001. The effect of implied orientation derived from verbal context on picture recognition. Psychological Science, 12, 153-156.

Talmy, L. 2000. Toward a Cognitive Semantics. Cambridge: MIT.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. 1995. Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Perani, D., Cappa, S. F., Fazio, F., & Rizzolatti, G. Unpublished Ms. Sentences describing actions activate visuomotor execution and observation systems.

Tomasello, M. 1992. First verbs: A case study of early grammatical development. Cambridge: Cambridge University.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory specific cortex. Proc. Natl. Acad. Sci. USA 97: 11125–11129.

Zwaan, R. A., Stanfield, R.A., & Yaxley, R. H. 2002. Do language comprehenders routinely represent the shapes of objects? Psychological Science, 13, 168-171.

Hedged Responses and Expressions of Affect in Human/Human and Human/Computer Tutorial Interactions

Khelan Bhatt (bhatkhe@iit.edu)
Martha Evens (evens@iit.edu)
Shlomo Argamon (argamon@iit.edu)
Computer Science Department, Illinois Institute of Technology
10 West 31st Street, Chicago, IL 60616

Abstract

We study how students hedge and express affect when interacting with both humans and computer systems, during keyboard-mediated natural language tutoring sessions in medicine. We found significant differences in such student behavior linked to whether the tutor was human or a computer. Students hedge and apologize often to human tutors, but very rarely to computer tutors. The type of expressions also differed—overt hostility was not encountered in human tutoring sessions, but was a major component in computer-tutored sessions. Little gender-linking of hedging behavior was found, contrary to expectations based on prior studies. A weak gender-linked effect was found for affect in human tutored sessions.

Introduction

How people interact with computers is of clear importance to the design of effective computer interfaces. The book *The Media Equation* (Reeves & Nass 1996) claims that people treat computer systems essentially the same as they treat people, though more recent work (Shechtman & Horowitz 2003; Goldstein et al., 2002) has raised serious questions about this conclusion. Differences between how people respond to human beings and how they respond to computers have been informally documented since the first experiments with natural language interfaces (Thompson, 1980). A better elucidation of the issues may improve intelligent systems design.

Specifically, understanding these issues better may aid in the development of more effective tutoring systems. In this paper, we study the differences between student reactions to our Intelligent Tutoring System (ITS), CIRCSIM-Tutor (Michael et al., 2003), and the human tutors on which it was modeled. Our goal is to characterize student hedges and expressions of affect and try to determine how our ITS could understand them and respond effectively.

We are motivated by experiments (Fox 1993) that suggest such differences for ITSs that carry out a natural language dialogue with the student. Fox carried out a "Wizard-of-Oz" experiment which showed students to be polite and friendly to human tutors when they met with them face-to-face, but decidedly rude to the same tutors when communicating with them over a slow computer link and told that a machine was tutoring them.

The current study has potentially important implications for the future development of our ITS. Investigation of how human tutors respond to student misery, frustration, and rage is the first step toward making systems more friendly and responsive. By contrast, our system's current response to student hedges and expressions of affect (as to any input it does not understand) is to tell the student what kind of input it is expecting. The result is dialogue like this:

Student: Clueless!
Tutor: Please respond with prediction table parameters.

Better understanding of how and when students express affect in tutoring sessions and the functions of such expressions in the discourse may lead to improvements in student modeling and hence tutoring effectiveness.

Background

Thompson's (1980) system was a pioneering natural-language interface designed to help U.S. Navy personnel load cargo onto ships. It thus attempted to delete all affective remarks, to avoid confusing the parser. Although the system was quite effective at its task, most of its affective input consisted of curses. By contrast, chat-oriented natural language interaction programs like ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975), can impress their users with simulated charm and intelligence, despite a lack of any deep understanding. Similarly, physicians experienced the natural language interface of Shortliffe's (1982) MYCIN and ONCOCIN programs as attractive, even though

input was restricted to one-word answers to questions.

The specific question of how to properly interpret student hedging in tutoring sessions was raised at the *NAACL Workshop on Adaptation in Dialogue Systems,* held as part of the 2001 meeting of the Association for Computational Linguistics. It was suggested that student hedges might provide useful information by reliably signaling student misconceptions. Our collaborators on the CIRCSIM-Tutor project at Rush Medical College are dubious about this suggestion, however. Ten years ago, after their first experiments with tutoring in cardiovascular physiology they resolved to stop commenting on hedges, because they felt that student hedging reflects personal communication styles more than any real confusion. Further experience has not changed their minds, although they respond with help and encouragement whenever they believe the student to be experiencing real distress (Bhatt 2004).

As well, there is an increasing recognition in the ITS community of the importance of affect. A full session at *Intelligent Tutoring Systems* 2002 was devoted to such issues (Aist et al. 2002; Kort & Reilly 2002; Vicente & Pain 2002). These papers all argue for the importance of responding to evidence of student distress. Our study is the first, to our knowledge, that explicitly studies student hedging of answers and expressions of affect by comparing human and computer tutorial sessions. Relevant in this context also is the recent general trend towards greater concern in the AI community with emotional aspects of intelligence, sparked mainly by the work of Breazeal and Brooks (Brooks et al. 1998; Breazeal 1998).

Goals and Hypotheses

We study response hedging and expressions of affect in human and machine tutoring sessions. This study incorporates both exploratory and hypothesis testing goals. The main exploratory questions that we investigated are as follows:

What kinds of hedged responses and expressions of affect do we see in human tutoring sessions?

What kinds of hedged responses and expressions of affect do we see in machine tutoring sessions?

How might the two kinds of tutoring interactions differ regarding student use of hedged responses and expressions of affect?

In addition, based on results in human/computer interaction (primarily Fox (1993) and Thompson (1980)), we formulate our main hypotheses:

H1a (Hedging Differs): *Student use of hedging differs depending on whether the tutor is a human or a computer system.*

H1b (Affect Differs): *Student use of affect differs depending on whether the tutor is a human or a computer system.*

The workshop discussion mentioned above also prompted us to investigate two subsidiary hypotheses about hedging, and how it may prove useful for student modeling:

H2a (Hedges Inform): *The presence of a hedge provides information regarding whether a student answer is right or wrong.*

H2b (Hedges Wrong)*: Hedged answers are almost always wrong and so provide near certain feedback for student modeling.*

Regarding the relevance of *H2b*, note that most computer tutoring systems cannot currently make use of 'weak' probabilistic information for student modeling, such as "hedged answers are 20% more likely to be wrong than non-hedged answers", but only more certain statements, such as "hedged answers are almost always wrong".

Gender-linked variation

Many previous studies, including Lakoff (1975) and Aries (1989), have reported that women hedge more than men, although interpretation of such claims is complex (Holmes 1984), since hedging can be a politeness or face-saving strategy, and not necessarily an expression of uncertainty. Of particular relevance are recent results on hedging in tutoring systems (Shah et al., 2002), which found that women hedge significantly more often than men when making initiatives in tutoring dialogues. If such differences are consistent, it should influence how tutoring systems interact with male and female students. We thus formulate:

H3a (Women Hedge): *Women hedge answers more often than men in tutoring interactions.*

Aries, Lakoff (1990) and Tannen (1990) all describe women as more likely to express emotion than men. Hence:

H3b (Women are Affectual): *Women use more affective expressions than men in tutoring interactions.*

Furthermore, Lakoff (1975) also describes women as apologizing more often. Thus we also consider whether:

H3c (Women Apologize): *Women apologize more often than men in tutoring interactions.*

Data Collection

Human/Human Tutoring Sessions

We collected transcripts of keyboard-to-keyboard human tutoring sessions (henceforth, *H/H sessions*) between students and their expert tutors on the subject of the baroreceptor reflex during November 1999. Sessions took place with the student and the tutor in separate rooms, communicating only via keyboard. The tutor for each session was either Joel Michael or Allen Rovick (both professors of physiology, the same tutor throughout each session), and the 25 subjects were paid volunteers, first year students at Rush Medical College enrolled in a physiology course. The data examined consists of over 51,000 words (over 12,000 lines) of student-tutor dialogue, from hour-long sessions (numbered K52-K76 in our corpus).

Human/Computer Tutoring Sessions

In November 2002, most of the first year class at Rush Medical College used CIRCSIM-Tutor (Michael et al., 2003) for one hour in a regularly scheduled laboratory session. Some students worked in pairs, some alone, so we wound up with only 66 transcripts (the *H/C sessions*), which we used as the basis for our findings about machine tutoring sessions. The system presents the same problems about the baroreceptor reflex as the human tutors and attempts to emulate their tutoring strategies. We have not yet attempted to analyze the differences between the single-user and paired sessions.

Methodology

Coding of Hedges

Hedges in the transcripts were hand-coded using a coding scheme based on the hedge types described in Shah's (2002) study of hedged initiatives. The first step was to examine transcripts of four H/H sessions (K52-K55) and to establish an initial categorization. This phase was performed collectively by Bhatt and Evens. Subsequently, the remaining twenty-two sessions were coded by each researcher independently. Each hedged instance was classed by one of the predefined types (Table 1). Inter-rater reliability was excellent, with a kappa of 0.97.

Following this initial coding and coder comparison, some hedge types were eliminated or aggregated into other types, and coding was standardized in all transcripts. Transcripts were electronically marked up using SGML tags, to facitlitate subsequent counting of hedges and hedge types for statistical analysis. The final list of hedge types, along with counts and examples of usage, is given in Table 1.

Table 1: Final list of hedge categories with definitions or examples of usage, with counts of occurrences as answers (**A**) and initiatives (**I**).

Hedge Type	A	I	Example
BELIEVE	6	0	*I believe*
EITHER_OR	2	0	*Either X orY*
EQUIVALENT	3	1	*it sounds as though*
EXPECT	12	0	*probably*
GUESS	10	1	*I guess*
KIND_OF	7	0	*Kind of*
MAYBE	4	4	*Maybe*
NOT_SURE	9	3	*I'm not sure*
Q1	61	11	Question mark after a statement
Q2	2	1	Question syntax with no "**?**"
SHOULD	1	0	*X should increase*
TAG	2	2	*It shouldn't X, should it?*
THINK	44	11	*I think*
THOUGHT	21	4	*I thought*
TRY	3	0	*I can try*

Table 2: Types of affect expressions in student responses and examples of usage, with counts of occurrences as answers (**A**) and initiatives (**I**).

Affect type	A	I	Example
AMAZEMENT	0	1	*Wow*
AMUSEMENT	0	1	*Ha ha*
APOLOGY	4	14	*Sorry*
COMPREHENSION	6	6	*I get it*
CONFUSION	1	7	*I'm a bit confused*
CONTEMPLATION	14	5	*Hmmm*
CURIOSITY	0	2	*I'm curious*
DIFFICULTY	0	2	*I'm having difficulty*
FEEDBACK	0	6	*That was helpful*
GRATITUDE	0	14	*Thank you*
GREETING	0	1	*Good morning*
PAIN	0	1	*Ouch*
REALIZATION	5	9	*Ahh*

Coding of Affect

For coding affect a similar procedure to that above was followed. Evens and Bhatt scanned the text comprising the sessions K52-K55 and searched for instances of student affect together, discussing potential instances. A set of categories was derived from these initial analyses, and the remaining sessions (K56-K76) were then coded independently by both researchers. The results were then discussed until a consensus was

reached on each instance. Table 2 lists the final categorization of the types of affect found in the data, with counts and examples. Transcripts were electronically marked up using SGML tags as above.

Parenthetically, identifying affect in student responses was quite straightforward. In fact, almost every expression of affect was explicitly signaled by the student. This is encouraging for the use of affectual cues by computer tutoring systems, since in a text-based medium it is very difficult, if not impossible, to deduce students' emotional states from implicit cues (such as sarcasm).

Results and Discussion

Hedging in Human Tutoring

Hedged answers occur on average 6.04 times per session (σ=3.77). The different kinds of hedges are given in Table 1. The two most common types by far (together accounting for more than half of all occurrences) are Q1, adding a question mark to an answer otherwise in statement form (possibly expressing a sort of "questioning intonation"), and THINK, expressing a modal likelihood assessment via grammatical metaphor.

The majority of hedged answers are correct (57.6%, N=151), and so hedging does not provide a clear-cut signal of misunderstanding on the part of the student, so the data do not support *H2b: Hedges Wrong*. However, an even larger majority of non-hedged answers are correct (80.1%, N=359). This difference is significant (one-sided p<0.001), supporting *H2a: Hedges Inform*. Indeed, wrong answers are almost twice as likely to be hedged than correct answers (42.7% versus 26.3%).

In contrast to other work, we found gender to make no significant difference in hedging answers, as women hedge answers an average of 5.46 times per session, whereas men do so 6.66 times, well within the statistical variation of our sample. Hence *H3c: Women Hedge* is not supported. No gender-linked difference was found for correctness of hedged answers either, with women and men averaging 59.1% and 56.2% correct for hedged responses, respectively.

Hedging in Machine Tutoring

Surprisingly, there was only a single hedge in all 66 H/C sessions, clearly supporting *H1a: Hedging Differs*. In this sole example the student hedges an answer with a spurious statistic "9/10" when "all", or no marker at all, would have been more correct:

S: 9/10 times the dr will dominate because the rr can't bring all the way back

Affect in Human Tutoring

Expressions of affect are fairly common in the H/H sessions; with large variations, however, between different students. Out of twenty-five sessions, twenty-two contained at least one instance of student affect, while three had none at all. The most common type is APOLOGY, with eighteen occurrences overall. Instances of affect occur 3.52 times per session (σ=2.65), with a very high level of variation between students.

Men and women express affect at similar overall rates, with average numbers of 3.66 and 3.38 occurrences per student, respectively, so *H3b: Women are Affectual* is not supported. On the other hand, although all thirteen of the sessions involving female students include at least one expression of affect, three of the male-student sessions do not. Fisher's exact test on these data gives p=0.096, so that we may (barely) reject the null hypothesis that the same fraction of men as women are likely to express affect in tutorial sessions. This supports a weaker version of *H3b*—although some men express a lot of affect, men are more likely than women to show no affect at all.

Considering just apologies (the overall most frequent expression of affect), χ^2 testing for two independent samples gives p=0.12, so the data do not permit rejection of the null hypothesis that men and women apologize at similar rates, and so we cannot support *H3c: Women Apologize*.

Affect in Machine Tutoring

There were more examples of affect than of hedging in the H/C sessions, but the 20 instances of affect found in 66 H/C sessions are still far fewer than the 88 instances found in just 25 H/H sessions. Moreover, only 12 sessions (18%) contained any affect at all, as opposed to 22 (88%) of the H/H sessions. Thus we find that our data clearly support *H1b: Affect Differs*.

Even more significant than the large difference in frequency of affect is the difference in the *kinds* of affect that students expressed when interacting with a computer system. We saw none of the kinds of affect listed in Table 2 that we found in the H/H sessions—affect-related expressions in the H/C sessions tended to be more confrontational than with a human tutor. Although some instances of affect did seem to be

genuine expressions of feeling, some seemed more designed to push and test the system. Glass (1999) reported even more hostile input to an earlier version of the system. We therefore classed such responses into 3 categories: Hostile (5 responses), Testing (4 responses), and Refusal-To-Answer (11 responses). For example, student T48 seemed to get annoyed with the system as these two "Hostile" excerpts indicate:

T: Why did you enter 'no change' for TPR?
S: you know why.
. . . .
T: Why is MAP still decreased?
S: I don't want to tell you.

T74 seems pretty annoyed too:

T: Why is MAP still decreased?
S: blalaal

However, student T60 is clearly trying to "test" the system:

T: Why did MAP change in the manner that you predicted?
S: In other words, <student's name> knows all...

So is T81, we think, but perhaps this was simple honesty:

T: Why did you enter 'no change' for TPR?
S: Nimesh said so

Conclusions

Our results clearly show strong differences in student use of hedges and expressions of affect, depending on whether they are being tutored by a human or a computer ITS. While all students hedge in sessions with human tutors, they do not hedge at all in the machine sessions (with one exception). This conclusion is also supported by experience with the Why2-ATLAS system (Rosé et al. 2002); Carolyn Rosé told us that they do not see hedging either, though they looked for it since they had also observed it frequently in human tutoring sessions (Rosé, personal communication). The progress of speech-enabled tutoring (Bratt et al. 2002) is of great interest; it is possible that a difference in communication modality can affect student hedging behavior. As well, decoding students' affect may be easier from speech, due to tonal and prosody cues (Forbes-Riley & Litman 2004).

One specific result of importance to ITS is that hedging is not a clear indication of student uncertainty or misunderstanding, as had been believed. Indeed, examination of the types of hedges most used by students leads us to believe that hedges are more connected to issues of conversational flow and politeness, rather than expression of uncertainty. This interpretation is implied by the two most common forms of hedges in our data; Q1 uses a question mark to demand a response (confirmation?) from the tutor, while THINK expresses a modal assessment via a subjective metaphor, rather than a more direct modal verb or adjunct, thus requesting that the tutor respond to the student's mental state. Further research will be needed to examine this interpretation more closely.

As opposed to hedging, students do express affect to machines, though far less often than to humans. The real difference is in the *kind* of affect expressed, though—students do not apologize to computers, nor do they thank them or give them direct feedback; they do, however, express confusion and frustration. Together with our results on hedging, this leads us to suspect that the fact that students know they are interacting with a computer changes their attitude towards the conversation, contra Reeves and Nass (1996), and they are less concerned with helping to keep the flow going than they are in 'normal' conversation (Sacks et al. 1974).

In future work, we will look at hedging and affect in more human tutoring sessions. We wonder if the fact that Michael and Rovick practice the motivational techniques described by Lepper et al. (1993) influences the fact that they receive more positive affective input. This will help us to better understand how tutor style might encourage more useful hedging and expression of affect. Currently, we are concentrating on investigating the responses made by human tutors to student expressions of distress, in order to develop rules to make CIRCSIM-Tutor more friendly and responsive.

Acknowledgments

This work was supported by the Cognitive Science Program, Office of Naval Research, under Grants No. N00014-94-1-0338 and N00014-02-1-0442 to Illinois Institute of Technology, and Grant N00014-00-1-0660 to Stanford University. The content does not reflect the position of policy of the government and no official endorsement should be inferred.

This work would have been impossible without the expert tutoring of Joel Michael and

Allen Rovick of Rush Medical College and their determination to create effective machine tutors. Thanks also to the anonymous reviewers whose expert comments helped improve this paper.

References

Aist, G., B. Kort, R. Reilly, J. Mostow, & R. Picard. (2002). Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. *ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, San Sebastian, Spain.

Aries, E. (1989). Gender and communication. In P. Shaver and C. Hendrick, eds., *Sex and Gender, Vol. 7 of the Review of Personality and Social Psychology.* Newbury Park, CA: Sage. 149-176.

Bhatt, K.S. (2004). Classifying student hedges and affect in human tutoring sessions for the CIRCSIM-Tutor intelligent tutoring system. MS Thesis, Department of Computer Science, Illinois Institute of Technology, Chicago, IL.

Bratt, E.O., B.Z. Clark, Z. Thomsen-Gray, S. Peters, P. Treeratpituk, H. Pon-Barry, K. Schultz, D.C. Wilkins, & D. Fried. (2002). Model-based reasoning for tutorial dialogue in shipboard damage control. *Proceedings of ITS 2002,* San Sebastian, Spain, June. 63-69.

Breazeal, C. (1999). A motivational system for regulating human-robot interaction. In *Proceedings of AAAI98*, Madison, WI. 54-61.

Brooks, R., C. Breazeal, R. Irie, C. Kemp, M. Marjanovic, B. Scassellati, & M. Williamson (1998), Alternative essences of intelligence. *Proceedings of AAAI98*, Madison, WI. 961-967.

Colby, K. (1975). *Artificial paranoia.* New York, NY: Pergamon Press.

Elliott, Clark. (1998). Hunting for the holy grail With "emotionally intelligent" virtual *a*ctors, *ACM SIGART Bulletin,* 9(1) 20-28.

Forbes-Riley, K. and Litman, D. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceeedings of HLT-NAACL 2004,* May, Boston, MA. 201-208.

Fox, B. (1993). *The human tutorial dialogue project.* Hillsdale, NJ: Erlbaum.

Glass, M.S. (1999). *Broadening input understanding in a language-based intelligent tutoring system.* Ph.D. Thesis, Department of Computer Science, Illinois Institute of Technology, Chicago, IL.

Goldstein, M., G. Alsiö, J. Werdenhoff. (2002). The media equation does not always apply: People are not polite towards small computers. *Personal and Ubiquitous Computing. Berlin: Springer-Verlag* 6:87-96.

Holmes, J. (1984): Women's language: A functional approach. *General Linguistics* **24**(3):149-178.

Kort, B. & R. Reilly. (2002). An affective module for an intelligent tutoring system, *Intelligent Tutoring Systems* 2002.

Lakoff, R. (1975). *Language and woman's place*. New York, NY: Harper and Row.

Lakoff, R. (1990). *Talking power: The politics of language*. New York, NY: Basic Books.

Lepper, M. R., M. Woolverton, D. L. Mumme, and J-L. Gurtner. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. P. Lajoie and S. J. Derry (Eds.), *Computers as cognitive tools*, Hillsdale, NJ: Erlbaum, 75-105.

Michael, J.A., A.A. Rovick, A.A., M.S. Glass, Y. Zhou, & M. Evens (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments, 11(3),* 233-262.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge, UK: Cambridge University Press.

Rosé, C.P., D. Bhembe, A. Roque, S. Siler, R. Shrivastava, & K. VanLehn. (2002). A hybrid natural language understanding approach for robust selection of tutoring goals. In S.A. Cerri, G. Gouardères, & F. Paraguaçu (eds.) *ITS 2002,* LNCS 2363. Berlin: Springer-Verlag. 552-561.

Sacks, H., E.A. Schegloff, & G. Jefferson. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language,* **50:**696-735.

Shechtman, N., & L.M. Horowitz (2003). Media inequality in conversation: How people behave differently when interacting with computers and people. *CHI* 5(1): 281-288.

Shah, F., M.W. Evens, J.A. Michael, & A.A. Rovick. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions, *Discourse Processes,* 33(1) 23-52.

Shortliffe, E.H. (1982). The computer and clinical decision-making: Good advice is not enough. *IEEE Engineering in Medicine and Biology Magazine,* **1**(1), 16-18.

Tannen, D. (1990). *You just don't understand: Women and men in conversation.* New York, NY: William Morrow and Co.

Thompson, B. H. (1980). Linguistic analysis of natural language communication with computers. *Proceedings of the 8th International Conference on Computational Linguistics COLING 80)*, Tokyo, Japan, np.

Vicente, A. de, & H. Pain. (2002). Informing the detection of the students' motivational dtate: An empirical study. *Intelligent Tutoring Systems 2002*.

Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language interaction between mind and machine. *CACM* 9(1) 36-44.

Incorporating Self Regulated Learning Techniques into Learning by Teaching Environments

Gautam Biswas (gautam.biswas@vanderbilt.edu), Krittaya Leelawong, Kadira Belynne, Karun Viswanath, and Nancy Vye
Department of EECS & ISIS, Box 1824 Sta B, Vanderbilt University
Nashville, TN 37235 USA

Daniel Schwartz (daniel.schwartz@stanford.edu), Joan Davis
School of Education, Stanford University
Stanford, CA 94305 USA

Abstract

This paper discusses Betty's Brain, a teachable agent in the domain of ecosystems that combines learning by teaching with self-regulation mentoring to promote deep learning and understanding. Two studies demonstrate the effectiveness of this system. The first study focused on components that define student-teacher interactions in the learning by teaching task. The second study examined the value of adding meta-cognitive strategies that governed Betty's behavior and self-regulation hints provided by a mentor agent. The study compared three versions: an intelligent tutoring version, a learning by teaching version, and a learning by teaching plus self-regulation strategies. Results indicate that the addition of the self-regulation mentor better prepared students to learn new concepts later, even when they no longer had access to the self-regulation environment.

Introduction

The recent proliferation in computer-based learning environments has produced a number of tutoring systems (Wenger, 1987) and pedagogical agents (Johnson, et al., 2000). The typical intelligent tutoring system curriculum is problem-driven. The system selects problems for the user to solve, and provides feedback on the solutions generated. The tutoring paradigm has been very successful. At the same time, it often emphasizes localized feedback, and does not always help students practice higher-order cognitive skills especially in complex domains (e.g., picking what questions to ask or how to examine resources for learning). Problem solving in complex domains requires active decision-making by learners in terms of setting learning goals and applying strategies for achieving these goals. The current paper examines ways to address these latter goals using an "intelligent" learning environment.

Our goal has been to introduce effective learning paradigms that advance the state of the art in computer-based learning systems and support students' abilities to learn, even after they leave the computer environment. Our approach has been to create environments where students teach computer agents. This paper reports the results of two studies. One study explored different features of a specific learning by teaching environment, Betty's Brain. The second study manipulated the metacognitive support students received when teaching "Betty" and measured its effects on the students' abilities to subsequently learn new content several weeks later.

The cognitive science and education research literature supports the idea that teaching others is a powerful way to learn. Research in reciprocal teaching, peer-assisted tutoring, programming, small-group interaction, and self-explanation hint at the potential of *learning by teaching* (Palinscar & Brown, 1984; Cohen, et al. 1982; Papert, 1993; Chi, et al., 1994). Bargh and Schul (1980) found that people who prepared to teach others to take a quiz on a passage learned better than those who prepared to take the quiz themselves. The literature on tutoring has shown that tutors benefit as much from tutoring as their tutees (Chi, et al., 2001; Graesser, et al., 1995). Biswas et al. (2001) report that students preparing to teach made statements about how the responsibility to teach forced them to gain deeper understanding of the materials. Other students focused on the importance of having a clear conceptual organization of the materials. Additionally, teachers can provide explanations and demonstrations during teaching and receive questions and feedback from students. These activities seem significant from the standpoint of their cognitive consequences in improving understanding of complex concepts.

A key benefit of the learning by teaching process focuses on the need to structure knowledge in a compact and communicable format. This requires a level of abstraction that may help the teacher develop important explanatory structures for the domain. For example, many people find that preparing a conference presentation helps them decide which concepts deserve the "high level" status of introductory framing. The need to structure ideas not only occurs in preparation for teaching, but can also occur when teaching. Good learners bring structure to a domain by asking the right questions to develop a systematic flow for their reasoning. Good teachers build on the learners' knowledge to organize information, and in the process, they find new knowledge organizations, and better ways for interpreting and using these organizations in problem solving tasks.

Despite its potential benefits, learning-by-teaching can initially seem inefficient. For example, students may need to learn the right way to teach, which can slow down their learning of the subject matter in the short run. At the same time, learning-by-teaching may have long-term benefits in

that it helps students appreciate what a complete and communicable answer needs to look like, and they may learn how to consult resources to understand deeply enough that they can teach well. In this case, it seems important to evaluate not only how well students learn the target knowledge of the teaching episode, but also how well they are prepared to learn in the future as a result of learning-by-teaching (Bransford & Schwartz, 1999).

We have adopted a new approach to designing learning-by-teaching environments that ideally supports the learning outcomes described above, provide tools that enable users to visually organize and reason about their domain knowledge as they teach a computer agent, and include feedback to promote better self-regulation during the learning and teaching processes. A key challenge to the learning-by-teaching approach is that students are usually novices with regard to domain content and teaching tasks. To help with the domain content, our design includes content-integrated instruction that encourages students to access and think about resources, and check their reasoning during the teaching (and learning) process by interacting with the teachable agent and assessing its performance. To help with the teaching and learning aspects, we have made the computer agent more participatory in the learning process, and developed a Mentor agent that acts as a "meta-cognitive" coach, and provides strategy and content feedback about teaching with understanding, while avoiding the very specific localized feedback that is characteristic of many tutoring systems. Ideally the combination of the two can help students not only learn the content of a specific lesson, but also prepare students to learn in the future when they no longer have access to the system.

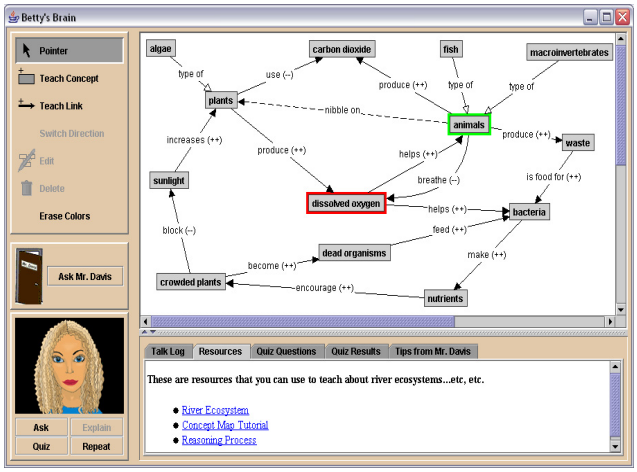
Implementing Learning by Teaching Systems

Our teachable agents (TAs) provide important structures to help shape the thinking of the *learner-as-teacher*. Each agent manifests a visual structure that is tailored to a specific form of knowledge organization and inference. In general, our agents try to embody four principles of design:

- Teach through visual representations that organize the reasoning structures of the domain (e.g., directed graphs and matrices).
- Build on well-known teaching interactions to organize student activity (e.g., teaching by "laying out," teaching by example, teaching by telling, teaching by modeling).
- Ensure the agents have independent performances that provide feedback on how well they have been taught (each agent uses a distinct AI reasoning technique, such as qualitative reasoning, logic, and genetic algorithms).
- Keep the start-up costs of teaching the agent very low (as compared to programming). This occurs by only implementing one modeling structure with its associated reasoning mechanisms.

Betty's Brain makes her qualitative reasoning visible through a dynamic, directed graph called a *concept map* (Novak, 1996). The fact that TAs represent knowledge structures rather than the referent domain is a departure from many simulation-based learning environments. Simulations often show the behavior of a physical system, for example, how an algal bloom increases the death of fish. On the other hand, TAs simulate the behavior of a person's thoughts about a system. Learning empirical facts is important, but learning to use the expert structure that organizes those facts is equally important. Therefore, we have structured the agents to simulate particular forms of thought that may help teacher-students structure their thinking about a domain.

Fig. 1 illustrates the interface of Betty's Brain. Students explicitly teach Betty using a graphical drag and drop interface to create and modify their concept maps in the top pane of the window. They use the *Teach Concept* button to create new concepts, and the *Teach Link* button to create relations between concepts. When teaching the agent about relationships, students use a popup template to specify the name (e.g., *breathe*, *produce*, *helps*) and type of relationship *(causal*, *type of*, and *descriptive*). For causal relations, students indicate whether the relation implies an *increase* (++) or *decrease* (–-). For example, in the map in Fig.1, the concept map implies an increase in *fish* will result in a decrease in *dissolved oxygen*. Note that the student generates all concept and relationship names. They are not chosen from a menu.



Once taught, Betty reasons with her knowledge and answers questions. Users can formulate their own queries using the *Ask* button, and observe the effects of their teaching by analyzing Betty's responses. Templates are provided to ask Betty two kinds of questions: (i) *If <concept A> increases (decreases) what happens to <concept B>?* and (ii) *Tell me all you know about <concept A>*. For the latter question, Betty enumerates all the concepts that are directly linked to *<concept A>*. For the former question, Betty uses qualitative reasoning methods to derive her answers to question through a chain of causal inferences. For example, using the concept map in Fig. 1, Betty can conclude that an increase in *algae* will cause *fish* to increase.

Betty also provides explanations for how she derives her answers by depicting the derivation process using multiple modalities: text, animation, and speech. Details of the rea-

soning and explanation mechanisms in Betty's Brain are presented elsewhere (Leelawong, et al., 2001).

We should clarify that Betty does not use machine learning algorithms to achieve automated learning from examples, explanations, and induction. Our focus is on the well-defined schemes associated with teaching that support a process of instruction, assessment, and remediation. These schemas help organize student interaction with the computer, much as people's well-defined schemas for spatial organizations helped to create the desktop metaphor for windows-based computer systems.

The system also includes sets of teacher-generated quiz questions. Betty can take the quiz, and students see how she performs and receive the correct answer. The quiz questions are structured to provide students cues on concepts and relations that are important in the domain of study. Examples of some quiz questions are shown in Fig. 2.

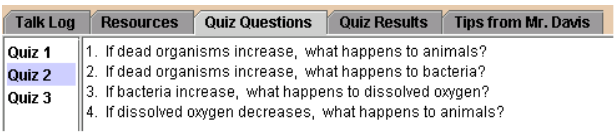


Figure 2: Quiz Questions

A Prior Study without a Self-Regulation Mentor

To study the effectiveness of Betty's Brain we conducted an experiment on 50 high-achieving fifth grade students from a science class in an urban public school located in a southeastern US city. The students were asked to teach Betty about river ecosystems. We examined the effects of the interactive features of the teachable agent environment using a 2×2 between-subjects design. One group of students could submit their agent to take a Quiz (and receive feedback on the correct answer). A second group could Query their agent by generating their own questions and seeing how Betty chains through her map to reach the answer (there was no expert feedback on the answer). The third condition, which could neither Query nor Quiz the agent, was basically using a graphing package. Students who had both Query and Quiz features could ask Betty questions and see her performance on the quiz questions. Students were given instructions on how to use the system, and then they used the software for 3 one-hour sessions. To help students learn what to teach, reference materials were made available during and in between their teaching sessions with Betty.

We hypothesized that having the query feature would help students debug their own thinking and reasoning in the problem domain, and this would result in maps with more inter-linked concepts. Betty's answers and her explanations would make explicit the process of reasoning across chains of links in a concept map. For the Quiz condition, we expected that students would map backward from the quiz questions and use the feedback they received about her answers to produce more accurate concept maps.

Analysis of the scope of students' maps and the types and accuracy of links contained therein are presented in (Leelawong, et al., 2002). On the positive side, students who used the Query and/or Quiz mechanisms understood causal relations better than the students who did not. This was reflected in their concept maps, which had a larger proportion of causal links than the No Quiz and No Query group. As predicted, students who had access to the Query feature had the most inter-linked maps and most elaborate reasoning chains. The Quiz feature was effective in helping students decide the important domain concepts and types of relationships to teach Betty.

We also noted some negative aspects to our system. Our observations of students during the study suggested that students who had the quiz feature were too focused on *"getting the quiz questions correct"* rather than *"making sure that Betty (and they themselves) understood the information"* (Davis, et al., 2003). The activity logs of the students who used the quiz showed a pattern of quick one-link corrections followed by a retake of the quiz. The query mechanism and resources were used sparsely, and it is unlikely they gained a deep understanding of causal structures. On the other hand, the Query-only group spent more time with Betty's explanations and reading resources. Surprisingly, students who had the query feature without the benefit of quiz feedback produced as many valid relevant causal links as the conditions with the quiz and quiz and query feature. This demonstrated the value of explicitly illustrating the reasoning process (by having Betty explain her answers) so that students understand causal structures.

Reflections on these results made us rethink our design and implementation of TA environments. A primary concern was the student's focus on getting quiz questions right without trying to gain an understanding of interdependence and balance in river ecosystems. We realized that interactions between the student-as-teacher, Betty, and the quiz feature had to be improved to facilitate better learning. Further, in exit interviews, students emphasized that they would have liked Betty to be more active and exhibit characteristics of a good student during the teaching phase (Davis, et al., 2003). Several students suggested that we should "*do some sort of game or something and make the system more interactive,*" and "*Betty should react to what she was being taught, and take the initiative and ask more questions on her own.*" Consistent with this feedback, we noted that the first version of Betty was passive and only responded when asked questions. We believed that to create a true learning by teaching environment, Betty needed to better demonstrate qualities of human students. A tutor gains deeper understanding from interactions with a tutee (Chi, et al., 2001) that includes answering the tutee's questions, explaining materials, and discovering misconceptions. Betty should be designed to benefit her users in the same way.

Self-Regulated Learning and Betty's Brain

As mentioned earlier, an important realization from this first study was that we were dealing with young children who were novices in teaching practice and in domain knowledge content. To accommodate this, the

learning environment was redesigned to provide appropriate scaffolds and proper feedback mechanisms to help students overcome their initial difficulties in learning and teaching about a complex domain. The scaffolds took on three primary forms. First, we made improvements in the online resources available for learning about river ecosystems. We reorganized the resources to emphasize the concepts of interdependence and balance. This changed the partitioning of the resources to the three primary cycles that govern ecosystem behavior: (i) the oxygen/carbon dioxide cycle, (ii) the food chain, and (iii) the decomposition cycle. A hypertext implementation allowed direct access to sections and subsections. An advanced keyword search technique provided access to information using keywords. (Students in the study below found the resources to be much more useful, and used them extensively while teaching Betty.)

The second change is that we redesigned the quiz so that the questions would support users in systematically building their knowledge about river eco-systems. The questions were no longer randomly sampled from the full domain, but they gradually introduced more complex questions. Furthermore, the first item in each quiz was a comprehensive question that covered all of the domain concepts and relations associated with a particular cycle. This prevented students from taking a sequential approach of building the concept map to answer one question at a time. We also improved the feedback the students received.

These two changes were important, but we doubted they would be sufficient in supporting users in becoming better learners and teachers, nor did they address our users requests for a more *"life like"* Betty. Therefore, our third change, and most relevant to the study below, was to make Betty more reactive to what she was being taught, as well as to use self-regulation strategies in her interactions with her student-teacher. Along with this, we added a mentor agent to the system to help users observe and develop metacognitive and self-regulation strategies to support active and independent learning. Self-regulated learning should be an effective framework for providing feedback because it promotes the development of higher-order cognitive skills (Corno & Mandinach, 1983), and it is critical to the development of problem solving ability (Pintrich & DeGroot, 1990).

Our new design adopted some aspects of the framework of *self-regulated learning,* described by Zimmerman (1989) as situations where students are "*metacognitively, motivationally, and behaviorally participants in their own learning process.*" Self-regulated learning strategies involve actions and processes that can help one to acquire knowledge and develop problem solving skills (Pintrich & DeGroot, 1990). Zimmerman describes a number of self-regulated learning skills that include goal setting and planning, seeking and organizing information, keeping records and monitoring, and self-evaluation. We developed mechanisms by which Betty forced the student to conform to the self-regulation strategies. In parallel, the Mentor agent included resources that helped students develop these skills during their learning and teaching.

This resulted in a number of changes to Betty's Brain. For example, when a student begins the teach phase by constructing the initial concept map, both the Mentor and Betty make suggestions that the student *set goals* about what to teach, and make efforts to gain the relevant knowledge by studying the river ecosystem resources. The Mentor continues to emphasize the reading and understanding of resources, whenever the student has questions on *how to improve their learning*. The user is given the opportunity to *evaluate her knowledge* while studying. If she is not satisfied with her understanding, she may *seek further information* by asking the Mentor for additional help. While teaching, the student as teacher can interact with Betty in many ways, such as asking her questions (*querying*), and getting her to take *quizzes* to evaluate her performance. Users are given a chance to predict how Betty will answer a question so they can check what Betty learned against what they were trying to teach.

Some of the self-regulation strategies manifest through Betty's persona. These strategies make Betty more involved during the teach phase, and drive her interactions and dialog with the student. For example, during concept map creation, Betty spontaneously tries to demonstrate *chains of reasoning*, and the conclusions she draws from this reasoning process. She may query the user, and sometimes remark (right or wrong) that an answer she is deriving does not seem to make sense. This is likely to make users reflect on what they are teaching, and perhaps, like good teachers they will assess Betty's learning progress more often. At other times, Betty will prompt the user to *formulate queries* to check if her reasoning with the concept map produces correct results. There are situations when Betty emphatically refuses to take a quiz because she feels that she has *not been taught enough*, or that the student has not given her *sufficient practice by asking queries* before making her take a quiz.

After Betty takes a quiz offered by the Mentor agent, she discusses the results with the user. Betty reports: (i) her view of her performance on the particular quiz, and if her performance has improved or deteriorated from the last time she took the quiz, and (ii) the Mentor's comments on Betty's performance in the quiz, such as: "*Hi, I'm back. I'm feeling bad because I could not answer some questions in the quiz. Mr. Davis said that you can ask him if you need more information about river ecosystems.*") The Mentor agent's initial comments are general, but they become more specific if errors persist, or if the student seeks further help (*"You may want to study the role of bacteria in the river"*).

In addition to self-regulation advice that included information on how to be a better learner and better teacher, the domain content feedback from the Mentor agent was directed to make the student think more about interdependence among concepts. Students seeking specific help were first directed to relevant sections in the resources for further study and reflection, rather than being told what was wrong in their concept maps. When the Mentor provided specific feedback, it was about *chains of events* to help students better understand

chains of events to help students better understand Betty's reasoning processes.

Overall, we believe that the introduction of self-regulation strategies provides useful scaffolds to help students learn about a complex domain, while also developing metacognitive strategies that promote deep understanding and abilities to learn in the future. One of the achievements of the new system is that students retain control rather than being told what to do (e.g., they need to request help from the mentor and they teach Betty). Only when the student seems to be hopelessly stuck, does the Mentor spontaneously intervene to help students advance in their learning (and teaching) task.

A Study of the Added-Value of Self-Regulation

A new experiment with fifth graders was designed to compare the Teachable Agent system with the self regulation mentor (SRL) against two other approaches: (i) A learning by teaching (LBT) version that was similar to the Query & Quiz version before, and (ii) An externally-guided learning system (ITS) designed with a pedagogical agent. In the ITS version, the pedagogical agent asked students to create concept maps that could answer a set of quiz questions (therefore, there was no *teaching* component), and the agent would provide feedback on how to correct their map when their quiz answers had errors. All three groups had access to identical resources on river ecosystems and the same query and quiz features. To evaluate student learning, we examined pre-posttest scores, how they used the system, the quality of their final maps, and their ability to reproduce the maps subsequently. Importantly, several weeks later, we asked the students to learn about the Nitrogen cycle, which had not been covered during the initial instruction. This permitted us to determine which group had been better prepared to learn, once they no longer could rely on the scaffolds of their respective version. Our expectation was that the SRL students would do better on this latter measure of preparation for future learning (Bransford & Schwartz, 1999), because they had learned how to "take charge" of their own learning.

Experimental Procedure

A fifth grade classroom was divided into three equal groups of 15 students each using a stratified sampling method based on standard achievement scores in mathematics and language. The students worked on a pretest with twelve questions before they were separately introduced to their particular versions of the system. The three groups worked for six 45-minute sessions over a period of three weeks to create their concept maps. All groups had access to the online resources while they worked on the system.

All three conditions had the same quiz questions while working with the system, and they had access to the query feature and Mentor agent (Mr. Davis), though he appeared with different capacities. The task given to the ITS group was to create concept maps that correctly answered the 16 questions that were divided up into three quizzes. They had the same interface to create and modify their concept maps as the other groups, but Betty did not exist in the ITS system. The ITS feedback came from the Mentor, who told students if their map held the correct answers to the quiz questions and provided hints on how the students could correct their maps. The two other groups, LBT and SRL, were told to teach Betty and help her pass a test so she could become a member of the school Science club. Both of these groups had access to the three quizzes. The LBT group only received mentor feedback about the quality of Betty's specific answers to the quiz. The SRL group received more extensive feedback from the Mentor, but only when they queried him. Coupled with the Mentor, the SRL Betty was also endowed with self-regulation strategies that governed her behavior. Therefore, the SRL condition was set up to develop more active learners by promoting the use of self-regulation strategies.

At the end of the six sessions, every student took a post-test that was identical to the pretest. Two other delayed post-tests were conducted about seven weeks after the initial experiment: (i) a *memory test*, where students were asked to recreate their ecosystem concept maps from memory (there was no help or intervention when performing this task), and (ii) a *preparation for future learning transfer test*, where they were asked to construct a concept map using on-line resources and answer questions about the land-based nitrogen cycle. Students had not been taught about the nitrogen cycle, so they would have to learn from resources during the transfer phase. (All three conditions simply used the concept mapping interface, resources, and "correct/incorrect" feedback from the mentor on several quiz questions.)

For learning about river ecosystems, students in all conditions improved from pre- to posttest on their

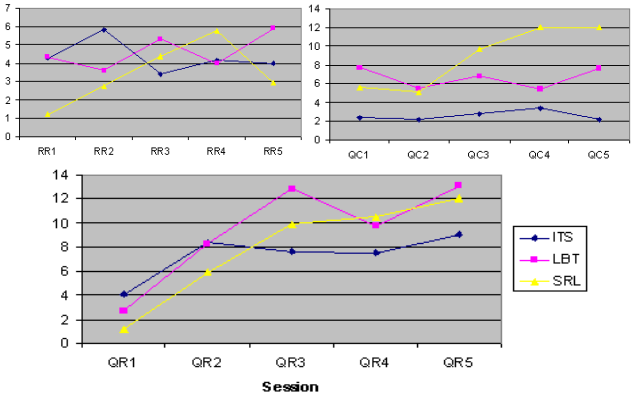


Figure 3: Resource Requests (RR), Queries Composed (QC), & Quizzes Requested QR) per session.

knowledge of interdependence (p's<.01, paired T-tests), but not ecosystem balance. There were few differences between conditions in terms of the quality of their maps. However, there were notable differences in their use of the system during the initial learning phase. Fig. 3 shows

the average number of resource, query, and quiz requests per session by the three groups. It is clear from the plots that the SRL group made a slow start as compared to the other two groups. This can primarily be attributed to the nature of the feedback; i.e., the ITS and LBT groups received specific content feedback after a quiz, whereas the SRL group tended to receive more generic feedback that focused on self-regulation strategies. Moreover, in the SRL condition, Betty would refuse to take a quiz unless she felt the user had taught her enough, and prepared her for the quiz by asking questions. After a couple of sessions the SRL group showed a surge in map creation and map analysis activities, and their final concept maps and quiz performance were comparable to the other groups. It seems the SRL group spent their first few sessions in learning self-regulation strategies, but once they learned them their performance improved significantly.

For the delayed memory test, the table below presents the mean number of expert causal links and concepts in the student maps. Results of ANOVAs using Tukey's LSD to make pairwise comparisons showed that the SRL group recalled significantly more links that were also in the expert map (which nobody actually saw).

Student Map Included:	SRL Mean (se)	LBT Mean (se)	ITS Mean (se)
Expert Concepts	6.7 (.6)	6.4 (.5)	5.8 (.6)
Expert Causal Links	3.3[a] (.6)	1.7 (.6)	2.0 (.6)

[a] Significantly greater than LBT, $p < .05$

We thought that the effect of SRL would not be to improve memory, but rather to provide students with more skills for learning subsequently. When one looks at the results of the test of preparation for future learning, the differences between the SRL group and the other two groups are significant. The table below summarizes the results of the transfer test, where students read resources and created a concept map for the land-based nitrogen cycle. There are significant differences in the number of expert concepts in the SRL versus ITS group maps, and the SRL group had significantly more expert causal links than the LBT and ITS groups. When learning about the river ecology, the SRL students had received some guidance in how to use resources productively and how to think about the quality of their map. This guidance transferred to learning about the nitrogen cycle.

Student Map Included:	SRL Mean (sd)	LBT Mean (sd)	ITS Mean (sd)
Expert Concepts	6.1[a] (.6)	5.2 (.5)	4.1 (.6)
Expert Causal Links	1.1[ab] (.3)	0.1 (.3)	0.2 (.3)

[a] Significantly greater than ITS, $p < .05$;
[b] Significantly greater than LBT, $p < .05$

Conclusions

The results demonstrate the significant positive effects of SRL strategies in understanding and transfer in a learning by teaching environment. Students in all three groups demonstrated the same learning performance in traditional learning tasks, but the SRL group outperformed the other two in the far transfer test. We believe that the differences between the SRL and the other two groups would have been more pronounced if the transfer test study had been conducted over a longer period of time. Lastly, we believe that the concept map and reasoning schemes have to be extended to include temporal reasoning and cycles of behavior to facilitate students' learning about the concept of balance in ecosystems.

References

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology, 72*(5), 593-604.

Biswas, G., Schwartz, D., Bransford, J., & TAG-V. (2001). Technology Support for Complex Problem Solving: From SAD Environments to AI. In Forbus & Feltovich (eds.), *Smart Machines in Education* (pp. 71-98). Menlo Park, CA: AAAI Press.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* , **24**, 61-101. Washington DC: American Educational Research Association.

Chi, M. T. H., De Leeuw, N., Mei-Hung, C., & Levancher, C. (1994). Eliciting self explanations. *Cognitive Science, 18*, 439-477.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from Human Tutoring. *Cognitive Science, 25*(4), 471-533.

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of peer tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237-248.

Corno, L., & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychology, 19*(2), 88-108.

Davis, J. M., Leelawong, K., Belynne, K., Bodenheimer, R., Biswas, G., Vye, N., et al. (2003). Intelligent User Interface Design for Teachable Agent Systems. *Proc. Intl. Conf. on Intelligent User Interfaces*, Miami, Florida, 26-34.

Graesser, A. C., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychologist, 9*, 359-387.

Leelawong, K., Davis, J., Vye, N., Biswas, G., Schwartz, D., Belynne, K., et al. (2002). The Effects of Feedback in Supporting Learning by Teaching in a Teachable Agent Environment. *Proc. Fifth Intl. Conference of the Learning Sciences*, Seattle, Washington, 245-252.

Leelawong, K., Wang, Y., Biswas, G., Vye, N., & Bransford, J. (2001). Qualitative reasoning techniques to support learning by teaching: The Teachable Agents project. *Proc. Fifteenth Intl. Workshop on Qualitative Reasoning*, San Antonio, Texas, 73-80.

Novak, J. D. (1996). Concept Mapping as a tool for improving science teaching and learning. In D. F. Treagust, R. Duit & B. J. Fraser (eds.), *Improving Teaching and Learning in Science and Mathematics* (pp. 32-43). London: Teachers College Press.

Palincsar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension -monitoring activities. *Cognition and instruction, 1*, 117-175.

Papert, S. (1993). *The Children's Machine: Rethinking school in the age of the computer*. New York, NY: Basic Books.

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.

Zimmerman, B. J. (1989). A Social Cognitive View of Self-Regulated Academic Learning. *Journal of Educational Psychology, 81*(3), 329-339.

Acknowledgements

This work has been supported by a NSF ROLE 0231771 & 0231946. The assistance provided by John Bransford and John Kilby is gratefully acknowledged.

Error-Reduction and Simplicity: Opposing Goals in Classification Learning

Mark Blair (mrblair@indiana.edu)
Indiana University, Department of Psychology
1101 E. Tenth Street, Bloomington, IN 47405 USA

Abstract

Studies of real world experts show that they use different and subtler regularities than novices to make effective classifications. In laboratory studies of learning however, participants have a strong preference for simple cue sets, even at the expense of accuracy. The present experiments investigate participants' ability to use subtle stimulus dimensions in order to eliminate category exceptions. Results show that some participants were able to use the optimal 3-cue set, but many could not. When there were two optimal cue sets, one with 2 dimensions and one with 3, participants favored the simpler set, even though it meant ignoring an obvious and diagnostic cue. Overall there were wide individual differences, with almost every cue set adopted by some participants. Current theories of attention that posit rapid shifts of learned attention offer promise in accounting for the results.

For any organism to adapt successfully it must become sensitive to meaningful regularities in the environment. Humans have developed flexible learning systems, allowing them to rapidly adjust to changing environments. Learning concepts, representations of classes of stimuli that 4require an equivalent response, conserves resources by reducing the amount of information that needs to be processed from the environment, and also allows for generalization to related, novel circumstances. Representing a complex environment, with abundant interdependencies and subtle regularities requires a rich set of concepts.

To perform this function, the human perceptual system can attend selectively, become sensitized to highly complex stimulus dimensions, and even create novel functional features. These processes affect the perception of a stimulus and therefore alter the representation of that stimulus. Differences in cue use and representation across experts and novices appear in many areas such as biology (Boster & Johnson, 1989), physics (Chi, Feltovich, & Glaser, 1981), computer programming (Davies, 1994), wine tasting (Solomon, 1997), bird identification (Johnson & Mervis, 1997) and pocket billiards (Blair & McBeath, 2001). Despite an abundance of differences between expert and novice differences in cue use, laboratory studies of learning (specifically "knowledge restructuring") have shown that participants have had a strong resistance to using new information and more complicated cue sets, even though they would afford better performance (Lewandowsky, Kalish & Griffiths, 2000).

To the extent that learning is error driven, exceptions in the cue set provide a powerful motivator to incorporate new dimensions, however, there is also a pressure toward simplicity. Additional dimensions, which may a space in which the categories separate, can be expensive to represent. Completely altering the dimensions used for categorization can require more energy than using unique stimulus-level elements to classify exceptions. This leads to the memorization of exceptions, rather than a refining of the cue set. It is clear that category learning is influenced by opposing forces; one to enlarge, and one to reduce the dimensionality of the cue set.

The present research examines the complementary processes of the expansion and reduction of the cue set toward effective and efficient representation. The goal of the present study is to verify that participants can and do optimize their cue sets using both expansion and reduction when learning categories with subtle dimensions. Previous research has demonstrated these complimentary processes by manipulating the stimulus set to provide a new dimension (Blair & Homa, 2003b). These studies produced wide individual differences, with many participants incorporating new dimensions to eliminate category exceptions, and many others choosing to rely on simple cue sets which result in many exceptions and significant error. The Blair and Homa (2003b) studies also showed that participants can shift to optimal spaces if they are less complex. Overall, in these studies participants demonstrated both the flexibility found in studies of expertise and the insistence on simplicity found in studies of knowledge restructuring.

Many real world category learning problems do not involve learning new information never experienced before, but rather involve learning to be sensitive to stimulus features that have existed all along, but may have been overlooked for more salient dimensions. For example in bird identification, color is an obvious perceptual cue, used by experts and novices alike. To tell the difference between a Hepatic Tanager and a Summer Tanager, both of which are predominantly red, one must notice the color of the bill and whether the bird has a gray ear patch or not. These are features that novices are prone to miss. The present studies use two obvious dimensions as well as a subtle third dimension as a direct analogy to those cases.

Experiment 1

In Experiment 1, a sequential presentation same-different task was employed to examine the discriminability of the three dimensions used in the remaining experiments, and an additional dimension (color) used in a related set of studies (Blair & Homa, 2003b). If the stimuli are to be used in later experiments, they should be of roughly equal discriminability, with the exception of tail bumpiness, which should be significantly less discriminable than the other three dimensions.

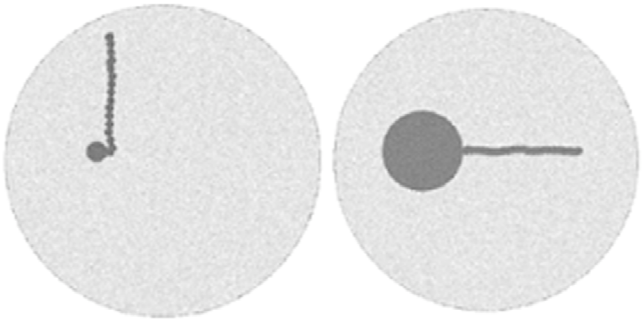


Figure 1: Two Example Stimuli. Tail angle, head diameter, and bumpiness of the tail are the three primary dimensions of variation. Stimuli could take on any one of 11 values on or between the two extremes shown here.

Method

Participants Participants were 26 undergraduates from Arizona State University who participated for course credit. All participants were naïve as to the purpose of the test.

Stimuli Examples of the stimuli are shown in Figure 1. These stimuli were created in the graphics program Adobe Photoshop. Designed to look like something seen under a microscope, these fictitious microorganisms offer a ready analogy to expertise domains such as medical diagnostics. Stimuli varied in one of four different dimensions: head diameter, tail angle, color, and tail bumpiness. Head diameter varied from 25 to 100 pixels. Tail angle varied from 0 to 90 degrees. Color, in RGB values, ranged from 65-75-230 to 139-75-30. These colors were set so that they were of equivalent luminosity, that is, they look the same shade of gray if viewed as a black and white image. The tail bumpiness was created by using brushes with "spacing" set from 60 to 90. The full range of the variation was broken up into 11 equal sized steps, thus dimensions were always one of those 11 values. In addition to possible variations in the four principle dimensions, the tail of each microorganism was hand drawn on top of a line of the correct angle and thus represented a source of stimulus-specific variation. A Gaussian noise filter (30-unit) was also applied to each stimulus. Afterward, a "crystallization" filter (3-pixel) was applied to the stimulus area around, but not including, the microorganism.

Procedure The entire experimental task, including instructions, was displayed on computer. The experimental task was a sequential same-different task in which participants were shown two stimuli, one at a time, and asked to judge whether they were the same or different. Participants were shown two example stimuli and the four main dimensions of variation were indicated. Participants were instructed that only variations on the four consequential dimensions should elicit a 'different' response, any other variations were to be ignored. On each trial, the first stimulus was shown for 2000 msec, and then the screen went blank for 1000 msec. Finally, the second stimulus appeared, and remained on screen until the participant responded. There were 144 trials and of the 144 stimulus pairs, 72 were 'same' pairs and 72 were 'different'. Of the 72 'same' trials, 36 were pairs showing exactly the same stimulus and 36 were pairs of stimuli that had the same values on the consequential dimensions, but were created separately. Of the 72 'different' pairs, there were three stimuli from each of three levels of variation (1, 2 and 4 units) from each of the four main dimensions (head diameter, tail angle, color, and tail bumpiness). The lower, upper and middle parts of the range of variation were used for each trio of stimuli associated with a level of variation. For example, for 1 unit variations on head diameter, the three stimulus pairs might include a pair with 25 and 32-pixel heads, a pair with 93 and 86-pixel heads and a pair with 48 and 54-pixel heads. The values of the remaining three dimensions, which did not vary between members of a stimulus pair, were randomly assigned. They were approximately equally distributed across the possible range of values.

Results and Discussion

To establish that the three primary dimensions are of roughly equal discriminability and that they are all more discriminable than the subtle dimension, a single-factor repeated measures ANOVA was run, using change type (diameter, angle, color, bumpiness, values and same). Results, depicted in Figure 2, showed a significant main effect of change type, $F(5,120)=104.80$, $p<.0001$. Scheffe post-hoc tests revealed that color was not significantly different from either angle or diameter, but that performance on differences in angle were detected significantly more often than differences in diameter, showing that these three dimensions are roughly, but not perfectly equal. The

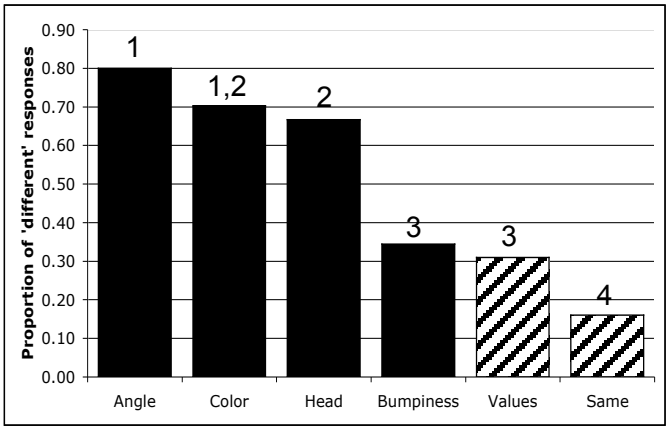


Figure 2: Data from Experiment 1. Groups with different numbers above the bar are significantly different from one another. The striped bars indicate trial blocks where a 'different' was incorrect.

proportions of 'different' responses for each of these three dimensions were significantly higher than for 'same' trials, establishing that participants have the ability to detect these changes. Performance on these three dimensions was also significantly better than on bumpiness, thus, supporting its classification as a subtle dimension. Despite worse performance on the subtle dimension, participants responded 'different' significantly more frequently on trials with changes on this dimension, than on 'same' trials. Finally, performance on 'bumpiness' trials did not differ significantly from performance on 'values' trials, where changes in the way the stimuli were drawn preserved the values of the four key dimensions. Because participants were instructed to ignore the kinds of changes that occur on 'values' trials, they were likely responding without being aware of what kind of changes they saw. The equivalence of performance on these two trial types suggests that participants may be responding with an equivalent lack of awareness on 'bumpiness' trials. This further supports the classification of the bumpiness dimension as subtle.

Experiment 2

Experiment 2 used the subtle dimension of tail bumpiness. This procedure makes this experiment an apt analog to real world categories that experts must master. The effect of pointing out the subtle dimension to participants was investigated using three between-subjects conditions: Help, No-Help (NH) and No-Help/Help (NH-H). In the Help condition participants were told in the instructions at the beginning of the experiment that tail bumpiness is an important cue in helping to classify the stimuli correctly. In the NH condition, participants were never told about tail bumpiness. In the NH-H condition, participants were given this information at the beginning of Stage 2.

Method

Participants Participants were 83 Arizona State University undergraduates enrolled in an introductory psychology course. They participated for course credit. They were randomly assigned to one of three conditions: Help (n=33), NH (n=24), NH-H (n=26).

Stimuli The stimuli had three principal dimensions of variation: head size, tail angle and tail bumpiness as described in Experiment 1.

Structure The two categories used were linearly separable only in three dimensions; therefore the best lower dimensional bounds always had exceptions. The best two-dimensional linear decision boundary left 14% of the exemplars as exceptions and the best single dimension linear decision boundary left 30% as exceptions. Figure 3 shows the training stimuli plotted in the angle/diameter space and also bumpiness/angle/diameter space.

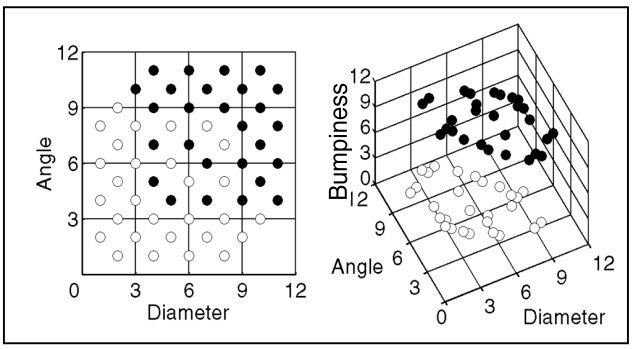


Figure 3: The stimulus space used in Experiment 2, plotted for stimuli with and without the subtle bumpiness dimension.

Procedure The experiment had two stages with each stage consisting of a learning phase and a transfer phase. The learning phase included four blocks of 56 trials. Within a trial block every stimulus in the learning set was presented once. The presentation order was randomized for each trial block, and for each participant. On each trial, the participant was shown a stimulus and asked to classify it as 'normal' or 'deviant'. Once the participants indicated their choice by pressing the appropriate key ('n' or 'd') on the keyboard, the correct answer was presented next to the stimulus for 2000 msec. If the participant answered incorrectly, the feedback was red instead of black. After the learning phase, participants performed a transfer task in which they classified two cycles of 16 novel stimuli. The transfer set presented after the three-dimensional stimulus set consisted of 8 stimuli on either side of the best three-dimensional linear decision boundary at a range of values. These values were such that if participants were using the best combination of angle and diameter dimensions, or any one of the three dimensions alone, they would achieve 50% correct. Participants were not given feedback during the transfer task. Stage 2 was a repetition of Stage 1, except for the instructions, as determined by the condition. For all conditions, the instructions encouraged participants to achieve perfect classification for both Stage 1 and Stage 2.

Results and Discussion

Overall performance was equally good whether or not participants were told about the usefulness of the subtle dimension (Help, M=87%; NH, M=88%; NH-H, M=88%). The dimension of primary interest is tail bumpiness. Single group t-tests against zero revealed that NH-H condition showed significant increase in the use of tail bumpiness in the second stage, $t(25)=2.34$, $p<.05$, but the other conditions showed no significant change in tail bumpiness from Stage 1 to Stage 2. The instructions clearly increased the use of tail bumpiness. The NH group had the lowest use of tail bumpiness, with the NH-H group significantly increasing their use of tail bumpiness, after being instructed to do so.

Table 1: Percentages of participants for which each cue set accounts for most of the transfer test responses from Experiment 2.

Condition	Stage	Bump	Angle	Diam	D/A	B/D/A
NH	1	36%	11%	18%	29%	7%
NH-H	1	29%	14%	14%	32%	11%
H	1	43%	14%	6%	26%	11%
NH	2	16%	32%	13%	26%	13%
NH-H	2	56%	11%	11%	4%	19%
H	2	32%	22%	7%	15%	24%

Individual differences in the adoption of the various cue sets were assessed by the transfer test. Participants were sorted according to the cue set that best matched their responses. Because all dimensions were present in the transfer stimuli, it was possible for a participant's responses to fit two different cue sets equally well. In such cases both cue sets were counted and the percentage reported was calculated across all preferences not all participants. That said, the large majority of participants preferred only one cue set. The percentages of preferences for each cue set are reported in Table 1.

Overall, many participants seemed to make use of the subtle dimension of tail bumpiness early in training; even without having it brought to their attention. Use of the subtle dimension was increased by instructions however, and though participants in the NH condition showed no performance deficit by the end of Stage 2, the transfer test reveals that they used tail bumpiness least. The general findings of this experiment mirror the Blair & Homa (2003b) studies using an obvious 3rd dimension, instead of the subtle one, namely some participants incorporated the 3rd dimension while others stuck with one or two sub-optimal dimensions. One question still open is whether or not more participants would adopt a cue set which separates the categories if it had fewer dimensions. This question is addressed in Experiment 3.

Experiment 3

In Experiment 2, it was shown that many participants rapidly detected and used diagnostic information, even though it was subtle. Some participants can readily use a three-dimensional cue set if the categories are made separable. For Experiment 3, a new stimulus space was created by changing the values on the subtle dimension. Like the space in Experiment 2, this space was not linearly separable using the two regular dimensions (head diameter and tail angle) but was separable by also considering the subtle dimension (tail bumpiness). Unlike the space for Experiment 2, this space was also separable when considering only head diameter and tail bumpiness. The separable 2-D and 3-D spaces are shown in Figure 4. These categories allow participants to collapse their cue set to only two dimensions with no loss of accuracy. The primary objective of this experiment was to assess the degree to which participants are able find the most efficient cue set if it required using fewer dimensions.

The instruction manipulation was dropped for this experiment; all participants were in the equivalent of the NH condition. Also, individual differences in this experiment were expected to be more important because there were multiple effective strategies. Accordingly, a larger number of participants were tested.

Method

Participants Participants were 96 Arizona State University undergraduates enrolled in an introductory psychology course. They participated in the experiment to fulfill a course requirement.

Structure In order to detect the use of the various possible 2-D spaces during transfer, the stimulus values for the training and transfer stimulus spaces were altered from Experiment 2. As in the Experiment 2 space, individually, diameter and angle were 70% predictive. The bumpiness dimension was 73% predictive. In 2-D space (head diameter and tail angle) this stimulus space is identical to the Experiment 2 space, that is, 8 of the 56 stimuli (14%) were exceptions. In 3-D space, the two categories were linearly separable. Also, in the 2-D space defined by head diameter and tail bumpiness the categories were linearly separable. The two spaces where the categories are linearly separable are shown in Figure 4. The category structure (mean within-category distance divided by mean between-category distance) was .60 for the categories represented in the diameter/bumpiness space and .66 for the categories when represented in the 3-D space.

Procedure The procedure was identical to Experiment 2 except the transfer tasks, which involved 2 cycles through a transfer set with 12 stimuli.

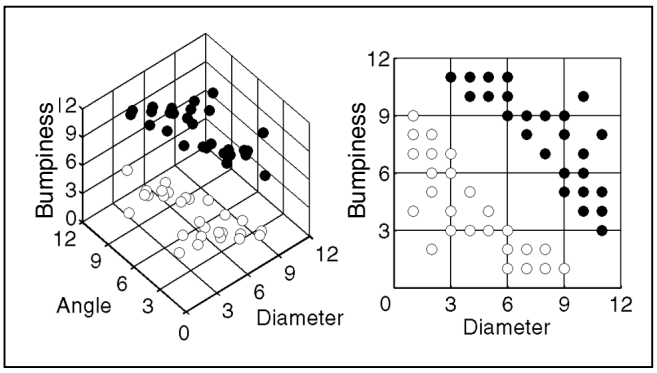


Figure 4: This stimulus space used for Experiment 3. This space was modified from the space for Experiment 2 so that the categories are separable in both the B/A/D space (plotted on the left) and the B/D space (plotted on the right).

Results and Discussion

Overall, performance improved across trial blocks and participants averaged 83% correct on the final trial block. In the transfer task, the measures of the three single dimension cues (bumpiness, angle and head diameter) and three two dimension cue sets (bumpiness/angle, bumpiness/diameter and diameter/angle) are not independent and require some explanation. Dimension use, as in Experiment 2, is measured as the proportion of trials that would be supported by that dimension or pair of dimensions. For each of the six cues, 8 of the 12 transfer trials present a stimulus on which the values of the cues suggest one category over another, and 4 of the 12 transfer trials present a stimulus for which the cue is neutral. Each single dimension and its dual-dimension opposite (e.g., diameter and bumpiness/angle) use the identical eight trials, and so are perfectly negatively correlated. This means that if a participant chose according to the value of tail bumpiness on .88 proportion of the trials, then they must also have chosen according to head/angle on the remaining .12 proportion of the trials. The other cues are only partially correlated. The overall attention use is forced to sum to 1.5 for the three single dimension cues and to 3.0 for all six dimensions. As an example, a participant might have a balanced attentional profile, equally distributing attention to all dimensions in combination (the 3-D solution). This profile would conform to each cue set on .50 proportion of the trials. A different participant might not use the subtle tail bumpiness cue at all. This profile would yield a 1.0 proportion for D/A (and 0 for bumpiness) and, assuming equal weight to the remaining dimensions, .25 for B/A and B/D (and therefore .75 for diameter and angle individually).

In the first transfer task, the use of the head dimension is consistent with the highest proportion of trials (.63) with D/A (.61) and B/D (.52) the next two highest. The emphasis on these three dimensions increased after the second stage of training (.70, .62 and .58 respectively). Both the increase in head and the increase in B/D (or, alternatively, the decrease in use of angle) were significantly different than chance, $t(95)=2.89$, $p<.01$ and $t(95)=3.20$, $p<.01$. Participants seem to increasingly ignore the information given by angle in favor of reliance on head diameter, even though angle provides easier discriminations (Experiment 1).

As in Experiment 2, a table summarizing individual data was created. The number of participants for which a cue or cue set received the highest use during the final transfer test was recorded. Twelve participants had equal endorsement for two different cue sets. In these cases, each cue was given credit, so the total number of endorsements is 108, even though there were only 96 participants. The percentage of the total number of preferences is shown for each cue in Table 2.

Table 2: Percentages of participants for which each cue set accounts for most of the transfer test responses from Experiment 3.

Stage	Bump	Angle	Diam	D/A	B/D	B/A	B/D/A
1	6%	16%	29%	24%	17%	4%	5%
2	1%	6%	38%	23%	23%	1%	8%

This tabulation reveals that 30% of the participants used a space in which the categories were separable, with nearly three times as many preferring the simpler B/D space to the B/D/A space by the end of training. This more than the 13% in the equivalent condition from Experiment 2. The reduction in angle use that is equivalent to the adoption of the B/D cue set is made more dramatic in that the angle dimension yielded the most accurate discriminations in Experiment 1. Further, both tail angle and tail bumpiness are features of the same component of the stimulus. On the other hand, there were still two thirds of the participants who used only obvious dimensions and failed to adopt a cue set which separates the categories. The preference for simplicity may extend to the discriminability of the cues as well as their number.

General Discussion

The experiments reported here investigated participants' ability to use subtle perceptual cues to disambiguate overlapping categories. In Experiment 1, it was established that the primary dimensions of variation, head diameter and tail angle, were roughly equated for discriminability, and the subtle dimension of tail bumpiness was much less discriminable. Experiment 2 demonstrated that some, but not all participants were able to use the subtle dimension in conjunction with the other dimensions to eliminate exceptions to the obvious dimensions. Experiment 3 showed that participants favored using a 2-cue optimal cue set over a 3-cue one.

There are two aspects of the present data that are challenging from a modeling perspective. The first is the widespread disregard for diagnostic cues. In some participants, this shows up as an inability to use the bumpiness cue or perhaps even a total reliance on just one obvious cue. These expedient cue sets could not yield enough information to result in perfect performance without additional memorization of exceptions. Even for participants who chose one of the cue sets that separated the categories, there was a strong preference for the simpler (2-d) set. Use of the angle dimension, which was part of the optimal 3d cue set, and which was independently diagnostic, decreased with training. The second aspect of the data that is challenging is the broad individual differences. Nearly every possible cue set was used in all phases of the reported experiments. Category learning models which adjust attention weights based on gradient descent on error (i.e., the backprop algorithm) will tend to converge on a set of optimal weights, rather than showing dramatic differences in predicted performance (see discussion in Kruschke, 2001). These optimal weights will also be positive for any informative dimensions. More recent models of attention in associative learning that incorporate rapid shifts of attention

in conjunction with annealed learning rates show more promise of fitting our data. The rapid shifts of attention can lead to shifts away from dimensions before associations can form, and the annealed learning results in a progressive discounting of error, leading to participants getting frozen in a sub-optimal space. Kruschke and Johansen's (1999) RASHNL model has fit similar data in probabilistic category learning task.

The present study is related to recent work on human concept learning that is based on a simplicity principle (Feldman, 2003). This principle suggests that the ease with which categories can be learned is related to how incompressible or complex the categories are, that is, the length of its minimum description. Maximally complex categories, for example one consisting of a mailman, a speedboat, and a jelly doughnut, have no regularities at all, and the minimum description is simply a list of its members. Simpler categories, for example: big blue triangle, big red circle, and big yellow square, can be compressed down to exclude some of the data from the examples, leaving a smaller description, in this case "big" things. In the context of the current work, it could be said that incorporating the information from the bumpiness dimension, while increasing the number of dimensions, decreases the overall category description length, because the exceptions do not have to be explicitly encoded. Participants in Experiment 2, some of who showed increasing use of the bumpiness dimension, but also decreasing use of the angle dimension, also seemed to use a simplicity principle. Angle use decreased not because it was uninformative, but because it was not part of the minimal description. In addition to reworking their description of their overall category regularities, participants can and do augment their category representation by identifying and memorizing specific exception stimuli. In other words, simply adding any exceptions to the representation they have already formed. Several results, including the present data, suggest that this strategy is not uncommon in some typical category learning paradigms (Blair & Homa, 2001). This focus on individuals rather than on category level regularities seems to occur even in some separable categories if they are small and weakly structured (Blair & Homa, 2003). These results highlight the potential disparity between mathematical complexity and psychological complexity, and emphasize the importance of understanding how the cognitive system implements complexity minimization. A precise account of the relative costs of adding or shifting dimensions versus memorizing exceptions will certainly involve a better understanding of how attentional, perceptual, and memorial processes interact as classification expertise develops.

Acknowledgments

This work was completed in partial fulfillment of the Ph.D. requirements in psychology at Arizona State University. I would like to thank Steve Goldinger, Mike McBeath and Sue Somerville who comprised my excellent dissertation committee, and especially Don Homa, my mentor and committee chair. I would also like to thank Jacqueline Blair for helpful comments on an earlier draft of this paper and Chris Summers for help in conducting these experiments.

References

Blair, M. & Homa, D. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition 29*. 1153-1164.

Blair, M. & Homa, D. (2003). As easy to memorize as they are to categorize: The 5-4 categories and the category advantage. *Memory & Cognition 31*. 1293-1301.

Blair, M. & Homa, D. (2003b). Integrating Novel Dimensions to Eliminate Category Exceptions: When More Is Less. *Manuscript submitted for publication.*

Blair, M. & McBeath, M. K. (2002). Bangers in billiards - expert/novice differences in shot force. *Manuscript in preparation.*

Boster, J. S., & Johnson, J. C. (1989). Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist, 91*(4), 866-889.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152.

Davies, S. P. (1994). Knowledge restructuring and the acquisition of programming expertise. *International Journal of Human Computer Studies, 40*(4), 703-726.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, 12, 227-232.

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology-General, 126*(3), 248-277.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*(6), 812-863.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1083-1119.

Lewandowsky, S., Kalish, M., & Griffiths, T. L. (2000). Competing strategies in categorization: Expediency and resistance to knowledge restructuring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1666-1684.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*(1), 49-96.

Solomon, G. E. A. (1997). Conceptual change and wine expertise. *Journal of the Learning Sciences, 6*(1), 41-60.

The Effect of Temporal Delay on the Interpretation of Probability

Amber N. Bloomfield (a-bloomfield@northwestern.edu)
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208 USA

Abstract

The studies reported here investigate the interaction between probability and delay. In the first study, the fits of a range of high and low probability words were calculated for numerical probabilities presented with either a short or long delay. Results show that participants in the long delay condition felt that high probability words fit small numerical probabilities better and that low probability words fit large numerical probabilities better than did participants in the short delay condition. In a second study, participants were presented with money offers that were both delayed and risky. Findings indicate that delay is given less weight at low probabilities, and probability is given less weight at large delays when probabilities are mid-range. Combined, these data suggest that a trade-off occurs between giving attention to delay and giving attention to probability in judgments. One component of this arises from long time delays "dampening down" the influence of probability level, but the complete nature of the interaction between probability and delay remains to be explored.

Introduction

In everyday decision making, individuals must determine the value that various outcomes have for them. Often, even if people have a clear idea of the value that an outcome has for them in general (such as a week in Paris), they must assess its value in terms of different types of uncertainty associated with the outcome. One type of uncertainty arises from the outcome having a less than 100% likelihood of occurring (i.e., it is probabilistic). This type of uncertainty is normatively applied by translating an outcome into its *expected value* (EV): multiplying the value of the outcome by its probability of occurring. Another type of uncertainty associated with outcomes is temporal delay. Adjustment of an outcome due to temporal delay is referred to as *temporal discounting.*

Choice involving risk (i.e., probabilistic outcomes) and intertemporal choice have several parallel anomalies (Prelec & Loewenstein, 1991). These anomalies include common difference and ratio effects, immediacy and certainty effects, magnitude effects and sign effects. Common difference effect occurs when a pair of delayed outcomes which an individual is indifferent between produce a decisive preference for him or her when a common delay is added to both. For instance, a person might be indifferent between $25 now and $40 is one week, but may express a preference for the $40 if a one-week delay is added to both options. Similarly, common ratio effect occurs when two probabilistic options which a person is indifferent between produce a solid preference when their probabilities are multiplied by a common probability. A person might be indifferent between a 5% chance of $10 and 2% chance of $15, but prefer the $15 option if both probabilities are multiplied by 50%. Immediacy and certainty effects involve the overweighting of immediate outcomes in intertemporal choice and the overweighting of certain outcomes in risky choice. Magnitude effects occur when large amounts are discounted to a lesser (temporal discounting) or greater (discounting for risk) degree than are small amounts. Sign effects involve a tendency towards risk-aversion for gains and risk-seeking for losses in risky choice, and a steeper discounting of gains than losses in intertemporal choice.

Given these parallels, it is not surprising that some researchers have suggested that discounting for risk and discounting for delay arise from the same source. For instance, Benzion, Rapoport and Yagil (1989) argue that, in addition to the time value of money (characterized as the accepted interest rate) delay introduces a *risk premium*, which arises from the implicit risk associated with delay. By this interpretation, the temporal discounting stems from implicit risk combined with the rational time value of money. Alternatively, Rachlin and Raineri (1992) argue that probability can be expressed as waiting time, by estimating the number of trials until a win (60% chance ≈ 6 out of 10 trials are wins), and adding together the amount of time between each trial preceding the first win to calculate overall waiting time. These two ideas both argue for a fundamental source of uncertainty (either risk or delay) that leads to both types of discounting.

It is likely that because of the focus on equating probability and delay little attention has been paid to how they affect each other, both in determining the value of outcomes and directly. Only a few studies have actually presented participants with outcomes that are both delayed and probabilistic. Keren and Roelofsma (1995) argue for two different types of uncertainty present in intertemporal choice: internal (involving doubts about one's ability to predict future tastes/needs) and external (concerning doubts about whether promised future payments will be honored). Internal uncertainty is the type of uncertainty typically associated with intertemporal choice. External uncertainty is probabilistic uncertainty, which they argue is also a component of any temporal delay. They found that the immediacy effect could be derailed by making the options probabilistic (the immediate option was no longer over-weighted), and that adding a delay to a certain option weakened the certainty effect (the certain option was no longer as over-weighted, and more people preferred a risky

option that had a higher payoff). Keren and Roelofsma did not find a significant interaction of delay and probability, and suggested that the two factors are additive. However, they used only two levels of probability in their experiment examining the immediacy effect, and only one level of temporal delay when examining the certainty effect.

Keren and Roelofsma's (1995) description of the external uncertainty component of delay does suggest that the subjective interpretation of a given probability when a delay is introduced should be lower, because the uncertainty associated with a delay should make the relevant outcome seem even less likely to be received. While Keren and Roelofsma only used delay and probability to make qualitative departures from certainty and immediacy, respectively, a more continuous effect of delay on probability should be evident if external uncertainty increases with delay.

In the following studies, the effect of delay on the interpretation of probability is explored. In the first study, a range of probabilities are paired with one of two levels of delay for all participants, and 10 probability words are rated as to their fit of each of the numerical probabilities. In the second study, multiple levels of probability and delay are combined to examine their effects on the value of two different monetary outcomes.

Study 1

In Study 1, a direct method of examining the influence of delay on probability was employed. This design was based on past work by Budescu, Karelitz and Wallsten (2003) examining how numerical probabilities are mapped on to linguistic probability words/phrases. The method of presentation was reversed, so that participants were asked to rate degree of fit of 10 probability words for each of 10 numerical probabilities. It was predicted that, if there is an external uncertainty component of delay, this would lead to numerical probabilities presented with the longer temporal delay to elicit higher fit ratings for the low probability words and lower fit ratings for the high probability words.

Methods

Materials

Instructions Participants were asked to respond to each numerical probability item by rating each probability word for that numerical probability on a scale from 1 to 8, with 1 indicating that the word fit the numerical probability "not at all" and 8 indicating that the word "absolutely" fit the numerical probability.

Stimuli Participants were randomly assigned to receive all numerical probability statements with either a short (6 months) or long (3 years) delay. Participants were presented with 10 numerical probabilities (5% - 95% in steps of 10%) embedded in the following statement: "If someone told you 'you have a _% chance of winning $9,864 in 6 months/3 years,' to what degree do you feel each of the following words fits the probability this person stated?" For each statement, participants rated 10 probability words on the basis of their fit. The probability words, (from lowest to highest probability-mapping, according to Budescu, et al., 2003) were: Impossible, Improbable, Unlikely, Doubtful, Toss-up, Possible, Probable, Good chance, Likely and Certain. The numerical probabilities were presented in a different randomized order for each participant.

Procedure Participants received the instructions for the task and responded to the test items via computer. During the task, participants were presented with each numerical probability statement followed by the 10 probability words. Participants typed in their rating for each word on the keyboard. After responding to all 10 numerical probability statements, participants were presented with the debriefing.

Participants Participants were 32 Northwestern undergraduates who participated to fulfill partial course requirement (17 in the 6 month delay condition, 15 in the 3 year delay condition).

Results

The mean rating of each probability word for each numerical probability in the two conditions was translated into proportion of total fit (by dividing the mean score by eight). These proportions were then collapsed across the three low probability words, not including impossibility (Improbable, Unlikely, and Doubtful) and the three high probability words, not including certainty (Probable, Good chance and Likely) to create a composite Overall-Low and Overall-High fit for each numerical probability.

A regression performed on the Overall-Low composite fits revealed a significant effect of probability ($t(19) = -24.94$, $B = -.979$, $p<.001$), a marginally significant effect of condition ($t(19) = 2.05$, $B = .080$, $p = .058$) and a significant interaction between probability and condition ($t(19) = 2.51$, $B = .099$, $p<.05$). Figure 1 displays the Overall-Low fits for the 6 month and 3 year conditions across numerical probabilities.

Overall-Low fits decreased as probability increased (as

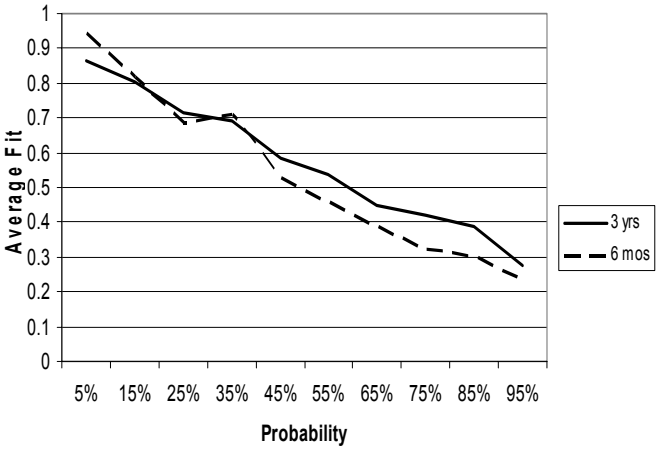


Figure 1: Overall-Low Proportions Across Probabilities

would be expected), and were higher for the 3-year delay condition (as was predicted). However, the interaction between condition and probability reveals a more complicated picture: While participants in the 3 year delay condition have higher Overall-Low fits for the higher numerical probabilities than do participants in the 6 month delay condition, there is also a tendency for these participants to rate the fit of low probability words as *lower* for the smaller numerical probabilities.

A regression performed on the Overall-High composite fits revealed a significant effect of probability ($t(19) = 25.47$, $B = .981$, $p<.001$) and a significant interaction between probability and condition ($t(19) = -2.90$, $B = -.111$, $p<.05$). Figure 2 presents the Overall-High fits across probabilities for both conditions. For the Overall-High ratings, the nature of the interaction between condition and probability is more pronounced. Participants in the 3-year condition provided lower fit ratings of the high probability words for the higher numerical probabilities *and* higher fit ratings for the lower numerical probabilities.

The findings of Study 1 suggest that delay does not have a uniform effect on the interpretation of probabilities. Rather, the effect of delay is determined by both the level of the numerical probability and the "level" of the probability word. The longer delay increases the fit of the low probability words to the numerical probabilities, as predicted, but only for those probabilities in the mid-range or higher. For smaller numerical probabilities, the longer delay *decreases* the participants' ratings of low probability words. Similarly, the longer delay decreases participants' ratings of high probability words for probabilities in the mid-range or higher, as predicted, but *increases* these ratings for the lower numerical probabilities. It seems that a longer delay decreases the "positive-ness" of the high probabilities but also the "negative-ness" of the low probabilities. At long delay, participants do not seem to uniformly interpret probabilities as lower, but do seem to uniformly interpret probabilities as less extreme.

Given the findings of Study 1, it is apparent that delay does influence how numerical probabilities are understood, and the direction of this effect depends on the level of the numerical probability. Because this interaction exists, it is important to look at how probability and delay combine in determining the value of outcomes that are both delayed and probabilistic. In Study 2, the effects of delay and probability, and the interaction between the two, on the value of monetary outcomes are investigated.

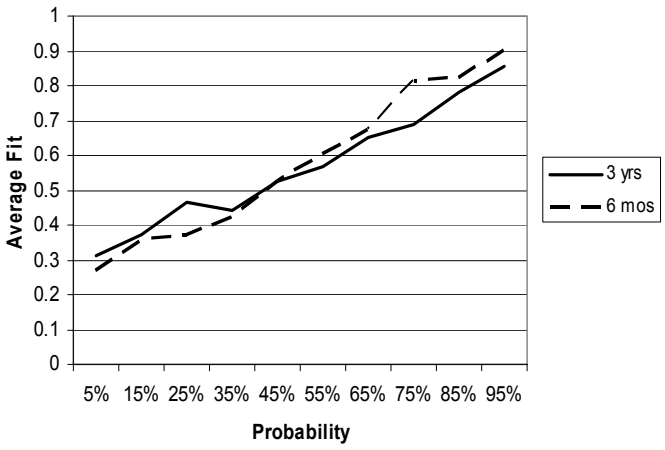


Figure 2: Overall-High Proportions Across Probabilities

Study 2

Methods

Materials

Instructions Participants were instructed to respond to each item by providing a *certainty equivalent* (CE) for the money offer. A certainty equivalent is an amount that will be received immediately and for certain, such that participants feel they would be indifferent between this amount and the presented money offer.

Stimuli Participants responded to a total of 70 money offers. For each of two payoff amounts ($10,000 and $1 million) participants responded to five gambles with one of five probabilities (5%, 30%, 55%, 80% or 95%), five delayed payments with one of five time delays (6 months, 1 year, 3 years, 5 years or 10 years) and 25 *delayed gambles*, combining each probability level with each delay level once. Gambles, delayed payments and delayed gambles for both payoffs were presented in a different random order for each participant.

Procedure Participants received the instructions for the task and responded to the test items via computer. During the task, participants were presented with each money offer and typed in a certainty equivalent using the keyboard. Participants were not given any feedback on their performance. At the end of the 70 money offers, participants were read debriefing information by the experimenter.

Participants Participants were 18 Northwestern undergraduates who participated to fulfill partial course requirement. The data from two additional participants was excluded due to a failure to follow directions (more than ¼ of their responses were greater than the payoff of the money offer or they responded "0" to one or more items).

Results

Because several participants had one or two responses that were greater than the payoff amount (errors occurring from accidentally typing an extra "0" into the computer), all analyses were performed on median responses as opposed to mean responses.

Weighting and Discounting Factors Using the responses to the delayed payments, it was possible to calculate the temporal discounting factor (k) for each participant. This factor represents the extent to which, for each day of delay, an individual devalued the payoff amount. The formula for k

is derived from the formula for temporal discounting developed by Mazur (1987): k = (V/dv)-(1/d), where V is the undiscounted value of the outcome, d is the delay in days and v is the provided subjective value of the outcome. A k of 0 implies no discounting of the outcome for delay. For each participant, the median k across all delayed payments was obtained.

There was a significant effect of amount on the median k values: participants tended to have larger ks for payoffs of $10,000 than for payoffs of $1 million ($t(17) = 2.85$, $p < .05$). The mean k for the $10,000 delayed payments was .0005, while that for the $1 million delayed payments was .0002. Larger ks imply greater temporal discounting, and greater discounting of smaller payoffs is consistent with the magnitude effect discussed in Prelec and Loewenstein (1991). Although the difference between the k factors for the two amounts seems quite small, such a difference would result in an 8% decrease in value for $10,000 delayed by 6 months compared to only a 4% decrease for $1 million delayed by 6 months.

Using the responses to the gambles, the probability weighting factor (h) was calculated for each participant. This factor represents the extent to which an individual's weighting of probabilities in their responses corresponds to expected value (h = 1 means responses are perfectly in line with expected value). An h greater than 1 demonstrates risk-aversion (the certainty equivalent is less than the expected value of the gamble), while an h between 0 and 1 shows risk-seeking (the CE is more than the gamble's expected value). The formula used to calculate h was derived from the probability weighting formula provided by Rachlin and Raineri (1991): h = pV/v(1-p) – p/(1-p). Here, p is the probability of acquiring the outcome amount V, and v is the provided subjective value of the outcome. For each participant, a median h was obtained.

There was no significant effect of amount on h, ($t(17) = 1.51$, $p > .05$). The mean h for $10,000 gambles was 1.66, and 5.62 for the $1 million gambles. The large difference is due to one participants' extremely risk-averse responses for the $1 million gambles (median h for $10,000 = 1; for $1 million, median h = 1.11). This is consistent with Green, Myerson and Ostaszewski (1999), who found that magnitude effects in probability discounting are often small or non-existent. However, it is worth mentioning that, of the 14 participants that had different probability weighting factors for the $10,000 and $1 million gambles, 10 had larger h values for the $1 million gambles, which is consistent with the discussion of Prelec and Loewenstein (1991).

Overall Analyses for Delayed Gambles Participants' certainty equivalents were transformed to proportion of payoff amount (e.g., $5000 for a $10,000 payoff was .50) for overall data analyses. Again, the median rather than the mean of these proportions was used for analyses to control for extreme responses. A regression with amount, probability and delay as predictors revealed significant effects of probability ($t(49) = 23.15$, $B = .95$, $p<.001$) and delay ($t(49) = -2.97$, $B = -.12$, $p<.01$), and a marginally significant interaction between probability and delay ($t(49) = -1.88$, $B = -.08$, $p = .068$), on the median proportion CE. The effect of probability on response was as expected (greater CEs provided for larger probabilities), although participants over-weighted 5% to a far greater degree than has been found in past studies. Delay also had the predicted effect, with smaller CEs provided for larger delays. Figure 3 displays the overall findings of proportion CE for the five delays and the five probabilities. Because amount had no significant influence on proportion CEs, they are collapsed across amount.

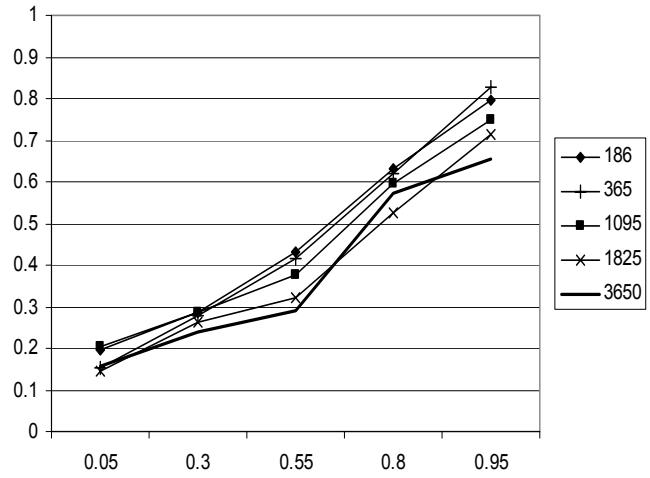


Figure 3: Proportion Certainty Equivalents by Delay and Probability

The delay by probability interaction is most apparent for the two longest delays (5 and 10 years). Participants show little increase in proportion CE with the increase from 30% to 55% probability for delayed gambles at a 5 or 10 year delay. In addition, participants show little sensitivity to delay at the two smallest probabilities.

These findings suggest an interaction between probability and delay of the type found in Study 1. Probability appears to be given less weight at longer delays. While the lines representing gambles at 5 and 10 year delays show no increase in CEs between 30% and 55%, the lines for gambles at 6 month, 1 year and 3 year delays show a relatively constant increase with probability. The influence of delay on probability is not evident until delay reaches high levels.

The findings of Study 2 are not so one-sided, however. In deciding the value of delayed and probabilistic outcomes, participants seem to be dividing their attention between probability and delay. It is not simply that probability is given less weight when longer delays are present: probability levels are given less weight at long delays when probabilities are *small*. Further, as is evident from examining the points at 5% and 30%, little attention is given

to delay at the two smallest probability levels (it is not until a 55% probability of winning that participants really begin to differentiate between the delay levels in their responses). This interpretation does not necessarily require that long delays induce smaller subjective probabilities. Rather, in the influence of delay on probability, at long delays participants could simply be using probability less in determining their responses. This idea of trading off attention between option components is supported by analyses performed for the $10,000 and $1 million gambles separately: delay was not a significant predictor for the $1 million gambles ($t(24) = -1.61$, $B = -.09$, $p > .10$, but was significant for the $10,000 gambles ($t(24) = -2.67$, $B = -.15$, $p < .05$). This could indicate that, when outcome amount was large, participants devoted attention to the amount and the probability level in their figuring of a certainty equivalent, leaving no attention for delay.

The present interaction demonstrates that probability is weighted less when a long delay is associated with the gamble, and that delay is weighted less when probability of winning is low. Study 2 provides evidence that probability and delay influence each other in determining the value of payoffs.

Discussion

Study 1's findings indicate that delay influences the interpretation of probability, such that low probabilities are interpreted as less *unlikely*, *improbable* and *doubtful*, and high probabilities are interpreted as less *probable*, representative of a "*good chance*," and *likely*. A large delay seems to take attention away from probability at both high and low levels, suppressing the negative-ness of the low probabilities, and the positive-ness of the high probabilities (in effect, "dampening " the impact of probability).

Study 2's findings support the idea that delay influences the interpretation of probability: Probability (when it is mid-range) is given less weight in participants' judgments when delays are long. However, the relationship between delay and probability seems to be more complicated: delay is also given little weight when probability is very small. This suggests that what may actually be going on in the evaluation of the delayed gambles is a tradeoff between the attention given to probability and the attention given to delay. Thus, when delay is very large, and probability is mid-range, probability is given less attention than at smaller delays. Conversely, when probability is small, delay is given less attention than when probability is higher. The finding that delay is not a significant predictor for $1 million gambles also suggests that when amount is very high, there may not be enough attention left over for delay to figure into participants' responses.

Whether or not the "dampening" effect was present in Study 2 is not completely clear. Although the interaction of delay and probability can be interpreted as less attention given to probability when delay is very long, the effect of probability level on the weight given to delay was unanticipated. Further, if the dampening effect of large delays on probability was present one would have expected to see less weight given to probability, not just for the mid-range probabilities, but for the higher probabilities as well. However, if the interaction of delay and probability arises largely from the influence of delay on probability, it would be extremely difficult to see in the data from Study 2. As was pointed out above, probability had a *much* greater impact on responses than either delay or the interaction of delay and probability. In fact, it was not unusual for participants to show no sensitivity to delay at all in responding to the delayed gambles, but rather simply respond in accordance with expected value.

Past studies have demonstrated that outcome amount and probability have a dominant/subordinate relationship, with outcome amount taking precedence (Liberman & Trope, 1998; Sagristano, Trope, & Liberman, 2002). Although Study 2 did not provide any way of looking at that particular relationship, its findings do suggest that a similar relationship may exist between probability and delay. Probability accounted for most of the variance across both outcome amounts, and was significant within both outcome amounts. Delay, on the other hand, accounted for a small amount of variance across amounts, and ceased to be a significant predictor when only the $1 million gambles were considered. This makes sense if payoff is more important to participants than probability, which is more important than delay. The possibility of a dominant/subordinate relationship between probability and delay should be more directly examined in future studies using a method similar to that used by Liberman and Trope (1998).

If probability is indeed more important to participants making decisions about delayed gambles than is delay, techniques to highlight delay could be used. In Study 2, for all delayed gambles, participants were given the probability information first. This could have decreased the role of delay in participants' responses. Further, payoff amounts were expressed as round numbers ($10,000 and $1 million) for which it would be relatively easy to calculate expected value. A current study is investigating the influence between delay and probability when items are counterbalanced as to which information is presented first, and when payoffs are not round (e.g., $10,135). It is hoped that this study will produce responses that are more sensitive to delay, and allow a clearer picture of the interaction between delay and probability.

Another question that remains unanswered is the manner in which probability level influences the interpretation of a given delay. Study 2's findings suggest that the effect of delay may be dampened by the presence of a small probability. If the external uncertainty portion of the temporal delay affects probability interpretations, perhaps the external uncertainty associated with risk changes the interpretation of delays, by highlighting the delay's external uncertainty. Although it is difficult to find a factor to pair with delay that will parallel the relationship between numerical probabilities and probability words, one could be constructed presenting linguistic descriptions of durations

(e.g., "brief time" or "very long wait") to examine the influence of small and large probability levels on the interpretation of delays associated with monetary outcomes. It is possible that for a very small probability people will not differentiate as much between different levels of delays as they do with a large probability.

A final area yet to be examined is that the two uncertainty components of delay (internal and external) mentioned in Keren and Roelofsma (1995) may be able to be separately manipulated. For instance, it is easy to imagine situations where a delay could imply greater external uncertainty (e.g., a promise from an unreliable source), but not necessarily greater internal uncertainty. Conversely, while a spring vacation in Cancun might seem very valuable to me now, I have good reason to believe that it will have less value for me when I am ten years older, though I have no reason to think that I am less likely to receive that trip in ten years as opposed to a trip to Spain. Looking at how delay influences the value of different outcomes which emphasize or increase its internal or external uncertainty component is necessary to fully understand why delay decreases value.

Conclusion

The present studies demonstrate that there is an effect of delay on the interpretation of probabilities and an interaction between delay and probability on the value of monetary outcomes. When delay is longer, probabilities are interpreted as less extreme, at both higher and lower levels. Further, mid-range probabilities are weighted less at longer delays when valuing monetary outcomes, and delay is weighted very little when probabilities are small, suggesting that attention is traded off between delay and probability, depending on the levels of each. Because choices in life often involve delays and likelihoods less that 100%, it is worthwhile to explore the way people combine these two types of uncertainty and, especially for temporal delay, to gain a better understanding of the roots of delay's influence on value.

References

Benzion, U., Rapoport, A., & Yagil, J. (1989). Discount rates inferred from decisions: An experimental study. *Management Science*, *35*(3), 270-284.

Budescu, D., Karelitz, T., & Wallsten, T. (2003). Predicting the directionality of probability words from their membership functions. *Journal of Behavioral Decision Making*, *16*, 159-180.

Green, L., Myerson, J., & Ostaszewski, P. (1999). Amount of reward has opposite effects on the discounting of delayed and probabilistic outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 418-427.

Keren, G., & Roelofsma, P. (1995). Immediacy and certainty in intertemporal choice. *Organizational Behavior and Human Decision Processes*, *63*(3), 287-297.

Liberman, N, & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology, 75,* 5-18.

Mazur, J.E. (1987). An adjusting procedure for studying delayed reinforcement. In M.L. Commons, J.E. Mazur, J.A. Nevin, and H. Rachlin (eds.), *Quantitative Analyses of Behavior: Vol. 5. The Effect of Delay and of Intervening Events on Reinforcement Value* (pp. 55-73). Hillsdale, N.J.: Erlbaum.

Prelec, D., & Loewenstein, G. (1991). Decision making over time and under uncertainty: A common approach. *Management Science*, *37*(7), 770-786.

Rachlin, H., & Raineri, A. (1992). Irrationality, impulsiveness, and selfishness as discount reversal effects. In G. Loewenstein and J. Elster (eds.), *Choice Over Time* (pp. 93-118). New York: Russell Sage Foundation.

Sagristano, M., Trope, Y., & Liberman, N. Time dependent gambling: Odds now, money later. *Journal of Experimental Psychology: General, 131*(3), 364-376.

Enhancing Simulation-Based Learning through Active External Integration of Representations

Daniel Bodemer (d.bodemer@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Str. 40
72072 Tuebingen, Germany

Abstract

Discovery learning with computer simulations is a demanding task for many learners. Frequently, even fostering systematic and goal-oriented learning behavior does not lead to better learning outcomes. This can be due to missing prerequisites such as the coherent mental integration of different types of representations comprised in the simulations and in the surrounding learning environment. Prior studies indicated that learning performances can be enhanced by encouraging learners to interactively and externally relate different static sources of information to each other before exploring dynamic and interactive visualizations. In an experimental study addressing the domain of mechanics it was largely confirmed that the active external integration of representations can improve simulation-based learning outcomes.

Introduction

Computer-based learning environments increasingly comprise simulations in terms of dynamic and interactive visualizations to illustrate complex processes and abstract concepts. These simulations may be highly interactive in that they allow learners to change input variables by entering data or by manipulating visual objects and to observe the consequences of these changes in the dynamic visualizations as well as in additional representations such as numeric displays, formulas or text labels.

The conceptual model underlying the simulations has frequently to be inferred by the learners in processes of discovery learning, which correspond to the steps of scientific reasoning: defining a problem, stating a hypothesis about the problem, designing an experiment to test the hypothesis, carrying out the experiment and collecting data, evaluating the data, and (re-)formulate a hypothesis. The use of simulations frequently aims at inducing active learner behavior and constructive learning processes (e.g., de Jong & van Joolingen, 1998; Rieber, Tzeng & Tribble, in press). Learners have to self-regulate their learning behavior in order to discover the underlying conceptual model, which is assumed to lead to the acquisition of deeper domain knowledge (e.g., Schnotz, Boeckheler, & Grzondziel, 1999). However, it has shown that learners encounter difficulties in all phases of the discovery learning process. For example, learners have problems formulating useful hypotheses, designing appropriate experiments, and evaluating the output variables adequately (e.g., de Jong & van Joolingen, 1998; Njoo & de Jong, 1993; Reigeluth & Schwartz, 1989; Reimann, 1991). Moreover, many learners have difficulties in planning their experiments in a systematic and goal-oriented way and therefore interact with the simulations rather randomly (e.g., de Jong & van Joolingen, 1998; Schauble, Glaser, Raghavan, & Reiner, 1991).

Additional problems may be caused by the dynamic visualization of the simulated concepts. On the one hand the externalization of dynamic processes may prevent learners from performing cognitive processes relevant to learning on their own (e.g., Schnotz et al., 1999). On the other hand dynamic visualizations may overburden the learners' cognitive capabilities due to large amounts of continuously changing information, particularly if the output variables are represented as non-interactive animations that do not provide learners with the possibility to adjust the playback speed or to watch single frames (e.g., Lowe, 1999). In order to cope with these requirements, learners frequently make use of a strategy that limits their processing to selected aspects of a dynamic visualization, which are often not the most relevant aspects of the visualization, but rather those that are most perceptually compelling (cf. Lowe, 2003).

In order to support simulation-based discovery learning it has been suggested to structure the learners' interactions with the learning environment (e.g., van Joolingen & de Jong, 1991). Typically, these support methods guide learners to focus on specific variables of the underlying model, to generate hypotheses about relationships between these variables, to conduct experiments in order to test the hypotheses, and to evaluate the hypotheses in light of the observed results. Furthermore, various instructional support methods have been developed to facilitate specific processes of discovery learning, such as offering predefined hypotheses or providing experimentation hints (e.g., Leutner, 1993; Njoo & de Jong, 1993; Swaak, van Joolingen & de Jong, 1998). However, empirical results regarding these methods of instructional guidance are ambiguous (cf. de Jong & van Joolingen, 1998). Learners frequently did not make sufficient use of the instructional support to increase their learning outcomes.

One way to explain these findings is that learners lack prior knowledge necessary to benefit from complex visualizations. Learners who do not know enough about the domain of the visualized and simulated concept have problems processing complex dynamic visualizations and to interact with them in a goal-oriented way, even if they have enough information about useful learning behavior (cf. Leutner, 1993; Lowe, 1999; Schauble et al., 1991). Another reason – which is not independent from prior knowledge – is the difficulty of interconnecting multiple representations. Usually, simulations are embedded in multimedia learning

environments and presented in combination with symbolic external representations such as text and formulas. These different kinds of representations may complement each other, resulting in a more complete representation of the illustrated concept (e.g., Ainsworth, 1999; Larkin & Simon, 1987). Both Mayer (1997, 2001) in his theory of multimedia learning and Schnotz and Bannert (1999, 2003) in their integrative model of text and picture comprehension place emphasis on the importance of integrating textual and pictorial information into coherent mental representations during multimedia learning. However, learners are frequently not able to systematically relate multiple external representations to each other. As a consequence, these learners fail to integrate the different external representations into coherent mental representations, resulting in fragmentary and disjointed knowledge structures (e.g., Ainsworth, Bibby, & Wood, 2002; Seufert, 2003). Accordingly, to facilitate simulation-based learning it seems to be important not only to support learners in dealing with the dynamics and the interactivity of the simulations, but also to help them in relating the dynamically visualized information to corresponding information of other external representations.

To facilitate learning with multiple external representations it has been repeatedly suggested to present textual and pictorial information in a spatially integrated format instead of presenting them separately from each other in a "split-source" format (e.g., Chandler & Sweller, 1991, 1992; Mayer, 1997, 2001; Tarmizi & Sweller, 1988). According to cognitive load theory (Sweller, 1988; Sweller, van Merriënboer, & Paas, 1998) this can reduce unnecessary visual search resulting in a decrease of cognitive load and thus better learning. Another suggested method to support learners in making connections between different sources of information is to link the features of multiple representations by various symbolic conventions such as using the same color for corresponding entities in different representations (e.g., Kalyuga, Chandler, & Sweller, 1999; Kozma, 2003; Kozma, Russell, Jones, Marx, & Davis, 1996). While these instructional suggestions have the potential to reduce cognitive load, they do not directly support learners in constructing meaningful knowledge. Learners may nevertheless remain rather passive, concentrating on surface features of the visualizations and they may still be unable to mentally process and integrate the represented information in an adequate way (cf. Ploetzner, Bodemer & Feuerlein, 2001; Seufert, 2003).

Bodemer, Ploetzner, Feuerlein & Spada (in press) tried to initiate more active processes of coherence formation by encouraging learners to systematically and interactively integrate different multiple representations in the external environment. Learners were provided with spatially separated pictorial and symbolic representations on the screen and were asked to relate components of familiar representations to components of unfamiliar representations by dragging the symbolic represented elements and dropping them within the visualizations (see Figure 1).

This external process corresponds largely to the mental process of structure mapping as described by Gentner (1983; Gentner & Markman, 1997) and Schnotz and Bannert (1999). While (inter-)actively relating different sources of information is intended to directly support coherence formation, the simultaneous construction of an integrated format is supposed to gradually reduce unnecessary cognitive load (e.g., Chandler & Sweller, 1991, 1992). Bodemer et al. (in press) were able to demonstrate that – compared to the presentation of information in a pre-integrated or in a split-source format – learning outcomes can be improved significantly when learners actively integrate static information before interacting with dynamic visualizations

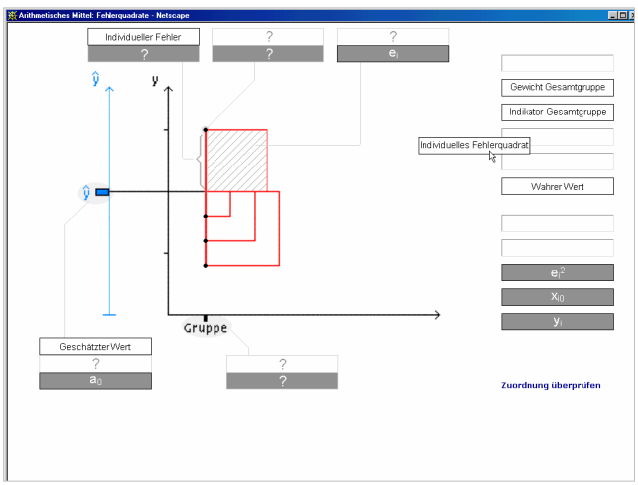


Figure 1: Active integration of information while learning statistics (cf. Bodemer et al., in press).

Bodemer et al. (in press) found the largest benefit of active integration when teaching extremely complex statistics concepts. In this paper an experimental study will be described which investigates possible benefits of active integration in another application domain with a slightly lower degree of complexity. It is hypothesized that also in less complex domains learners who actively integrate multiple representations will outperform those learning with a pre-integrated format. However, the advantage of active integration should rise with the degree of complexity of the learning material.

In order to avoid influences of assessment on the processes of discovery learning, Bodemer and his colleagues assessed the learning outcomes only after the learners had interacted with the dynamic visualizations. Thus they could not identify if knowledge has been acquired already during the process of active integration or afterwards during the process of discovery learning or both. In the study described below the learners' knowledge has been assessed both after integrating static representations and after interacting with dynamic visualizations. It is hypothesized that already the active integration of static representations can lead to better

learning outcomes. Additionally, learners who integrate multiple representations actively should improve comparatively more during simulation-based discovery learning.

Method

In this experimental study the participants learned various mechanics concepts in two consecutive learning phases. In the first learning phase they were provided with symbolic representations and static versions of dynamic and interactive visualizations. In the second learning phase they explored dynamic and interactive visualizations in a self-guided way.

Design

The experiment used a 2 x 2 factorial design with repeated measures on the second factor. The first factor addressed two levels of *information integration*, which was varied in learning phase 1: (1) presentation of the information in a pre-integrated format and (2) active integration of information. In the first condition the learners had to deal with visualizations that were already labeled while in the second condition the learners had to establish a relationship between the symbolic representation and the visualizations by dragging and dropping the symbolic representations onto the visualizations. The within-subjects factor was time of assessment: After the integration of multiple representations (test 1) and after the exploration of dynamic and interactive visualizations (test 2).

Participants

Forty-eight students (22 males and 26 females, aged 19 to 31) of the University of Tuebingen were randomly assigned to each of the two experimental conditions. They were paid for their participation. To prevent a high level of prior knowledge students of Mathematics and Physics were excluded as participants.

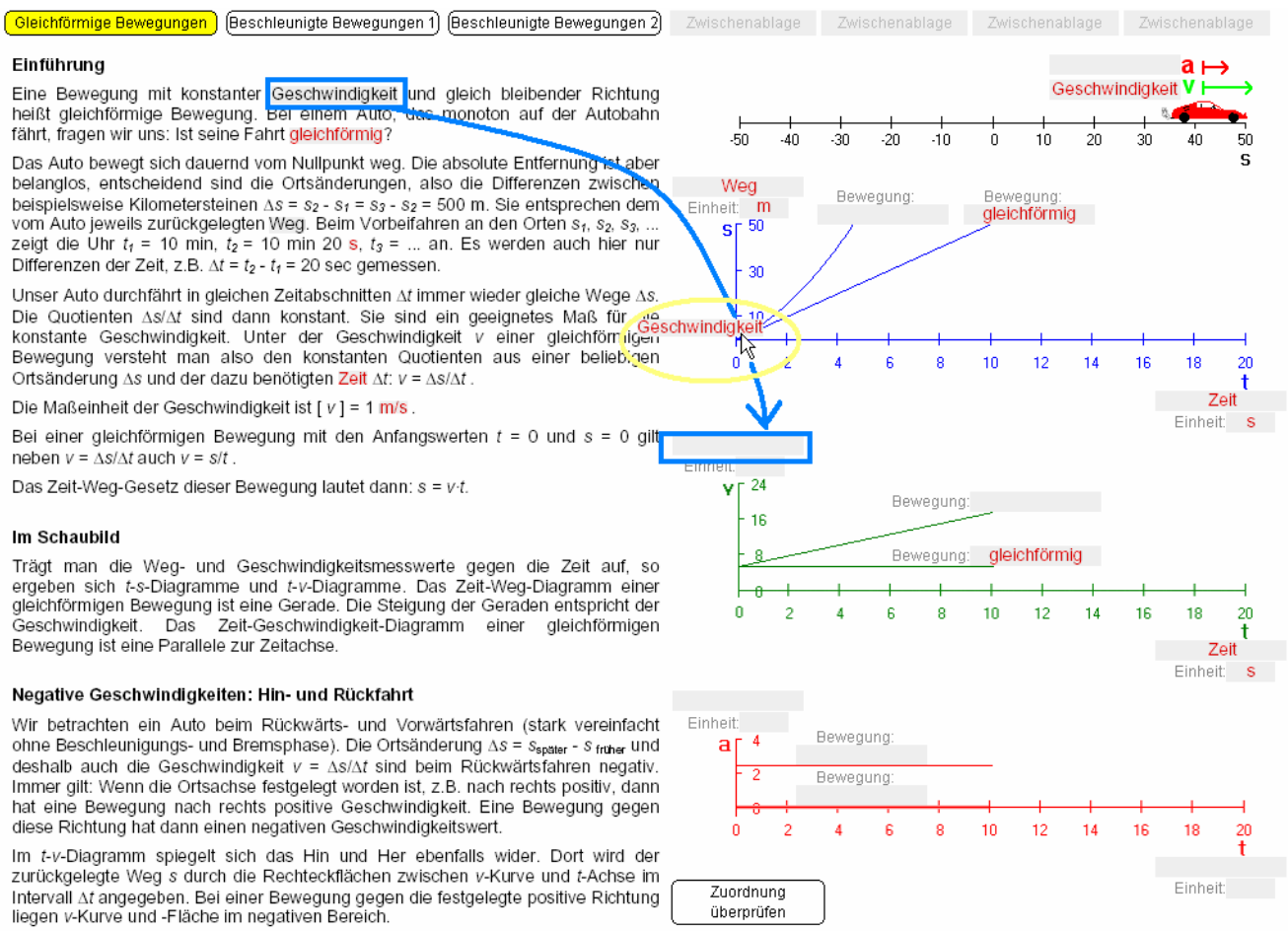


Figure 2: Active integration of information about mechanics concepts (learning phase 1).

Material

The *application domain* was comprised of various mechanics concepts, such as uniform and accelerated motion in one dimension. The *instructional material* consisted of two parts corresponding to the two learning phases:

(1) an instructional text accompanied by static visualizations, presented in the first learning phase on a computer (cf. Figure 2). The instructional text covered the left side of the screen and comprised three pages between which the learners could switch back and forth. The right half of the screen showed static versions of dynamic and interactive visualizations comprising the sketch of a moving car with corresponding velocity and acceleration vectors, a position-time graph, a velocity-time graph, and an acceleration-time graph. The presentation differed according to the two experimental groups of the first factor. In the group with *pre-integrated information* components of the visualizations were labeled with textual and algebraic information; whereas in the *active integration* group the learners interactively related the textual and algebraical information from the instructional text to the visualizations and thus created an integrated format on their own.

(2) dynamic and interactive visualizations, which were presented in the second learning phase (cf. Fig. 4). The visualizations were taken from the interactive learning environment PAKMA (Blaschke & Heuer, 2000). They correspond to the graphs of learning phase 1 with the addition that they could be modified by interactively changing variables and by running animated motion sequences.

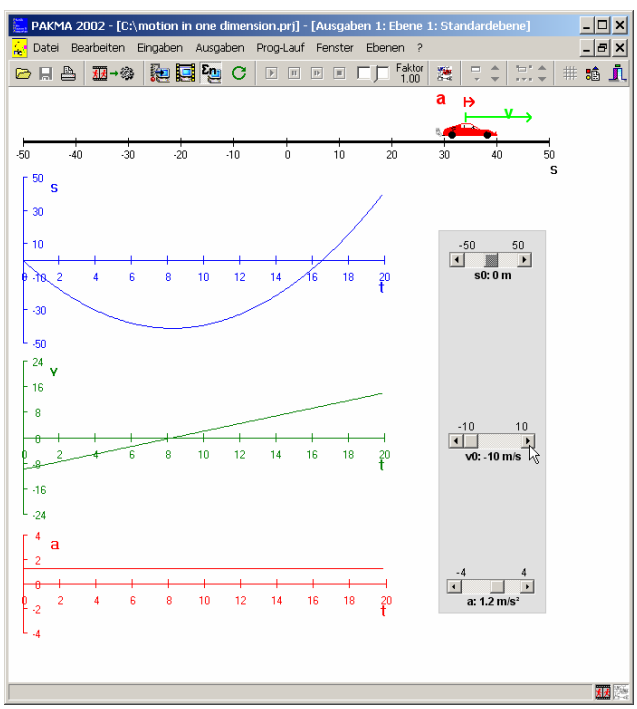


Figure 3: Dynamic simulation displaying motion in one dimension (learning phase 2).

The *test material* consisted of a knowledge test, given to the learners prior to the first learning phase, and two tests, which assessed the knowledge after each of the two learning phases. The tests were made up of different types of questions, which all required reasoning and transfer, and contained graphical elements in either the question or the answer or both: (1) questions which addressed transformations from textual to graphical representations, (2) questions which addressed transformations from graphical to textual representations, and (3) questions which addressed transformations within graphical representations. The pre-test and the first post-test consisted of six questions (two of each type); the second post-test consisted of 12 questions (four of each type). The participants' answers were scored by two independent raters.

Procedure

At the beginning of the experiment, all participants took the pre-test (20 minutes). Thereafter, learners of the condition *active integration of information* could train dragging and dropping of objects in a neutral domain (2 minutes). In learning phase 1 the participants were provided with the static versions of the dynamic and interactive visualizations accompanied by the instructional text (30 minutes). The information was either provided in a pre-integrated format or required learners to actively integrate it on their own. Then the learners took post-test 1 (20 minutes), followed by learning phase 2, in which the participants explored the dynamic and interactive visualizations without instructional guidance (15 minutes). Finally, the learners took post-test 2 (40 minutes). All participants had to spend the same time on the tasks.

Results

With regard to the pre-test there were no statistically significant differences between the groups for any of the test categories. The results of the post-tests are presented in the following. Table 1 shows the means and the standard deviations for the three types of questions: textual-graphical, graphical-textual, and graphical-graphical. Table 2 shows the results of a multivariate (Wilks-Lambda) and univariate two-way analyses of variance with repeated measures on the factor *time of assessment*.

The analyses of variance revealed a significant effect of *information integration* for those test questions which addressed transformations from graphical to textual representations. Learners with active integration performed better than with pre-integrated information in all categories of both tests; however, with regard to the two other types of questions the comparisons failed to reach statistical significance. The factor *time of assessment* had a significant effect on the test categories graphical-textual and graphical-graphical as well as across all types of questions. However, there were no interaction effects indicating that learners of both groups improved their knowledge during the exploration of the dynamic and interactive visualizations to approximately the same degree.

Additionally performed *t*-tests revealed that, on average, learners with active integration already achieved better learning outcomes after the first learning phase. Against the expectations, these differences between the groups slightly diminished in the second assessment after learning phase 2.

Table 1: Relative solution frequencies and standard deviations in both post-tests for the different questions.

Information integration		**textual-graphical**		**graphical-textual**		**graphical-graphic.**	
		Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Pre-integrated	M	.74	.71	.21	.40	.55	.66
	SD	.26	.28	.28	.26	.26	.28
Actively integrated	M	.84	.78	.37	.52	.67	.69
	SD	.24	.26	.26	.23	.29	.22
Overall	M	.79	.74	.29	.46	.61	.67
	SD	.25	.27	.28	.25	.28	.25

Table 2: The results of the multivariate and univariate two-way analyses of variance.

Source of variance	**Dependent variable**	**df**	**F**
	Between subjects		
Information integration	Across all types of questions	3, 44	1.48
	textual-graphical	1, 46	1.83
	graphical-textual	1, 46	4.32*
	graphical-graphical	1, 46	1.07
	Within subjects		
Time of assessment	Across all types of questions	3, 44	10.05**
	textual-graphical	1, 46	1.48
	graphical-textual	1, 46	27.91**
	graphical-graphical	1, 46	4.05*
Time of assessment x Information integration	Across all types of questions	3, 44	.56
	textual-graphical	1, 46	.13
	graphical-textual	1, 46	.39
	graphical-graphical	1, 46	1.51

Note: * $p < .05$, ** $p < .01$

Discussion

This paper investigated the benefit of an instructional support method to support learning with dynamic simulations in multimedia learning environments. Learners were encouraged to interactively and externally relate different static sources of information to each other before exploring dynamic simulations. In an experimental study the active integration of multiple representations was compared to the presentation of information in a pre-integrated format as suggested by Chandler and Sweller (1991, 1992) and Mayer (1997, 2001). The application domain was mechanics. It was hypothesized that learners who initially integrate multiple representations actively achieve better learning outcomes as found by Bodemer et al. (in press) for the domain of statistics.

The results largely confirmed that encouraging learners to actively integrate symbolic and static representations during multimedia learning can improve learning. Moreover, it shows that active integration of information – compared to the presentation of information in a pre-integrated format – can lead to the acquisition of knowledge already during learning with static symbolic and pictorial representations, and not only in combination with dynamic and interactive visualizations.

Contrary to expectations learners who actively integrated different representations were not able to improve comparatively more during simulation-based discovery learning. This may be due to the relatively low amount of additional information provided by the dynamic and interactive visualizations compared to their static versions. The static graphs already contained dynamic information by representing time on one axis. Ainsworth and van Labeke (2003) state that dynamic representations that express the relation between a variable and time do not contain more information than the same representation in a static form. Except for the illustration of the car with the corresponding velocity and acceleration vectors this applies to the dynamics of the simulation used in this study. However, the simulations contained additional information by providing the possibility to change variables interactively. But the number of changing options was very limited compared to the dynamic and interactive visualizations used by Bodemer et al. (in press).

The results differed with respect to the codalities of the test items. It appeared, that not only the retrieval cue codalities have to be considered (cf. Brünken, Steinbacher, Schnotz & Leutner, 2001); but also the codality of the learners' response effects the test result. Active integration of information was particularly helpful for answering questions that required transformations from graphical to textual representations.

Future research should consider the different codalities of test items as well as differences of visualizations and simulations with respect to the dynamics and the interactivity. Moreover, the learners' prior knowledge and the complexity of the learning task have to be accurately analyzed in further studies because they seem to significantly affect the use of actively integrating multiple representations.

Acknowledgements

The work reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG).

References

Ainsworth, S. (1999). The functions of multiple representations. *Computers and Education, 33*, 131-152.

Ainsworth, S., Bibby, P. A., & Wood, D. J. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences, 11*(1), 25-62.

Ainsworth, S. & van Labeke, N. (in press). Multiple forms of dynamic representation. *Learning and Instruction.*

Blaschke, K., & Heuer, D. (2000). Dynamik-Lernen mit multimedial-experimentell unterstütztem Werkstatt-Unterricht [Learning dynamics in multimedia projects]. *Physik in der Schule, 38*(2), 1-6.

Bodemer, D., Ploetzner, R. Feuerlein, I. & Spada, H. (in press). The active integration of information during learning with dynamic and interactive visualizations. *Learning and Instruction.*

Brünken, R., Steinbacher, S., Schnotz, W., & Leutner, D. (2001). Mentale Modelle und Effekte der Präsentations- und Abrufkodalität beim Lernen mit Multimedia [Mental models and the effects of presentation and retrieval mode in multimedia learning]. *Zeitschrift für Pädagogische Psychologie, 15*, 15-27.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology, 62*, 233-246.

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179-201.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45-56.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351-371.

Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction, 13*(2), 205-226.

Kozma, R.B., Russell, J., Jones, T., Marx, N., & Davis, J. (1996). The use of multiple, linked representations to facilitate science understanding. In S. Vosniadou, E. De Corte, R. Glaser & H. Mandl (Eds.), *International perspectives on the design of technology supported learning environments* (pp. 41-61). Hillsdale, NJ: Erlbaum.

Larkin, J.H., & Simon, H.A. (1987). Why a diagram is (sometimes) worth ten thousands words. *Cognitive Science, 11*, 65-99.

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: effects of adaptive and non-adaptive instructional support. *Learning and Instruction, 3*, 113-132.

Lowe, R. K. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education, 14*(2), 225-244.

Lowe, R. K. (2003). Animation and learning: Selective processing of information in dynamic graphics. *Learning and Instruction, 13*(2), 157-176.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*(1), 1-19.

Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press.

Njoo, M., & de Jong, T. (1993). Supporting exploratory learning by offering structured overviews of hypotheses. In D. M. Towne & T. de Jong & H. Spada (Eds.), *Simulation-based experiential learning* (pp. 207-223). Berlin: Springer Publishers.

Ploetzner, R., Bodemer, D., & Feuerlein, I. (2001). Facilitating the mental integration of multiple sources of information in multimedia learning environments. In C. Montgomerie & J. Viteli (Eds.), *Proceedings of the World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 1501-1506). Norfolk, VA: Association for the Advancement of Computing in Education.

Reigeluth, C. M., & Schwartz, E. (1989). An instructional theory for the design of computer-based simulations. *Journal of Computer-Based Instruction, 16*(1), 1-10.

Reimann, P. (1991). Detecting functional relations in a computerized discovery environment. *Learning and Instruction, 1*, 45-65.

Rieber, L. P., Tzeng, S.-C. & Tribble, K. (in press). Discovery learning, representation, and explanation within a computer-based simulation: Finding the right mix. *Learning and Instruction.*

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences, 1*, 201-239.

Schnotz, W., & Bannert, M. (1999). Einflüsse der Visualisierungsform auf die Konstruktion mentaler Modelle beim Text- und Bildverstehen [Influence of the type of visualization on the construction of mental models during picture and text comprehension]. *Zeitschrift für Experimentelle Psychologie, 46*(3), 217-236.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*(2), 141-156.

Schnotz, W., Boeckheler, J., & Grzondziel, H. (1999). Individual and co-operative learning with interactive animated pictures. *European Journal of Psychology of Education, 14*(2), 245-265.

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction, 13*(2), 227-237.

Swaak, J., van Joolingen, W. R., & de Jong, T. (1998). Supporting simulation-based learning: The effects of model progression and assignments on definitional and intuitive knowledge. *Learning and Instruction, 8*(3), 235-252.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Tarmizi, R. A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology, 80*(4), 424-436.

van Joolingen, W. R. & de Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science, 20*(5-6), 389-404.

Simple and Complex Extralinguistic Communicative Acts

Francesca M. Bosco (bosco@psych.unito.it)
Katiuscia Sacco (sacco@psych.unito.it)
Livia Colle (colle@psych.unito.it)
Romina Angeleri (angeleri@psych.unito.it)
Ivan Enrici (enrici@psych.unito.it)
Gianluca Bo (gianluca.bo@metis-ricerche.it)
Bruno G. Bara (bara@psych.unito.it)
Centro di Scienza Cognitiva e Dipartimento di Psicologia, Università di Torino
via Po 14 - 10123 Torino, Italia

Abstract

The present research aims to investigate the distinction between simple and complex communicative acts in the context of extralinguistic communication. We propose that, within the same pragmatic phenomena studied - which are standard communicative acts, deceit and irony - a simple communicative act is easier to comprehend than a complex one. Our proposal is based on the different complexity of inferential processes involved in comprehending communicative acts. We provide empirical evidence in support to our hypothesis with an experiment on children aged 5;5 to 8;6 years. We consider our results as favoring a unitary model of communication, where 'linguistic' and 'extralinguistic' are similar expressive channels underlying the same cognitive faculty.

Introduction

Philosopher John Searle (1975) introduced the classical distinction between direct and indirect speech acts. A direct speech act consists of a sentence where a speaker means exactly and literally what she is saying, for instance [1] 'Please pass me the salt', proffered by the speaker to obtain the salt, located on the table, from her table-companion. On the contrary, an indirect speech act consists of a sentence by which the speaker communicates to the hearer more than what she is actually saying. For instance [2] 'Do you mind passing me the salt?' or [3] 'My soup is lacking in salt', proffered by a speaker in order to obtain the same goal as in the previous example.

Searle claims that the primary illocutionary force of an indirect speech act is derived from the literal meaning *via* a series of inferential steps. The hearer's inferential process is triggered by the assumption that the speaker is following the Principle of Cooperation (Grice, 1975), together with the evidence of an inconsistency between the utterance and the context of enunciation. According to Searle, the hearer tries first to interpret the utterance literally, and only after the failure of this attempt, due to the irrelevance of the literal meaning, does he look for a different one, which conveys the primary illocutionary force. In this view, an indirect speech act is intrinsically harder to comprehend than a direct one. Indeed, understanding a direct speech act such as [1] is straightforward, that is, it does not require inferences, while understanding indirect speech acts, such as [2] and [3] relies on some kind of common knowledge. However, the length of the inferential path is not the same for each indirect speech act. For instance [3], an example of non conventional indirect speech act, requires a greater number of inferences than [2], an example of conventional direct speech act.

Some authors have criticized this position for different reasons (Clark, 1979; Sperber & Wilson, 1986; Recanati, 1995). In particular, Gibbs (1986; 1994) shows that a speaker can use an indirect act when she thinks that there might be obstacles against the request she intends to formulate: for example, when the speaker does not know whether the hearer owns the object she desires, he can use a conventional indirect request. The context specifies the necessity of using a conventional indirect and thus helps the hearer to understand the intended meaning more quickly. Gibbs suggests that, in such a circumstance, the partner infers the meaning of a conventional indirect speech act *via* an habitual shortcut that facilitates its comprehension.

In addition, Bara and Bucciarelli (1998) point out that for 2;6-3 year old children conventional indirects, such as 'Would you like to sit down?', compared to direct speech acts, such as 'What is your name?', are equally easy to comprehend. On the contrary, the same children have difficulties with non-conventional indirects: for instance they find it hard to understand that the answer 'It's raining' to the proposal 'Let's go out and play' corresponds to a refusal.

On the basis of the Cognitive Pragmatics theory, by Airenti, Bara and Colombetti (1993a), Bara, Bosco and Bucciarelli (1999) advanced an alternative explanation that constitutes the theoretical basis for the present research. The authors propose to abandon the distinction between direct and indirect speech acts and to adopt a new one between simple and complex speech acts, based on the increasing complexity of the inferential processes underlying their comprehension. This distinction has the distinct advantage of applying not only to standard speech acts, but also to non standard ones, like irony and deceit.

The aim of the present research is to extend the distinction between simple and complex speech acts to extralinguistic communication, and to provide empirical evidence in support of our hypothesis.

Cognitive Pragmatics theory

Airenti et al. (1993a) have presented the bases for a theory of the cognitive processes underlying human communication that holds for both linguistic and extra-linguistic communication. A major assumption of Cognitive Pragmatics is that intentional communication requires behavioral cooperation between two agents; this means that when two agents communicate they are acting on the basis of a plan that is at least partially shared. The authors call this plan a *behavior game*. The behavior game is a social structure mutually shared by the participants of the dialogue. Each communicative action performed by the agents realizes the moves of the behavior game they are playing. The meaning of a communicative act (either linguistic or extralinguistic or a mix of the two) is fully understood only when it is clear what move of what behavior game it realizes. Consider for example the following communicative exchange:

[4] Susan: "Do you have 10 dollars?"
Mark: "Oh, I forgot my wallet"

Mark understands that Susan is asking him to lend her some money on the basis of the behavior game they are mutually sharing:

[5] [LEND-MONEY]:
• A gives money to B;
• B returns money to A.

A game provides a context for the assignment of meaning to a communicative action (Bosco, Bucciarelli & Bara, 2004). It is the sharedness of these knowledge structures that allows them to maintain conversational cooperation in spite of Mark's refusal to cooperate on the behavior level.

Simple and complex standard speech acts

The comprehension of any kind of speech act depends on the comprehension of the behavioral game bid by the actor[1]. Unless a communicative failure occurs, each participant in a dialogue interprets the utterances of the interlocutor on the ground she gives as shared between them. According to such a perspective, the difficulty in comprehension of different types of speech acts depends on the chain of inferences required to pass from the utterance to the game it refers to. Direct and conventional indirect speech acts do immediately make reference to the game, and thus they are defined as *simple speech acts*. On the contrary, non conventional indirect speech acts can be referred to as *complex speech acts* in that they require a chain of inferential steps, since the specific behavior game of which they are a move is not immediately identifiable (Bara & Bucciarelli, 1998). For example, to understand [1] and [2] it is sufficient for the partner to refer to the game [ASK-FOR-OBJECT]. In order to understand [3], a more complex inferential process is necessary: the partner needs to share with the speaker the belief that if the soup is lacking salt it is not good to eat and that if there is some salt on the table and somebody proffers [3] she probably wants it. Only then, can the partner attribute to the utterance the value of a move of the game [ASK-FOR-OBJECT].

In other words, if the problem is how to access the game, the distinction between direct and indirect speech acts is irrelevant. The comprehension of a speech act requires the comprehension of the game of which it is part: in order to understand the actor's communicative intentions, the partner has to find a meaningful connection between the actor's utterance and the behavioral game they are playing. In the case of simple speech acts there is an immediate correspondence between the utterance and the game, that is the utterance straightforwardly refers to the game. On the contrary, in the case of complex speech acts the comprehension of the link between the speech act and the game requires the partner to make longer inferential processes. The bigger the distance between the utterance and the communicative context shared by actor and partner, the more difficult the comprehension of the utterance itself. In sum, the difference in the difficulty of comprehension between simple and complex acts depends on the steps needed to refer the utterance to the game bid by the actor or already shared by the participants.

We find the notion of simple and complex speech act more useful rather than the one of direct and indirect speech act because, as Bara et al. (1999) propose, it can be extended to other non standard pragmatic phenomena, in particular irony and deceit.

Simple and complex deceits and ironies

A deceit occurs when the mental states that the actor entertains are covertly different from those she communicates. Bara et al. (1999) propose that the difficulty in its comprehension can vary depending on the complexity of the inferential chain necessary to refer the utterance to the behavioral game. Consider the following example:

[6] *Andrew is eating some biscuits from a plate in front of him. He hears Julia arriving, and then he pushes away the empty plate in front of him. Julia sees the empty plate and asks:* "Who has finished my biscuits?". *Andrew answers…*
(a) Simple: "I don't have the slightest idea"
(b) Complex: "I'm on a diet"

In our example, the deceitful speech act [6a] is simple because it consists in an utterance which denies the actor's private (and true) belief *(not-p)*, that would allow the partner to immediately refer to the game [BISCUIT-STEALING] that the actor wishes to conceal from the partner. Instead, a complex deceitful speech act, such as [6b], consists in an utterance which leads to the inference: if he is on a diet, he cannot eat biscuits, that is inconsistent with the game [BISCUIT-STEALING] that the actor wishes to deny. Thus,

[1] Since the theory holds both for linguistic and extralinguistic communication, we prefer to use the terms actor and partner instead the classical speaker and hearer.

to comprehend a complex deceit, an agent needs a longer inferential chain.

Cognitive Pragmatics theory claims that irony can be understood when compared with the belief provided by the behavior game shared between actor and partner (Airenti, Bara & Colombetti, 1993b; Bara, in press). According to Bara et al. (1999), bearing in mind the complexity of the inferential chain necessary to refer the utterance to the game bid by the interlocutors, it is possible to distinguish two kinds of irony, simple and complex. Consider the following example:

[7] *Alex takes out from a toaster two completely burned pieces of toast. Mary arrives and Alex asks with a puzzled expression*: "Am I a good cook?" Mary *answers*…

(a) Simple *:* "The best cook in the world!"

(b) Complex: "I'll hire you in my restaurant"

A simple ironic speech act, such as [7a], corresponds to the antiphrastic theory of irony (Grice, 1989): an actor expresses *p* to mean *not-p*. Thus, a simple irony immediately contrasts with a belief shared between the agents, in our example that Alex is not a good cook. On the contrary, a complex ironic speech act requires a series of inferences in order to detect its contrast with the belief shared by the agents. Consider our example, by producing the complex irony [7b], an actor proffers an utterance which implies the belief *p* (to employ someone in a restaurant, s/he has to be a good cook), contrasting with the belief *not-p* (the guy is not a good cook), shared between the two agents. Thus, a person needs a longer inferential chain to comprehend a complex deceit rather than a simple one.

Bosco and Bucciarelli (submitted) empirically supported the distinction between simple and complex speech acts: children aged from 6;7 to 10 years, find it easier to comprehend simple speech acts, rather than complex ones, within the same pragmatic phenomena investigated, i.e. standard speech acts, deceits and ironies.

The present research focuses on the difference of the inferential processes between communicative acts pertaining to the same pragmatic category and it did not analyze the difference among inferential processes existent among various kind of pragmatic phenomena. Details about how different types of mental representations underlie the comprehension of standard communicative acts, deceits and ironies can be found elsewhere, e.g. Bucciarelli, Colle and Bara (2003): such a work focuses on the *type of inference* underlying specific kinds of pragmatic phenomena; for instance, understanding an ironic act requires the detection of a contrast between the speech act and the background knowledge shared by the interlocutors. The present research deals instead with the *length of the inferential* processes underlying the comprehension of communicative acts within the same pragmatic phenomenon, i.e. simple vs. complex standard acts, simple vs. complex deceits, simple vs. complex ironies.

Experiment: simple vs. complex extralinguistic communication

As we have shown in the previous paragraphs, the difference between simple and complex acts has been demonstrated in the context of linguistic communication. Let us now focus on extralinguistic communication. By extralinguistic communication we refer to actions such as facial expressions, hand gestures and body movements when they are intentionally performed to share a communicative meaning. These means of expression are of special importance in that communication, in the first phases of life, heavily relies on such kinds of actions. Also, persons who have lost the ability to communicate through language, e.g. patients with aphasia, have to resort to extralinguistic means. Not to mention the various kinds of situations in which normal adults need to communicate but are forced not to use speech. For all these sort of reasons, we conducted a study in the context of extralinguistic communication. In particular, our aim is to analyze whether the distinction between simple and complex acts holds also in such a context: if language and gestures are comparable ways of communication, we should expect that the distinctions made in linguistic contexts holds also for extralinguistic communication.

According to Cognitive Pragmatics theory, communication is indeed a unitary cognitive faculty aimed at modifying and sharing mental states, while 'linguistic' and 'extralinguistic' are means of expression that an agent may use indifferently in order manifest to and share her communicative intentions. For instance, waving a hand or saying 'Hello' are two ways of greeting that are only superficially different; at a deeper level, they can be seen as two different realizations of a greeting act. Thus, as Bara and Tirassa (1999; 2000) propose, the difference between 'linguistic' and 'extralinguistic' communicative acts turns out to be a matter of cognitive processing rather than of intrinsic nature. Within such a perspective Bucciarelli et al. (2003) assume that the construction of the meaning of a communicative act is independent of the input modalities[2]. Empirically, they tested the prediction that a communicative act has in principle the same difficulty of comprehension, whether performed through speech or gestures. Their results show that children of different age groups comprehend each pragmatic phenomenon (simple and complex standard communicative acts, simple deceits and simple ironies) equally well in the two modalities.

The present experiment investigates the comprehension of different extralinguistic communicative acts. It consists of two experimental conditions: simple and complex. In both conditions, participants have to attribute communicative

[2] Note that on the bases of other theoretical approaches (Burling, 1993; Chomsky, 1987), nonverbal and verbal communication do have separate roots in phylogenies; on such basis, the prediction that what holds for linguistic communication holds also for extralinguistic communication would be false.

intentions to actors in videotaped stories. Our analysis focuses on a perspective of a third person who observes an actor and a partner in a communicative interaction. We now report one example for each of the investigated pragmatic phenomena. They are all extracted from our experimental protocol.

Standard extralinguistic communication
[8] *Ann has just finished preparing dinner and walks out of the kitchen holding a dish of pasta. In order to call Bob, who is listening to loud music, Ann moves her head as if to say 'Come on! Dinner's ready'.*
In the simple version of the task, Bob [8a] nods to show that he is coming. In the complex version of the task, Bob [8b] places his hand on his stomach as if to say 'I'm hungry'. The request of coming for dinner is part of the behavior game [FAMILY-DINNER], in which Ann prepares dinner, calls Bob when dinner is ready, and B answers. Bob's [8a] nodding is a simple standard communicative act because it is a straightforward answer to Ann's question and, thus, immediately relies on the game shared between the two agents. On the contrary, to understand that the complex standard gesture [8b] for 'I'm hungry' implies an acceptance, a person has to assume that if one is hungry then he wants to eat and that if one wants to eat then he has the intention of coming to dinner.

Extralinguistic deceit
[9] *Bill and his brother are playing with cushions in their room, when a lamp falls down and breaks into pieces. Mum comes into the room and, standing with her hands on her hips, she assumes a severe and questioning look as if to ask 'Who broke the lamp?'.*
In the simple version of the task, Bill [9a] opens his arms in order to state his innocence. In the complex version of the task, Bill [9b] takes a book and shows it to Mum in order to convince her he was reading. Bill's [9a] gesture is a simple deceit because it immediately denies the actor's private belief, allowing the partner to refer to the game [DOMESTIC-MISDEED]. On the contrary, [9b] is a complex deceit because it implies a belief (if one is reading a book he is not moving, then he cannot cause any damage) that is inconsistent with the game [DOMESTIC-MISDEED]. Thus, in order to understand this sort of deceit one needs to make a more complex inferential chain.

Extralinguistic irony
[10] *Alice pours some soup into her and Ben's plates and both assume a disgusted look. Alice looks at Ben as if she is waiting for a comment.*
In the simple version of the task, Ben, with an ironic expression [10a] licks his lips as if to say 'It's delicious!'. In the complex version of the task, Ben with an ironic expression, [10b] gives his plate to Alice as to ask to have some more soup. Ben's [10a] gesture is a simple irony because it immediately contrasts with the belief (the soup is not good) which is part of the game [HOME-COOKING] shared between Alice and Ben. On the contrary, the complex irony [10b] implies a belief (if one asks for more food, it is because the food is good) that contrasts with the belief 'the soup is not good' shared between the two agents.

In conclusion, within the same pragmatic category, comprehending a simple communicative act requires an easier inferential chain than that required for a complex act. Indeed, simple acts immediately refer to the behavior game shared by actor and partner, while complex acts do not. Thus, for each of the investigated pragmatic phenomena, we predict that simple communicative acts are easier to comprehend than complex communicative acts.

Our study was conducted on children of different age groups. Indeed, adult subjects possess a fully developed cognitive system and communicative competence, and thus should not show any interesting errors in comprehending the different kinds of pragmatic tasks. On the contrary, within a developmental perspective, we expect that the ability to comprehend each kind of communicative act improves with children's age.

Material and Procedures

The experimental material comprised 12 videotaped scenes, each lasting 20-25 seconds and showing two characters engaged in a communicative interaction. All communicative acts were completely extralinguistic, performed only through gestures. Of these 12 scenes, 4 represented standard communicative acts, 4 deceiving acts and 4 ironic acts. Each scene has been recorded in two versions, one simple and one complex (see the examples described in the previous paragraph). Thus we devised two experimental protocols, A and B. Each protocol contains only one version for each scene. In each protocol the scenes are represented in a different random order. Half of the participants dealt with protocol A, while the other half dealt with protocol B. Each child was randomly assigned to protocol A or B. Every child saw 4 scenes representing a standard communicative act (2 simple + 2 complex), 4 scenes representing a deceiving communicative act (2 simple + 2 complex) and 4 scenes representing an ironic act (2 simple + 2 complex).

At the end of each scene, children had to show that they had understood the communicative interaction by explaining to the examiner what had happened and what the actor's communicative intention was. Participants' responses were rated by 2 independent judges. For each item, judges assigned a score of Ø (completely wrong answer), 1 (only partially correct answer) or 2 (correct answer).

Participants

The protocol was administered individually to 300 children, divided into three age groups: 100 children ranging from 5 to 5;6 (mean age = 5;3 years), 100 children ranging from 6;6 to 7 (mean age = 6;9 years), and 100 children ranging from 8 to 8;6 (mean age = 8;2). Within each age group, there were 50 males and 50 females. Children came from nursery and primary schools of Turin.

Results

Our hypotheses were globally confirmed. Figure 1 shows the mean percentages of the correct responses over all children to the simple and complex items: in every type of investigated phenomena (standards, deceits and ironies), subjects understand the simple communicative acts better than the complex ones. More in detail, overall children understand simple standard communicative acts more easily than the complex ones (T Test: $t = 5.55$; $p < .0001$).

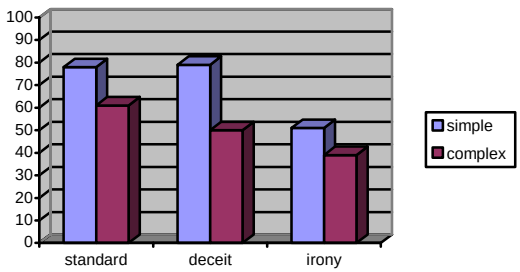


Figure 1: Histogram of the mean percentages of correct responses over all children.

As shown in Table 1, the same result holds for 5-year-olds (T Test: $t = 3.20$; $p < .002$), for 6-year-olds (T Test: $t = 2.67$; $p < .009$) and for 8-year-olds (T Test: $t = 3.76$; $p < .0001$). The same pattern of results holds also for simple and complex deceits. Simple deceits are easier, both overall subjects (T Test: $t = 10.19$; $p < .0001$), and within the various groups: for 5-year-olds (T Test: $t = 5.63$; $p < .0001$), for 6-year-olds (T Test: $t = 5.98$; $p < .0001$) and for 8-year-olds (T Test: $t = 6.15$; $p < .0001$). Finally, for ironic acts, simple ones are easier than complex ones over all subjects (T Test: $t = 3.26$; $p < .001$), for 6-year-olds (T Test: $t = 2.24$; $p < .03$) and for 8-year-olds (T Test: $t = 3.11$; $p < .003$), whereas there is no significant difference for 5-year-olds (T Test: $t = 0.65$; $p < .52$).

Table 1**:** Mean percentages of correct responses over all children and for single age groups.

Age groups	Standard		Deceit		Irony	
	Simple	Complex	Simple	Complex	Simple	Complex
5-5;6	67	50	70	42	37	34
6;6-7	79	65	83	53	56	39
8-8;6	86	68	83	56	64	46
Global	78	61	79	50	51	39

We also found significant data concerning children's performance improvement, in understanding every kind of task, in accordance with the increase of their age. Differences in performance among the three groups have resulted in both simple and complex standard speech acts (Anova: F ranging from 6.53 to 10.36; p ranging from .002 to .0001). Also the performances in comprehension of deceit improve as the age of the subjects increase, both for simple and complex deceits (Anova: F ranging from 3.32 to 6.44; p ranging from .002 to .03). The same result holds for simple ironies (Anova: $F = 13.23$; $p < .0001$) and for complex ironies (Anova: $F = 3.1$; $p < 0.05$).

Conclusions

In the present study we aimed to extend the analysis on simple vs. complex communicative acts to the domain of extralinguistic communication. The results globally confirm our predictions. Simple communicative acts are easier to comprehend than complex ones in all pragmatic phenomena investigated. This is true of all subjects, even within age groups. We explain such data considering the cognitive processes underlying the comprehension of the investigated tasks: in order to be understood complex communicative acts involve a higher inferential load than simple ones. Furthermore, as predicted, children's improve their performance in all investigated tasks, in accordance with the increase of their age.

Only in one case our data are not in line with our expectations: we did not detect significant differences in the comprehension of simple vs. complex ironies in the youngest group of 5-year-olds. A possible explanation is that irony is a too difficult pragmatic phenomenon to be fully understood by children of that age. For this reason irony comprehension results difficult in both cases (simple and complex): children gave such a few correct answers, that no significant difference emerged. This interpretation is consistent with results in literature which showed that only 6 year-old-children seem to fully grasp the intentions of ironic exchanges (Lucariello & Mindolovich, 1995). In line with such data Bucciarelli et al. (2003) found that irony – expressed both by linguistic speech acts and by gestures – is the most difficult pragmatic phenomena to comprehend, in comparison to standard communicative acts and deceits, for children aged 2;6 to 7 years. In addition, though still in line with our results, the authors found that only a small percentage (38%) of the 4;6-5;6 children in their study understood ironic gestures in an experimental setting.

Our results on children's ability to interpret extralinguistic gesture with a deceitful intent are in line with other experimental studies. For example Shulz and Cloghesy (1981) showed that only from 5 years of age children start to interpret pointing gestures with a deceitful intent, and that such an ability improves with the age. A related task has been studied by Call and Tomasello (1999). The authors investigated children's ability to deal with deceits - with the classic false belief task - in an extralinguistic form. The results are consistent with the verbal version of the task: only a few 4 year olds are able to complete the task, whereas most 5 year olds succeed.

Let us now consider our results in a wider perspective. Our main prediction was to detect an increasing difficulty in comprehension between simple and complex extra-

linguistic communicative acts in different pragmatic tasks. Such a prediction was grounded on the assumption that comprehension of simple and complex communicative acts can be explained by the complexity of the inferential chain involved in each of them, despite the communicative channel used to express them, i.e. linguistic or extralinguistic. Our results confirm such a perspective: we find the same trend of difficulty between simple and complex communicative acts that other studies underlined in linguistic comprehension (see Bucciarelli et al., 2003; Bosco & Bucciarelli, submitted). These similarities between linguistic and extralinguistic comprehension, which we found in each of the investigated pragmatic phenomenon, confirm that speech acts and extralinguistic communicative acts share the most relevant mental processes. Opposing viewpoints (e.g. Chomsky, 1987; Burling, 1993) consider linguistic and extralinguistic communication as two distinct phenomena, different in their intrinsic nature, and having separate roots in phylogenies. According to such a view, language is a complex module, independently evolved due to a non-finalized, genetic mutation. Our data seem to falsify the hypothesis of a separated line of development of language and communication, in favor of a unified theoretical framework in which linguistic and extralinguistic communication develop in parallel as different aspects of a unique communicative competence (Bara, in press).

Acknowledgments

This research was supported by the Ministero Italiano dell'Università e della Ricerca Scientifica (MIUR), FIRB Project, research code RBAU01JEYW.

References

Airenti, G., Bara, B. G., & Colombetti, M. (1993a). Conversation and behavior games in the pragmatics of dialogue. *Cognitive Science, 17*, 197-256.

Airenti, G., Bara, B. G., & Colombetti, M. (1993b). Failures, exploitations and deceits in communication. *Journal of Pragmatics*, *20*, 303-326.

Bara, B. G. (in press). *Cognitive Pragmatics*. Cambridge, MA: MIT Press.

Bara, B. G., & Bucciarelli, M. (1998). Language in context: The emergence of pragmatic competence. In A. C. Quelhas & F. Pereira (Eds.), *Cognition in context*. Lisbon: Istituto Superior de Psicologia Aplicada.

Bara, B. G., & Tirassa, M. (1999). A mentalist framework for linguistic and extralinguistic communication. In S. Bagnara (Eds.), *Proceedings of the 3rd European Conference on Cognitive Science* (pp. 285-290). Roma: Istituto di Psicologia del Consiglio Nazionale delle Ricerche.

Bara, B. G., & Tirassa, M. (2000). Neuropragmatics: Brain and communication. *Brain and Language, 71,* 10-14.

Bara, B. G., Bosco, F. M., & Bucciarelli, M. (1999). Simple and complex speech acts: What makes the difference within a developmental perspective. In M. Hahn & S. Stones (Eds.), *Proceedings of the XXI Annual Conference of the Cognitive Science Society* (pp. 55-60). Mahwah, NJ: Lawrence Erlbaum Associates.

Bosco, F. M., & Bucciarelli, M. (submitted). Simple and complex deceits and ironies.

Bosco, F. M., Bucciarelli, M., & Bara, B. G. (2003). Literal meaning and context categories in the attribution of communicative intentions. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the XXV Annual Conference of the Cognitive Science Society* (pp. 162-167). Boston, Massachusetts: Cognitive Science Society.

Bosco, F. M., Bucciarelli, M., & Bara, B. G. (2004). Context categories in understanding communicative intentions. *Journal of Pragmatics. 36*, 436-488.

Bucciarelli, M., Colle, L., & Bara, B. G. (2003). How children comprehend speech acts and communicative gestures. *Journal of Pragmatics, 35,* 207-241.

Burling, R. (1993). Primate calls, human language, and non verbal communication. *Current Anthropology, 34*, 25-53.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development, 70*, 381-395.

Chomsky, N. (1987). *Language and problems of knowledge:* The Managua Lectures. Cambridge, MA: MIT Press.

Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive Psychology*, *11,* 430-477.

Gibbs, R. W. (1986). What makes some indirect speech acts conventional? *Journal of Memory and Language, 25*, 181-196.

Gibbs, R. W. (1994). *The poetics of mind*. Cambridge: Cambridge University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press.

Grice, H. P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Lucariello, J., & Mindolovich, C. (1995). The development of complex metarepresentational reasoning: the case of situational irony. *Cognitive Development, 10*, 551-576.

Recanati, F. (1995). The alleged priority of literal interpretation. *Cognitive Science, 19*, 207-232.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press.

Shultz, T. R., & Cloghsey, K. (1981). Development of recursive awarness of intention. *Developmental Psychology, 17,* 465-471.

Sperber, D. & Wilson, D. (1986). *Relevance*. Oxford, UK: Blackwell.

Similarity and Categorisation: Getting Dissociations in Perspective

Nick Braisby (N.R.Braisby@open.ac.uk)
Department of Psychology, The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK

Abstract

Dissociations between similarity and categorization have constituted critical counter-evidence to the view that categorization is similarity-based. However, there have been difficulties in replicating such dissociations. This paper reports three experiments. The first provides evidence of a double dissociation between similarity and categorization. The second and third show that by asking participants to make their judgments from particular perspectives, this dissociation disappears or is much reduced. It is argued that these data support a perspectival view of concepts, in which categorization is similarity-based, but where the dimensions used to make similarity and categorization judgments are partially fixed by perspective.

Introduction

Explanations of categorisation have undergone a number of theoretical shifts (Medin, 1989), from classical to prototype models, and from prototype to theory-based models (e.g., Murphy & Medin, 1985). One of the key pieces of evidence against similarity-based models has been the finding that similarity and categorisation judgments can dissociate. For example, Rips (1989) found that participants judged an unknown item more similar to a coin yet more likely to be a pizza; and a bird transformed to look like an insect as more similar to an insect, yet more likely to be a bird.

Dissociations between similarity and categorization judgments appear directly to undermine similarity-based models of categorization. Prototype models, for example, assume that categorisation involves a similarity comparison between an object and a prototype in memory (e.g., Hampton, 1995). Exemplar-based models assume that a similarity comparison is made between an object and sets of exemplars in memory. In both kinds of model, categorization is taken to be a monotonic increasing function of similarity. That is, according to similarity-based models, it should not be possible for categorization to increase without a corresponding increase in similarity. These models thus deny the possibility of two kinds of change: i) a decrease in categorization accompanied by an increase in similarity; and ii) a decrease in categorization accompanied by no change in similarity.

In spite of the evidence and arguments in support of similarity-categorisation dissociations (henceforth, SCD), similarity-based models have maintained their appeal. Some of this can be attributed to the apparent success of similarity-based models in explaining much categorization data (cf. Hampton, 1998) even if SCDs remain as recalcitrant cases. But similarity-based models have also retained their appeal because the existence of SCDs has been questioned (despite other apparent demonstrations – e.g., Kroska & Goldstone, 1996; Roberson, Davidoff & Braisby, 1999). Smith & Sloman (1994), in seeking to replicate Rips' results, were able only to produce a SCD when participants were required to operate in a reflective, rule-based mode, by giving a concurrent verbal protocol. Similarly, Estes & Hampton (2002) only obtained a SCD when using a within-participants design; a between-participants design failed to show a dissociation. In contrast, Thibaut, Dupont & Anselme (2002) obtained SCDs in two experiments. Their participants were required to learn two artificial categories, exemplars of which were novel shapes. They found that participants tended to judge category membership according to the presence of a necessary feature, but similarity according to the presence of a salient characteristic feature.

Thibaut et al.'s results show that SCDs can arise without participants entering a reflective mode of categorization. However, they do not demonstrate that natural (as opposed to artificial) categories give rise to SCDs. That is, they have shown that participants can learn and use non-similarity-based categories, but not that natural categories are not similarity-based.

This paper seeks to add to this debate concerning SCDs and, more widely, similarity-based models of concepts by i) attempting to demonstrate a double similarity-categorisation dissociation; and ii) showing that such dissociations can be eliminated or diminished when judgments are given in context or perspective.

Previous work (e.g., Thibaut et al.) has shown that stimuli defined by the presence of both necessary and characteristic features may be categorized according to the necessary feature, and rated for similarity according to the characteristic feature. Of course, such work also implies the existence of two kinds of (potentially) borderline case: (a) an exemplar possessing the necessary but not the characteristic feature (N+C−); and (b) an exemplar possessing the characteristic but not the necessary feature (N−C+). According to the rationale extended by Thibaut et al., exemplar (a) should receive

a high categorisation but low similarity rating, and (b) should receive a low categorization but high similarity rating. Together these borderline cases could provide a double dissociation, and potentially more robust evidence of SCDs.

This paper also seeks to establish whether SCDs are context-sensitive. That is, it is possible that in context, categorization judgments are similarity-based, but that dissociations arise when categorization and similarity judgments are elicited in the absence of any specific context. If so, then the mixed evidence reported in the literature may stem from minor variations in task presentation, and it might be possible to retain a similarity-based model in which the weighting of features varies with context.

Experiment 1

This experiment sought to establish whether similarity and categorization judgments for biological categories dissociate for two kinds of borderline case: Appearance+Genetics− and Appearance−Genetics+.

Design

Task (Typicality, Categorisation), Appearance (A+,A−) and Genetics (G+,G−) were within-participants factors.

Method

Participants 40 undergraduate psychology students attending an Open University residential school volunteered to participate.

Materials Four food categories were chosen based on previous work (Braisby, 2001): salmon, apple, potato and chicken. Sixteen scenarios were constructed, as described below, so that there were four exemplars per category defined by the presence or absence of appearance and genetic properties: A+G+; A+G−; A−G+; and A−G−. The following shows how scenarios were constructed for the category 'apple'; the first set of brackets indicates wordings for G+ and G− conditions, and the second set indicates the A+ and A− wordings.

"You have just bought an apple from a reputable retailer. On examining its packaging closely, you find that it has been genetically modified [but it retains ALL/so that it has NONE] of the genetic properties specific to apples. On closer examination, you find that it [looks, feels, smells and even tastes JUST/does NOT look, feel, smell or even taste] like an apple."

Procedure All scenarios were presented and responses recorded using E-prime (Schneider, Eschman & Zuccolotto, 2002). Participants were given a practice example, and then asked to read the 16 scenarios. After each, participants first judged the category membership of the exemplar given the category label (e.g., apple), choosing either a Yes or No judgment. They then rated the exemplar for typicality on a 7-point scale relative to the category label. The typicality question is taken to be an index of similarity (cf. Hampton, 1998). Order of presentation of scenarios was random.

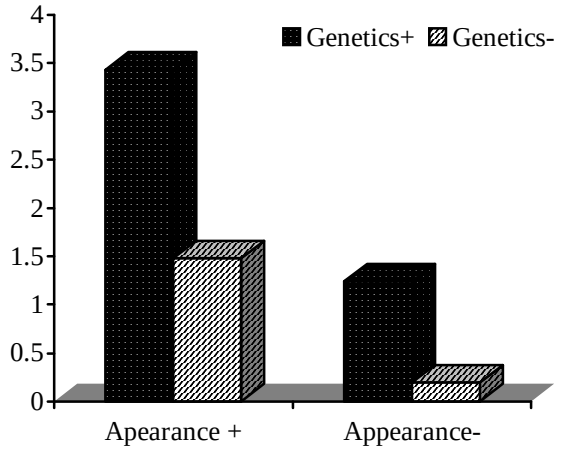


Figure 1. Overall ratings by Appearance and Genetics.

Results

Responses to the categorization question were summed over the four categories, yielding a scale of 0 to 4; the typicality question was transformed to the same scale (high scores imply high typicality and high categorization probability). Both typicality and categorization scores were analysed using ANOVA with Task (Typicality, Categorisation), Appearance (+,−) and Genetics (+,−) all within-participant factors.

There was no effect of Task ($p = 0.61$), but main effects of Appearance [$F(1,39) = 149.29$, $p < 0.001$; $\eta^2 = 0.79$] and Genetics [$F(1,39) = 109.59$, $p < 0.001$; $\eta^2 = 0.74$], interactions between Task and Appearance [$F(1,39) = 14.30$, $p < 0.01$; $\eta^2 = 0.27$], Task and Genetics [$F(1,39) = 11.63$, $p < 0.01$; $\eta^2 = 0.23$], Appearance and Genetics [$F(1,39) = 21.17$, $p < 0.001$; $\eta^2 = 0.35$], all subsumed by a marginal three-way interaction between Task, Appearance and Genetics [$F(1,39) = 3.88$, $p = 0.06$; $\eta^2 = 0.09$]. The key interactions between Appearance and Genetics, and between Task, Appearance and Genetics, are shown in Figures 1, and 2 and 3 respectively.

Pair-wise comparisons were conducted to examine the locus of the three-way interaction between Task, Appearance and Genetics. There was no effect of Task for either of the clear cases, i.e., either the A−G− or the A+G+ cases. However, there was an effect of Task, though in opposite directions, for the two borderlines. For the A+G− case, Typicality scores were markedly higher than Categorisation scores (Typicality = 1.93, categorization = 1.03, $t(39) = 3.21$, $p < 0.01$) while for the A−G+ case, Typicality scores were markedly lower

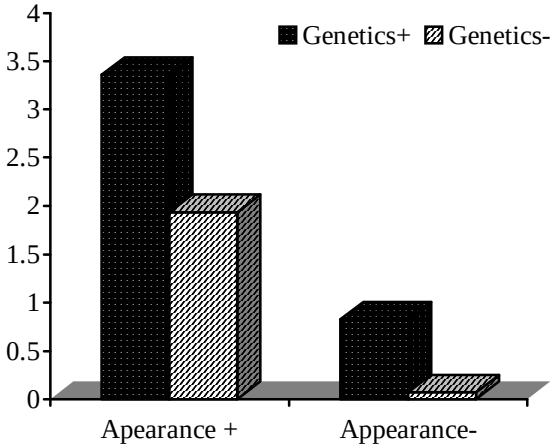


Figure 2. Typicality ratings by Appearance and Genetics.

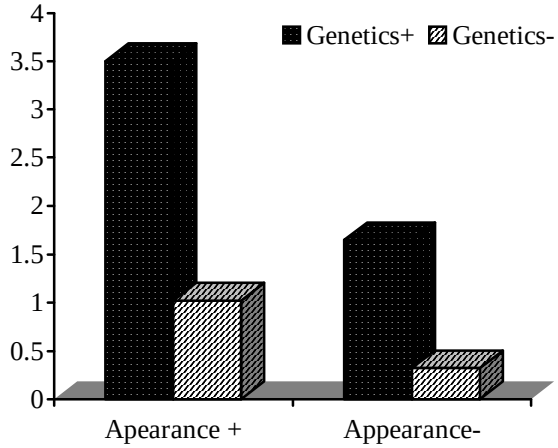


Figure 3. Categorisation ratings by Appearance and Genetics.

than Categorisation scores (Typicality = 0.83, categorisation = 1.65, $t(39) = 2.69$, $p < 0.05$).

Both Thibaut et al. and Estes & Hampton found that SCDs were due to only a subset of participants dissociatin their judgments. To examine this possibility, the number of times each participant gave dissociated pairs of judgments for the borderline items was calculated. A dissociated pair of judgments was defined in terms of differences between typicality and categorization responses for two borderline items within the same category, where the differences in the scores have different sign. For example, a participant might rate an A+G− as more typical than an A−G+, but categorise the former negatively and the latter positively. Though the difference in typicality scores is positive, the difference in categorization will be negative. Dissociations were also defined to include cases where participants gave differing categorization responses to the two borderlines within the same category, but gave the same typicality judgments. Each of these types of dissociation undermines the suggestion that categorization is a function of similarity.

Of the 40 participants, 25 (63%) gave no dissociated pair of judgments to the four pairs of borderlines with which they were presented; however, 5 (13%) gave dissociated judgments to all four pairs of borderlines.

Discussion of Experiment 1

Although previous research has claimed evidence of such dissociations, largely these have been single dissociations, i.e., items for which categorization points to category A and similarity to category B. In contrast, the present research dissociates these judgments in two ways. First, for A+G− items, typicality scores are higher than their corresponding categorization scores. For A−G+ items, typicality scores are lower than the corresponding categorization scores. Moreover, A+G− items are judged more typical than A−G+ items (1.93 and 0.83 respectively, $t(39) = 3.94$, $p < 0.001$); yet A+G− items are judged less likely to be in the category than A−G+ items (1.03 and 1.65 respectively, $t(39) = 1.73$, $p = 0.09$). Taken together, these findings present a challenge for similarity-based models, for it should not be possible for an item A to be more typical than item B and yet less likely to be a category member than item B.

It should also be noted that these materials do not present participants with fantastical transformations or discoveries. Nor do they tap artificial categories. Moreover, these data do not provide support for the idea that dissociations arise only under an especially reflective mode of categorisation. Response times were collected for both typicality and categorisation judgments. Typicality is often taken to be an index of an initial similarity computation, which can be over-ridden by a subsequent reflective categorisation. However, the response times in this experiment provide no support for this thesis: typicality response times averaged 5.13 seconds (including the time taken to read each scenario), yet categorisation averaged 3.86 seconds (again including reading time), a statistically significant difference [$t(39) = 5.40$, $p < 0.001$].

Lastly, though, it should be noted that the dissociations arise because of a subset of the participants, corroborating the findings of Estes & Hampton, and Thibaut et al.

In spite of not supporting similarity-based models, there is the possibility that categorisation and similarity judgments may align in context. That is, when in a specific context, it may be that participants make similarity-based categorisations, and dissociations arise only because these judgments are elicited out of context. The next two experiments sought to investigate this possibility by eliciting judgments in contexts

thought to emphasise either appearance or genetic properties. Both require participants to adopt a perspective in making their categorisation and typicality judgments (cf. Barsalou & Sewell, 1984)

Experiment 2

This experiment sought to establish whether similarity and categorization judgments dissociate for the two kinds of borderline, A+G− and A−G+, when participants are asked to adopt a perspective that emphasizes appearance properties.

Design

Task (Typicality, Categorisation), Appearance (A+,A−) and Genetics (G+,G−) were within-participants factors.

Participants 33 undergraduate psychology students attending an Open University residential school volunteered to participate.

Materials The same categories in experiment 1 were used. Scenarios were as in experiment 1, but were prefaced by the clause "Imagine that you are a Sculptor…". It was assumed, based on previous work, that participants would take this profession to signal the enhanced relevance of appearance properties.

Procedure An identical procedure to experiment 1 was followed. However, categorization and typicality questions were prefaced by the clause "Imagining yourself to be a sculptor…".

Results

Responses to the categorization and typicality questions (transformed as before) were analysed using ANOVA with Task (Typicality, Categorisation), Appearance (+,−) and Genetics (+,−) as within-subject factors.

There was no effect of Task ($p = 0.96$), but main effects of Appearance [$F(1,32) = 325.18$, $p < 0.001$; $\eta^2 = 0.91$] and Genetics [$F(1,32) = 35.30$, $p < 0.001$; $\eta^2 = 0.53$], interactions between Task and Genetics [$F1,32) = 6.82$, $p < 0.05$; $\eta^2 = 0.18$] and between Appearance and Genetics [$F(1,32) = 4.23$, $p < 0.05$; $\eta^2 = 0.12$]. However, there was no three-way interaction between Task, Appearance and Genetics ($p = 0.20$). The interaction between Appearance and Genetics is shown in Figure 4.

Although no three-way interaction was found, pair-wise comparisons were performed to examine the possibility that there might be differences between the Tasks for the two borderline items (since these were the source of the three-way interaction in experiment 1). Unlike experiment 1, there was no effect of Task for the A−G+ borderline, though there was a marginal effect for the A+G− borderline (Categorisation = 2.64, typicality = 3.17, $t(32) = 2.03$, $p = 0.05$).

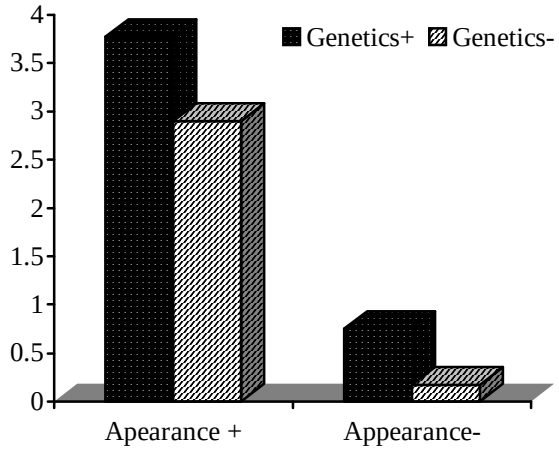


Figure 4. Overall ratings by Appearance and Genetics under a 'Sculptor' perspective.

As before, the response patterns of individual participants were examined to see how many gave dissociated judgments. Of the 33 participants, 23 (70%) gave no dissociated pair of judgments to the four pairs of borderlines with which they were presented; no participants gave dissociated judgments to all four pairs of borderlines.

Discussion of Experiment 2

Unlike experiment 1, these results provide no evidence of a dissociation between categorisation and similarity judgments. That is, the dissociation appears to have been eliminated by ensuring participants give their judgments from a specific perspective or context. The pairwise comparisons support this interpretation. Critically, the typicality and categorization scores do not violate the assumptions of similarity-based models: both A+G− and A−G+ cases differ in categorisation (2.64 and 0.94 respectively, $t(32) = 4.29$, $p < 0.001$) but also in typicality (3.17 and 0.56 respectively, $t(32) = 10.12$, $p < 0.001$). Hence the borderlines here do not provide evidence of dissociation – the increase in categorization is matched by an increase in typicality.

Experiment 3 seeks to establish whether the salience of genetic properties can be enhanced sufficiently to eliminate the dissociation in experiment 1.

Experiment 3

This experiment sought to establish whether similarity and categorization judgments dissociate for the two kinds of borderline, A+G− and A−G+, when participants are asked to make adopt a perspective that emphasizes genetic or biological properties.

Design

Task (Typicality, Categorisation), Appearance (A+,A−) and Genetics (G+,G−) were within-participants factors.

Participants 35 undergraduate psychology students attending an Open University residential school volunteered to participate.

Materials The same categories in experiment 1 were used. Scenarios were as in experiment 1, but were prefaced by the clause "Imagine that you are a Biologist…". It was assumed, based on previous work, that participants would take this profession to signal the enhanced relevance of biological properties.

Procedure An identical procedure to experiment 1 was followed. However, categorization and typicality questions were prefaced by the clause "Imagining yourself to be a biologist…".

Results

Responses to the categorization and typicality questions (transformed as before) were analysed using ANOVA with Task (Typicality, Categorisation), Appearance (+,−) and Genetics (+,−) as within-subject factors.

The pattern of results from the ANOVA exactly replicates that of experiment 2. There was no effect of Task ($p = 0.53$), but main effects of Appearance [$F(1,33) = 84.28$, $p < 0.001$; $\eta^2 = 0.72$] and Genetics [$F(1,33) = 125.02$, $p < 0.001$; $\eta^2 = 0.79$], an interaction between Task and Genetics [$F(1,33) = 13.93$, $p < 0.01$; $\eta^2 = 0.30$] and between Appearance and Genetics [$F(1,33) = 7.86$, $p < 0.01$; $\eta^2 = 0.19$]. As in experiment 2, there was no three-way interaction between Task, Appearance and Genetics. The interaction between Appearance and Genetics is shown in Figure 5.

As in experiment 2, although no three-way interaction was found, pair-wise comparisons were performed to examine the possibility that there might be differences between the Tasks for the two borderline items. As before, there was no effect of Task for the A−G+ borderline, but an effect of Task for the A+G− borderline (Categorisation = 1.03, typicality = 1.72, $t(34) = 2.51$, $p < 0.05$).

19 participants (54%) gave no dissociated pair of judgments to the four pairs of borderlines with which they were presented; only 1 (3%) participant gave dissociated judgments to all four pairs of borderlines.

Discussion of Experiment 3

As in experiment 2, this experiment suggests that the dissociation between categorization and similarity judgments reported in experiment 1 can be eliminated when judgments are given under a specific perspective.

While the pairwise comparisons support this interpretation, other comparisons suggest that the perspective has not exerted such a strong influence in this experiment as in experiment 2. Overall, the typicality scores do not differ significantly for the two borderlines cases, A+G− and A−G+ (1.72 and 1.71 respectively, $p = 0.96$); however, the two borderlines do differ in their categorization scores: 1.03 and 2.09 respectively, $t(34) = 2.55$, $p < 0.05$). Hence the borderlines in this experiment provide evidence contrary to similarity-based models, i.e., an increase in categorization is not matched by an increase in similarity. Nevertheless, relative to experiment 1, the perspective has served to eliminate the differences in typicality between the borderlines, even though differences in categorization remain.

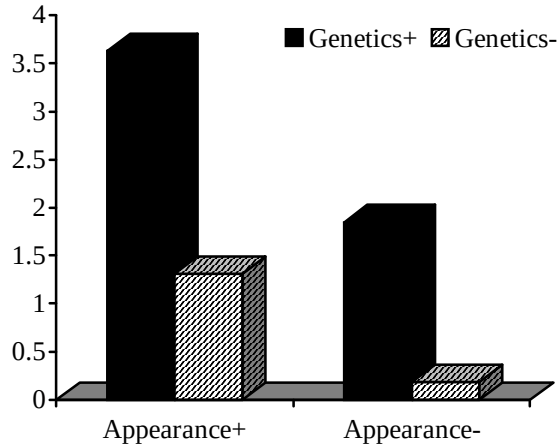


Figure 5. Overall ratings by Appearance and Genetics under a 'Biologist' perspective.

General Discussion

This paper provides evidence to support two main claims. The first is that similarity and categorization judgments dissociate for natural categories. The second is that such dissociations are perspective-dependent.

The data in experiment 1 reflect a double dissociation between similarity and categorization judgments, and serve to undermine similarity-based models of categorization. Though previous research has also uncovered dissociations, these have been single dissociations, and there have been difficulties in replicating those findings. Indeed, suggestions have been made that such dissociations arise only when categorization is highly reflective, only when designs are within-participant, and only for certain participants.

The data reported here contradict the first of these claims. That is, the categorization judgments obtained in these experiments have not been sought under a reflective mode – indeed response times show that participants take considerably less time to make these judgments than they do the corresponding typicality judgments. So, there is little evidence for these categorization judgments being particularly reflective.

These data confirm previous findings that dissociations arise because of a minority of participants.

The data do not speak to the claim that dissociations only arise in within-participant designs, although Thibaut et al.'s evidence contradicts such a claim. However, Thibaut et al.'s study arguably shows only that people can learn artificial categories for which similarity and categorization dissociations arise, not that these arise also for natural categories. This study appears to provide strong evidence that even for everyday, natural kind categories such dissociations arise. Moreover, the work presented here does not rely on identifying features that could be considered necessary or characteristic, a problem Thibaut et al. identify in previous work – indeed, the stimuli used here are defined without reference to particular features.

The second main claim is that by fixing perspective, dissociations between similarity and categorization judgments are reduced or eliminated. In experiment 2, making judgments from a 'sculptor' perspective eliminated the dissociation, whereas it was reduced in experiment 3.

How might these findings of perspective-dependence be explained? One possibility is that similarity-based models should be seen as models of categorization-in-context. Categorization and typicality judgments are often elicited out of context. Without the constraint of context, participants may call on different kinds of information to make the two kinds of judgment. In other words, models of categorization should first seek to model categorization-in-context and then attempt to explicate context-free categorization judgments. One possibility for such a perspectival account of concepts allows that categorization is similarity-based, but that the current perspective fixes the relevant dimensions to be used in the similarity computation (Braisby, 1998).

Another possible explanation is that in experiments 2 and 3, dissociations do not appear because the instructions used for categorization and similarity do not elicit those judgments. For example, it could be that both sets of instructions actually elicit a categorization judgment, and participants respond in the typicality task as best they can given that their judgment reflects a categorization, rather than an overt judgment of typicality. However, such an explanation is fraught with problems – for example, if conventional instructions do not determine the kind of judgment people make, then there is no obvious basis for deciding whether any previous research has really elicited categorization or typicality judgments.

In conclusion, the data reported here suggest that a simple similarity-based view of categorization is not right. However, when context is fixed, then the similarity-based models may fare much better. What is needed to augment such models is a mechanism by which the current perspective or context fixes the relevant dimensions on which categorizations and similarity judgments are made.

Acknowledgments

I would like to thank the Cognitive Science Group at the Open University for discussions of the ideas contained herein; any errors remain my own.

References

Barsalou, L. W., & Sewell D. R. (1984). Constructing representations of categories from different points of view. Emory Cognition Project Report No. 2. Emory University, Atlanta, GA.

Braisby, N. (1998). Compositionality and the modelling of complex concepts. *Minds and Machines, 8(4),* 479-508.

Braisby, N. (2001). Deference in Categorisation: Evidence for Essentialism? In Moore, J. D. & Stenning, K. (Eds.), Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society. Mahwah, NJ.: Lawrence Erlbaum.

Estes, Z. & Hampton, J. A. (2002). Similarity and essentialism in category judgments of natural kinds. Unpublished manuscript.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and projects.* Indianapolis, IN.: Bobbs-Merrill.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34,* 686–708.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65,* 137–165.

Kroska, A. & Goldstone, R. L. 1996). Dissociations in the similarity and categorization of emotions. *Cognition and Emotion, 10(1),* 27–45.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469-1481.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

Roberson, D., Davidoff, J., & Braisby, N. (1999). Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. Cognition, 71(1), 1–42.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime User's Guide. Pittsburgh: Psychology Software Tools, Inc.

Smith, E. E. & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22,* 525–538.

Thibaut, J-P, Dupont, M. & Anselme, P. (12002). Dissociations between categorization and similarity judgments as a result of learning feature distributions. *Memory & Cognition, 30(4),* 647–656.

A New Theory of the Representational Base of Consciousness

Andrew Brook {abrook@ccs.carleton.ca}
Paul Raymont {praymont@ccs.carleton.ca}
Cognitive Science
Carleton University
Ottawa, ON Canada

Abstract

Though we take mainly a philosophical approach, we hope that the results of our work will be useful to researchers on consciousness who take other approaches. Everyone agrees, no matter what their point of view on consciousness, that consciousness has a representational base. However, there have been relatively few well-worked-out attempts to say what this base might be like. The two best developed are perhaps the higher-order thought (HOT) and the transparency approaches. Both are lacking. Starting from the notion of a self-presenting representation, we develop an alternative view. In our view, a representation, a completely normal representation, is the representational base for not just for consciousness of its object (if it has one), but also itself and oneself as its subject. The unified picture of consciousness that results should assist research on consciousness.

Introduction

Though we take mainly a philosophical approach, we hope that the results of our work will be useful to researchers on consciousness using other approaches. Current views on consciousness can be divided by whether the theorist accepts or rejects cognitivism about consciousness. Cognitivism is the view that consciousness is just a form of representation or a property of information-processing systems that have representations (e.g., focussed attention). Anti-cognitivists deny this, claiming that inverted spectrum and zombie thought experiments show that consciousness could change while everything cognitive or representational stays the same. Whatever, researchers on both sides of this fence agrees that consciousness has a *representational base.* Whether or not consciousness *simply is* representational or cognitive, it at least *requires* representation and cognition.

However, there have been few well-worked out attempts to say what this representational base might be like. The two best developed are perhaps the higher-order thought (HOT) approach, in which the representational base of consciousness is a thought directed at one's own psychological state(s), and the transparency theory, in which one's conscious states are said to be, not objects of representation, but things that one knows about by inference from consciousness of the world around one, one's body, and so on.

As we will see, both approaches are lacking. We then introduce a notion of a self-presenting representation and attempt to build a better alternative around this notion. On our view, representations are self-presenting but more than that. A single representation is the basis not just of consciousness of the world and itself but also of oneself as its subject. This notion leads to a unified picture of consciousness.

The standard picture of representation

Many consciousness researchers accept the following as a principle of representation:

> RP: Representations represent something other than themselves and only something other than themselves.

If RP is right, our view is wrong and something like a HOT model or transparency model has to be right.

There are a host of problems facing HOT models (see Raymont, forthcoming). Perhaps the most serious arises from its separation of the representing state that confers consciousness from the state on which consciousness is conferred. The problem is that a representation can exist in the absence of its object. If so, a HOT that represents pain should be able to make things seem subjectively just as they would if one really were in pain – with no real pain. Rosenthal (1997, p. 744) at least bites this bullet but it is a pretty tough chew. Moreover, since in this case what is represented is not real, it is the representing state that has to be the conscious state. If so, there is nothing higher-order about consciousness. Indeed, the resulting conscious state would look remarkably like our self-presenting state.

So what about the transparency alternative? The basic idea behind transparency theory is that one is directly conscious only of what a representation represents and not the representation itself. We are conscious *via* representations, not *of* representations. Representations are transparent to us.

If so, consciousness of representing is an inference from the fact that we are conscious of what is represented. As Dretske puts it,

> You cannot represent something as F without, necessarily, occupying a state that carries the information that it is F (not G or H) that you are representing something as. [1995, p. 56]

All we know about our representing is what we can infer from how represented items appear. Dretske (p. 40) calls the resulting consciousness of our representing states *displaced perception*. The perception of an object is displaced by an inference onto the perception itself.

The transparency thesis faces some problems. First problem: when one is conscious of something by means of a given representation, one is thereby conscious of that repre-

sentation's content, i.e., of how that thing is represented by one's representation (visually, aurally, and so on). It is a *very* small step from this to consciousness *of* the representation. To take this step, various things may be needed but an inference from *what* one is representing does not seem to be one of them. One is conscious of not only *what* is represented but also *how* it's represented. The latter is an aspect of the representation itself.

Second problem: when one is conscious of *how* one is representing given contents (seeing them, imagining them, doubting them, remembering them), even Dretske allows that this knowledge does not come by way of an inference from representing something (1995, p. 57-8), though he does not say where it does come from. If transparency is not the case here, why anywhere?

Third problem: itches, pains and other bodily sensations. On the transparency view, feeling a pain, hurting, has to be nothing more than an inference from what the pain represents – some bodily damage or whatever. This is extremely implausible. (Dretske recognizes that he has a problem about pains, etc.: "this is a topic that I have neither the time nor (I admit) the resources to effectively pursue" [1995, p. 103].) All who accept RP face this problem.

Indeed, about the only phenomenon for which a transparency claim seems at all plausible is perception, especially vision, though even here transparency might have trouble with the difference between, say, seeing a corner and feeling it, in general with situations in which there is one content, two or more modes of representation. Put bluntly, the transparency thesis seems simply to be false, at least for most of consciousness.

As we said, if RP is right, something like a HOT model or transparency model are the only alternatives. If accepting RP leads to such problems, what happens if we deny it? It is not obviously true. Think again of pains and itches and other bodily sensations, not to mention feelings of tension and tiredness, mood states such as aimless anxiety or euphoria, and so on. What (other than themselves) are states like these about? Then there are altered states of consciousness. When consciousness is altered, what it is like to have those states certainly changes but there is no obvious candidate, other than the conscious states themselves, for what they are about.

There are two ways to reject RP. One would be to say that pains, etc., are not representations at all. This move is hopeless. In having pains (and mood states, altered states of consciousness, and so on), one is clearly aware *of something*, something about which a pain, for example, carries information. These are marks of representing. But what one becomes aware *of* in having a pain, what this state carries information *about*, is *itself*. Pains are representational by being *self*-representational; the qualities of which one becomes conscious by having these states are mainly qualities of the states themselves. So the other way of rejecting RP is to say that some representations at least are self-presenting.

Not only do we think that RP is false but our version of the opposing idea that representations can be self-presenting is fairly radical. Here is how our story goes.

Self-presenting Representations: The Representational Base of Consciousness

On our view, having a representation is all the representation that one needs to become conscious not only of what the representation is about and the representation itself, the standard view of self-presenting representations, but something more.[1] Consider the following sentence as uttered by MM:

1. I am reading the words on the screen in front of me.

Having the representation expressed by (1) can make MM conscious of the words on the screen, obviously. It can also make her conscious of the representation itself, that the words are being seen (not heard, imagined, touched, and so on). In addition, having this representation can make MM conscious of who is seeing the words, namely, herself. An ordinary single representation is all the representation that one needs to become conscious of that representation and also oneself as the subject of it. Let us call such a representation the *representational base* of becoming conscious of these items.

> **Representational base –** an act of representing that is all the representation that one needs to be conscious of it and of oneself as its subject

Almost any representation will do.

Imagining something unreal such as Pegasus will do just as well as perceiving an external object such as a computer screen. Indeed, even a representational state that had no object, and therefore could not make us conscious of anything other than itself, could still be the basis of becoming conscious of that state and of oneself. Moreover, a representation need not itself actually be recognized in order to provide a representational base for self-consciousness. Just recognizing what is represented in it would be an adequate representational base for one to be conscious of oneself (as conscious of that object).

Note carefully the term representational *base.* We are not saying that to have a representation is *to be* conscious of it. That would be a crazy view to hold. What we are saying is that having a representation provides everything *representational* needed to become conscious of having it and of oneself. Other things may be needed, too, shift of attention for example, or the conceptual resources to go from consciousness of something to consciousness of representing it.

Lest it be thought that the idea of a self-presenting representations is exotic, note that something as lowly as a bar code can be run quite a long way as an analogy. A bar code contains information about what it is 'about', usually the item's nature and price. But it also contains information about itself – a few of the bars are an integrity check on the

[1.] Among others, James held the standard view, as more recently have Kriegel (2003) and Tye (2003).

bar code itself. And it contains information about the thing that has it – it is physically mounted on the thing that has it. How far the analogy can be run does not matter here. What matters is that even a representation as simple as a bar code can be self-presenting.

Another homely example, a gauge, shows how our story goes.[2] A gauge presents information about something other than itself, namely, whatever it has the function of indicating information about. For example, an altimeter presents information about the distance to the earth's surface. However, an altimeter also presents information about itself, how far it is from earth, for example. And it is the gauge that presents this information about the gauge, not some higher-order gauge pointed at it! This is how we see conscious states. To be conscious of having a representation, all the representation that one needs is the state itself.

For the analogy of the gauge to be complete, the gauge would have to have one more function, It would have to represent the system that has it. Not a problem. Suppose that to provide information about altitude and itself, an altimeter has to port itself to a system. And suppose that to do so correctly, it has to recognize what sort of system it has been installed in. ('Ah, this is a Cessna Skylane.') Now we have at least a rough analogue of a representation presenting not just its object and itself but also its subject, the person who has it.

We don't have room to mount the full case for our notion of the representational base of consciousness but notice, if it is right, it would do some real work for us.

- It would entail that RP is false.
- It would entail that the idea behind transparency and displaced perception, the idea that we are not conscious of our own conscious states, is false and that the displaced perception move is unnecessary – the consciousness one has of one's own representations by having them is as direct and non-inferential as any consciousness of anything.

And,

- It would show that the HOT move is unnecessary. If a representation itself is all the representation we need to become conscious of that representation, there is no need for a higher-order representation of any kind.

Global Representation

Our notion of the representational base gives a single, unified account of the basis of three forms of consciousness – consciousness of the world, consciousness of one's own states, and consciousness of oneself. This nicely unified account is progress. Progress – but not the whole story.

So far we have talked exclusively about individual representations as understood by the tradition. As Kant already knew, however, the representations that serve as the representational base of consciousness are usually much 'bigger' than individual representations traditionally conceived. Indeed, in complex beings like us, we do not believe that there are any individual representations as traditionally conceived; but we will not go into that here.

The representations serving as the representational base of consciousness usually have multiple objects and encompass multiple representations (as traditionally conceived). Let us call such a representation a *global representation.* In a global representation, one is conscious of many objects and/or many representations as traditionally conceived, and one is conscious of them as a single complex object and/or a single complex representation.

> **Global representation** – representing many objects and/or many representations as traditionally conceived as a single complex object and/or a single representation.

Our points about the representational base can now be made using this notion. A global representation is all the representation that one needs to be conscious not just of its complex object, if it has one[3], but also of the representation itself, and of oneself as the 'the single common subject' of the elements of this representation (Kant, 1781/7, A350).

The structure of a global representation is complicated. Suppose that at the same time as one has the representation expressed by (1),

1. I am reading the words on the screen in front of me,

one also has representations expressed by,

2. I am puzzled by your comments
3. I am enjoying the music I hear outside
4. I believe our agreement was to meet at 6:00
5. Yesterday I thought I understood Kant's notion of the object
6. I wish the world were a fairer place

There are three elements of (1)-(6) that could be united in a single global representation.

- One could be conscious of the various represented objects here as a single complex object.
- One could be conscious of the various ways in which these objects are being represented as a single complex representation.

And,

- One could be conscious of oneself, the subject, as the single common subject of the whole business.

The next question is, How does a global representation serve as the representational base of consciousness of its complex object, itself, and its subject, the person who has it?

Joint Consciousness

Central to a global representation of objects is what we will

---

2. A gauge is one of Dretske's favourite examples.

3. Must global representations have objects (other than themselves)? Nice question; but not one for us here.

call *joint consciousness:*

> **Joint consciousness –** to be conscious of any of the objects of a global representation is to be conscious of other such objects.

It seems obvious that joint consciousness is a distinctive feature of a global representation. The notion of joint consciousness just stated clearly applies to consciousness of the world (plus such things as one's own bodily states), more exactly, to consciousness of intentional objects. What about consciousness of one's global representation and of oneself as its subject?

When one is conscious of representing in a global representation, here too there will be joint consciousness – to be conscious of some representings is to be conscious of others. And conscious of self? Here the idea that is plausible is a bit different. When one is conscious of oneself as subject of one bit of representing, one will usually be conscious of oneself as the subject of other representings, as their 'single, common subject', to use Kant's words again.

The next question. How could a global representation serve as the representational base of *joint* consciousness of representings and of oneself as their *common* subject? It seems plausible to hold that only certain elements of a global representation are needed for one to have an adequate representational base for consciousness of a single representing and of oneself as its subject. One will be conscious of doing an act of seeing by doing that act of seeing, whatever the rest of one's current global representation is like. And one will conscious of oneself *seeing*, which makes it likely that the act of seeing is the basis of this consciousness of oneself. If so, what is the representational base of *joint* consciousness of objects, of being *jointly* conscious of various representings, of being conscious of oneself as the *common* subject of a number of acts of representing?

Here we can only sketch what would have to happen for these forms of joint consciousness to occur. In the same way as an act of representing is the representational base for consciousness of that act and of oneself, it will have to be the representational base for consciousness of doing *other* acts of representing and of oneself as their *common* subject.

Moreover, the relationship will be symmetrical. One is not 'located' at any given representing. When consciousness of a representing carries one to consciousness of other representings, one has all the representings concerned equally. So it would be better to say that representings in general are the base for consciousness of representings in general. It is not any given representing that is the base of consciousness of oneself as common subject but representings in general.

Thinking of representations in the traditional way, it is hard to make sense of what we just said. But that may be the fault of the traditional conception. There is another way to think about representation within which what we just said makes fine sense.

Structure of a Global Representation

Instead of trying to make sense of the base of joint consciousness of representing and of oneself as subject in terms of relations *among* representations, let us try out the idea that these forms of consciousness are something that obtains *within* a single complex representation.

Our test case will be a person seeing something, hearing something, and tasting something, all parts of a global representation and global object, where the person is jointly conscious of the objects, the representings, and herself as subject. To be the base for such consciousness, how are the three acts of representing brought together in a global representation? Here are three possibilities:

1. The three acts and their objects become the object of a fourth, higher-order representation.

2. The three acts and their objects become parts of a single subsuming representation.

3. While their contents are taken up in a global representation, the three acts of representing do not survive even as parts of this state, though their objects remain distinct. They become three (for the moment let us call them) *modalities* of a single representation.

How might (3) work?

Consider what happens if one goes from a situation, at time *t*, that has objects *o*1 and *o*2 to a situation that has *o*1 but not *o*2. We could try to capture the change in two ways. We could say that where once there were two representations, r(*o*1) and r(*o*2), which were bundled in a mental structure of some kind, [r(*o*1) & r(*o*2)], we now have only the one representation, r(*o*1); one representation has been dropped from the bundle that existed at *t*. Or we could say that there was just one representation, r(*o*1 & *o*2), at *t* and it has been replaced by another single representation, r(*o*1).

The second view is simpler, since it does not involve postulating representations as parts of an encompassing representation. According to it, at *t* there was one representation that had a complex content. The content was complex because it had multiple contents, *o*1 and *o*2, as its parts. If it were a conscious representation, what would make it one representation is that to be conscious of any of its objects by means of it is to be conscious of other of its objects, too. Here, the part-whole relation obtains among objects, but there is no parallel multiplicity of representational states. On this approach, unlike the first, the representing state does not have 'smaller' or less complex representing states as parts.[4]

On this picture, globality obtains *within* a representation but not *among* representations. It does obtain among objects. Returning to the joint consciousness condition, where to be conscious of one thing is to be conscious of others, on our picture joint consciousness of objects can be present or absent but it is trivially present in a global representation,

---

[4.] If the representing state is a brain state, then it will have parts, but these parts will not be representing states.

simply because a global representation is a single representing state, r(o1 & o2). For conscious representing states of of this form, to be conscious of o1 by means of this state is also, by that very same act, to be conscious of o2. One representational state-token provides consciousness of both o1 and o2.

Consider an act of reference as an analogy. Suppose I refer to Toronto. Scarborough is part of that city; it is a part of the thing to which I referred. It does not follow that my act of referring to Toronto contains a numerically distinct reference to Scarborough. It was of course *possible* for me to refer simply to Scarborough, and thus to refer to part of the thing to which I actually referred, but it does not follow that I actually did so. The mere fact that Scarborough is part of the thing to which I referred does not entail that a reference to that borough figures as part of my act of referring. Similarly, the mere fact that o1 is part of (o1 & o2) does not entail that a distinct representation of o1 must figure as part of my act of representing (o1 & o2).

It may be objected that in advancing these observations and contrary to our intentions, we actually *provide the resources* for showing that a subject's global representation will contain parts that are themselves representations. Didn't we ourselves just say that representations are individuated by their objects? This suggests that for each object that we can individuate, including objects that are parts of a global object, there will be a corresponding representation individuated by that object.

This objection does not work. If we have relied on a claim that each representation is individuated by an object, we have not said how objects individuate a representation. What if only some objects individuate a representation? Let us introduce the idea of a *global object.*

> **Global object –** a group of represented objects that is the single complex object of a global representation

If we now say that only a global object individuates at least a conscious representation, we are clear of the objection.

To capture these claims about singularity of representation, let us generalize (3),

> 3. While their contents are taken up in a global representation, the three acts of representing [viz., seeing something, hearing something, and tasting something] do not survive even as parts of this state, though their objects remain distinct. They become three modalities of a single representation.

into (3′),

> 3′. A global representation at a given moment is a single representation not made up of multiple distinct representations and it has a complex object.[5]

Now that we have said what we mean by (3) and (3′), what about (1) and (2)? Though most philosophers hold to (1) or (2), they do so uncritically and seldom offer any support for the views. We do not know of any good argument for either one of them. To see where those adopting (1) or (2) might go wrong, recall the representation, r(o1 & o2).

Suppose that this representation occurs at t. Supporters of (1) and/or (2) may confuse the *untokened* (at t) representational type that would take part of r(o1 & o2)'s object (*viz.* o1) as its object, with a representational token, r(o1), held to be *part* of r(o1 & o2) and to have one of the latter state's objects as its sole object. There is no reason to suppose that there actually is such a distinct representational state nestled within r(o1 & o2). There could have been such a token at t, but the mere fact that we can entertain such a possibility – that is, the mere fact that we can think of an instantiation at t of that type – is no reason to conclude that there actually exists a token of that type at t.

Conclusion: (3′), the idea that at a given moment a global representation is not a group of representations but a single representation with a complex object, is a perfectly coherent point of view, one well supported by examples and analogies. Before we leave it, however, there are some objections that we need to answer. One of them is so obvious that it has probably occurred to most readers already.

Single Representation View: Objections

The obvious objection is this. In a global representation, we are conscious, within a single unified representation, of several sensory modalities. The phenomenal field is polymodal: it involves tactual data, visual data, auditory data, and so on.

(1) or (2) would try to account for the contributions of the different modalities to consciousness by saying that there are several distinct representations here – visual representations, auditory representations, and so on – that come together as parts of an encompassing global representation. On this view, the cognitive system constructs a variety of representations in different modalities. These representations are not simply superseded by a global state that combines their informational contributions in one representation. Instead, they are preserved as distinct representations within it. The polymodal complexity of the resulting global state is due to the presence in it of this range of representations.

Against this, we offer the following picture: we do not have several visual, aural, etc. representations, not conscious ones anyway. Rather, the information we represent is *formatted* visually, aurally, etc. The cognitive system receives some information in a visual format (reflecting, perhaps, the wave length of incoming energy) or tagged as visual, some formatted or tagged as aural, and so on. When this information appears in a global representation, its modality appears with it. But there is just one representation, the global representation. In this one state, diverse bits of information are formatted in a variety of modalities.

This view has the twin virtues of adequacy and parsimony. The onus rests with proponents of the more complicated view, in which a global representation is held to be an assemblage of other representations, to show that our account has failed to account for something. It is difficult to

[5] James (1890, vol. 1, pp. 145-61) held something like this view.

see what that could be.

Suppose that one is consciously representing things both visually and aurally. What accounts for something there being aural? This question must be answered either in terms of *how* the representation represents or in terms of *what* it represents. There are no other ways to specify the modality of a representation. But a global representation can incorporate either of these ways. There can be a diversity in what it represents; it can represent many things having many properties. And it can represent these contents in a variety of ways, visually, aurally, and so on.

In support of this contention, note that we are *compelled* to postulate a plurality of representations only when there is a particularly strong sort of incompatibility in how or what a cognitive system represents, that is, only when the system represents in ways that exclude one another. For example, theorists tend to attribute distinct representations (not all of which are held to be conscious) to a system in such cases as binocular rivalry, or to explain the fact that consciously seeing the Necker Cube in one way precludes consciously seeing it in the alternative way.

These cases and all others that we've been able to think of are not problems for us. To be a counter-example to our single representation/complex object picture, the elements would have to be: all conscious, and all available simultaneously to one conscious subject. The cases we've just considered do not meet these conditions.

In the Necker cube case, the conflicting representations are successive. In binocular rivalry, the representations may be simultaneous but one is not simultaneously conscious of them (Baars 1988, pp. 82-3; see p. 126). So even when we are driven to multiply representations in a subject at a time, the results are not a problem for our view.

Notice that cases such as binocular rivalry and the Necker cube arise *within* a modality. It is not clear that there are any strong incompatibilities across modalities. Certainly there aren't many. In the absence of strong incompatibilities, nothing compels us to posit *more than one* representation to make a place for different perceptual modalities.

Second objection. Think of a picture of a car in front of a house. It is plausible to say that this picture includes a picture of the car and a picture of the house, that the bigger picture of the car and house together contains little pictures of the car and of the house.

Is it so clear, though, that the picture of house and car together literally contains several distinct *pictures*, one for each item depicted? The belief that it does threatens to introduce an implausible multiplication of pictures. If one can discern ten thousand blades of grass in front of the house, then there would have to be ten thousand pictures in the larger picture. Clearly, at some level of decomposition, we stop positing a distinct picture for each part of the content that we are able to distinguish. Why not stop at the whole picture and say that it is the only picture, with no smaller pictures in it?

Third objection: If conscious representation consists of one big, non-compositional representation, how are acts of judging particular bits of content, forming beliefs about things based on particular bits of content, possible? Response: In the same way as information in the complex global object of a global representation can come 'marked' with various modalities (aural, visual, etc.), particular bits of the information in a complex global object can enter into particular information-processing activities: judging, remembering, and so on. These *activities* do not need to merge into a single representation and they can pick and choose information to work on *ad libitum*.

Even after responding to these objections, the support we have offered for (3´) is not decisive. It is strong, however, certainly strong enough to justify acceptance of the view.

Our theory that a single global representation is the representational base of consciousness has real potential. As we have seen, it provides a unified account of three major kinds of consciousness, consciousness of world, one's own representations, and oneself as subject. Though we can't show this here, it also opens the way to a nice account of: the unity of consciousness; the special features of consciousness of self; and the subject of consciousness (Brook and Raymont, forthcoming).

References

Baars, B. 1988, *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Brook, A. and P. Raymont. forthcoming. *A Unified Theory of Consciousness.* Cambridge, MA: MIT Press.

Dennett, D. 1978e. Mind writing and brain reading. In *Brainstorms.* Montgomery, VT: Bradford Books, pp. 39-50.

Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press/A Bradford Book.

Kant, I. (1781/1787) *Critique of Pure Reason* (trans. P. Guyer and A. Woods). Cambridge: Cambridge University Press, 1997. (References are in the pagination of the 1st (A) or 2nd (B) editions.)

Kriegel, U. 2003. Consciousness as intransitive self-consciousness. *Canadian Journal of Philosophy* 33:103-132.

James, W. 1890. *Principles of Psychology.* London: Macmillan.

Raymont, P. submitted. From HOTs to self-representing states

Rosenthal, D. 1997. A theory of consciousness. In N. Block, O. Flanagan, and G. Gűzeldere, eds. *The Nature of Consciousness.* Cambridge, MA: MIT Press.

Tye, M. 2003. *Consciousness and Persons: Unity and Identity*. Cambridge, MA: MIT Press/A Bradford Book.

Coherence in Perceptions of a Romantic Relationship

Aaron L. Brownstein (aaronb@usc.edu)
Department of Psychology, University of Southern California, Los Angeles, CA, 90089-1061

Stephen J. Read (read@usc.edu)
Department of Psychology, University of Southern California, Los Angeles, CA, 90089-1061

Dan Simon (dsimon@law.usc.edu)
School of Law, University of Southern California, Los Angeles, CA, 90089-0071

Abstract

When decision makers perceive all issues related to a decision as being consistent with their choice, they achieve *coherence*. Participants rated their agreement with different views of various issues related to a couple's relationship. Those who later decided whether the couple would get engaged or break up subsequently reinterpreted the issues to be consistent with their decision. Increasing the importance of the decision, highlighting coherent perspectives, or giving participants a prior preference did not strengthen the coherence shift, but coherence shifts did not occur without the chance to decide, suggesting that they occur in an all or nothing fashion. Individuals with higher need for cognition (Cacioppo & Petty, 1982) achieved a greater degree of coherence among facts associated with the relationship but not among more general beliefs about relationships.

Introduction

Early social psychologists attempted to develop a general model of cognitive functioning based on consistency maintenance (Festinger, 1957; Heider, 1946; Osgood & Tannenbaum, 1955; Newcomb, 1953). Those theories shared the assumption that cognition involves the interaction among elements, and that those elements tend to settle into stable states characterized by some type of *good form*, in which similar elements are interconnected and segregated from dissimilar elements. However, early consistency theories narrowly focused on small sets of two (Festinger, 1957) or three (Heider, 1946) elements at a time, and so were unable to represent the larger, more complex situations that people often encounter in their daily lives.

More recently, computer based models of multiple constraint satisfaction systems have begun to provide a mechanism for simulating maintenance of consistency throughout large, complex systems (Holyoak & Simon, 1999; Read, Vanman, & Miller, 1997; Simon, Snow, & Read, in press). In these models, units represent cognitive elements and links between units represent relations between elements, with excitatory links representing consistent relations and inhibitory links representing inconsistent relations. Dynamic processing is simulated by allowing units connected by excitatory links to increase each other's activation and units connected by inhibitory links to decrease each other's activation until the system settles in to a stable state of *coherence*, a kind of good form in which a subset of mutually consistent elements are highly activated.

Drawing on constraint satisfaction systems, researchers have begun to show that cognition involves imposing consistency on related concepts. Simon and colleagues (Holyoak & Simon, 1999; Simon, Pham, Le, & Holyoak, 2001; Simon, Snow, & Read, in press) developed a paradigm showing how, when people think about the issues related to a legal case and then render a verdict, their perceptions of the issues shift to become consistent with their eventual verdict, thereby achieving a coherent understanding of the whole case.

The finding that perceptions of issues shift to become consistent with an emerging decision violates two important assumptions of algebraic models of information integration, such as Bayes theorem and Anderson's (1962) Information Integration Theory. Those models assume that (1) the value of one element is not affected by the values of other elements and that (2) the value of an element is not changed when it is integrated with other elements to arrive at a conclusion. Nevertheless, Simon and colleagues found that during decision making, evaluations of issues related to a legal case shift to become consistent with the emerging decision and with each other.

The present research adapts Simon and colleagues' paradigm to test for coherence outside the legal context. In the first phase (pretest) participants read vignettes describing different couples and ambiguous events in their relationships, and then rated their agreement with statements giving different interpretations of the events. Some of the statements were *factual* items that involved interpreting the meaning of an event and others were *belief* items that involved interpreting the general implications of an event. For example, one vignette described how Eric and Daniella spent a day with Daniella's aunt Rachel, enabling her to observe the couple's interactions and subsequently report that she thought their relationship was going well. A factual item related to that vignette asks participants to rate the extent to which they think Aunt Rachel's optimistic impression was correct and a belief item asks them to rate the extent to which they think that, in general, it is possible to get a good sense of a couple's relationship by observing them for a day.

In the second phase participants read a longer story about a couple, Jenny and Mark, that combined all of the issues

raised separately in the pretest. For example, one part of the story involved Jenny and Mark spending a day with Jenny's aunt Rose, so that the aunt formed an impression that the relationship was going well. In the third phase (posttest) participants decided whether Jenny and Mark would get engaged or break up and then rated their agreement with the different interpretations of the issues, now that they were all embedded in the context of a single story. For example, on the posttest a factual item asked participants to rate the extent to which they thought Aunt Rose's optimistic impression of Jenny and Mark's relationship was correct and a belief item asked them to rate the extent to which they think it is possible to get a good sense of a couple's relationship by observing them for a day.

We predicted a *coherence shift*, so that from pretest to posttest interpretations of ambiguous issues would shift to become consistent with the decision. Thus, we expected participants who decided that Jenny and Mark would get engaged to be increasingly likely to interpret issues in a manner suggesting that they would stay together, and we expected participants who decided that they would break up to be increasingly likely to interpret issues in a manner suggesting that they would not stay together.

Simon, Snow, & Read (in press) tested several variations on their paradigm in a legal context, but could not affect the strength of the coherence shift, so in the present research we introduced four manipulations designed to moderate the strength of the coherence shift. First, we tried to increase the strength of the coherence shift by increasing the perceived importance of the decision. In a *decision* (control) condition, participants simply read about Jenny and Mark's relationship and indicated whether they thought the couple would get engaged or break up. To increase the perceived importance of the decision, we asked participants in a *gift* condition to decide not simply whether they thought the couple would get engaged or break up, but whether they would spend a substantial amount of money to buy an engagement present. We predicted that the added importance associated with the decision would induce participants to think about the choice more extensively, leading to stronger coherence shifts than in the control condition.

The second way we tried to increase the strength of the coherence shift was by outlining the two coherent perspectives on the story. As in the control condition, participants were asked to decide whether they thought the couple would get engaged or break up, but in an *outline* condition they were also asked to imagine that they had talked to two friends who gave their perspectives on the couple's relationship. The friends' perspectives were presented in two lists, with one interpreting each of the ambiguous issues as suggesting that the couple would get engaged and the other interpreting each issue as suggesting that they would break up. We predicted that outlining the two coherent perspectives would help participants reach their own coherent perspective more quickly and proceed to achieve more extensive coherence shifts than participants in the control condition.

The third way we tried to increase the coherence shift was by giving participants a prior preference for one of the alternatives. Russo, Medvec, and Meloy (1996) gave participants an extraneous reason to prefer one of a pair of alternatives (called an "endowment") and then presented information on the alternatives one attribute at a time, asking participants to rate the extent to which it favored one alternative or the other, until they were ready to choose one. Russo, Medvec, and Meloy found that participants "distorted" the attribute information so that it favored the endowed alternative, thereby further increasing their preference for it until they chose it. We thought that if decision makers distort information to favor a prior preference within the pre-decision phase, their attitudes might also start shifting to become consistent with a prior preference within the pre-decision phase, producing a stronger coherence shift by the time they reach the post-decision phase. Therefore, we introduced two new conditions in which participants were given a prior preference for one alternative. As in the control condition, participants were asked to decide whether they thought the couple would get engaged or break up, but in an *endowment-engage* condition they were also asked to imagine that they knew Jenny and Mark and thought they *should* stay together, while in an *endowment-breakup* condition they were asked to assume that they thought Jenny and Mark *should* break up. We predicted that, compared to participants in the control condition, participants in the endowment-engage condition would be more likely to decide that the couple would in fact get engaged, participants in the endowment-breakup condition would be more likely to decide that the couple would break up, and that participants in both endowment conditions would report stronger coherence shifts.

A fourth manipulation was designed to decrease the strength of the coherence shift. We thought that, just as the degree of importance of a decision may affect the amount of processing, the degree of involvement in a decision task may also affect the amount of processing and the strength of the coherence shift. Therefore, we predicted that decreasing participants' involvement in the decision task would decrease the strength of the coherence shift. In a pair of *assigned decision* conditions participants were asked to think about whether to buy an engagement present (as in the gift condition), but they were not allowed to make their own choices; instead, participants in an *assigned-buy* condition were asked to assume that they had decided to buy the necklace and participants in an *assigned-not-buy* condition were asked to assume that they had decided not to buy the necklace. We predicted that depriving participants of the ability to reach their own decisions would decrease their involvement and amount of processing, leading to weaker coherence shifts in the assigned decision conditions than in the gift condition.

We also tested whether two personality dimensions moderate the strength of the coherence shift. First, need for cognition (NFC; Cacioppo & Petty, 1982) involves individual differences in the "tendency to engage in and

enjoy thinking" (p. 116). We thought that participants with greater NFC would think about the experimental materials more extensively, leading them to report stronger coherence shifts than participants with lower NFC. Second, personal need for structure (PNS; Neuberg & Newsom. 1993) may involve individual differences in desire for simple structure. We thought that participants with greater NFS would be more likely to impose a coherent good form on the information given to them, thereby achieving greater coherence than participants with lower NFS.

Methods

Participants and Design

This experiment ran on the Internet. Participants were recruited by e-mailing notices to people who asked to be notified of new experiments and giving each participant an entry in a lottery for a cash prize. Participants were randomly assigned to the decision (n = 134 usable data sets), gift (108), outline (118), assigned-buy (105), assigned-not-buy (132), endowment-engaged (110), and endowment-breakup (98) conditions.

Materials and Procedure

The pre-test had 12 vignettes, including seven that involved romantic relationships and five that involved legal cases (distracters). The seven relationship vignettes concerned (1) a woman (Michelle) who had broken off several previous relationships and may appear to have "a problem with commitment," (2) a woman (Joanne) who declined to talk to her boyfriend about the future of their relationship, (3) a woman (Aunt Rachel) who spent a day with her niece and her niece's boyfriend and thought that their relationship was going well, (4) a woman (Lisa) who was two hours late for a date with her current boyfriend because she was consoling a previous boyfriend who was upset about his mother's illness, (5) a woman (Suzy) who was too busy at work to join her boyfriend and his parents for dinner, (6) a woman (Candice) who joined a gym after her boyfriend disparaged people who don't exercise, and (7) a woman (Rona) who brought her boyfriend to a family party.

After each vignette, there were 1 to 4 statements interpreting the facts in the vignette or expressing related beliefs. Some of the statements expressed attitudes consistent with the view that the couple in the vignette would stay together (e.g. *Aunt Rachel's favorable view of the relationship was correct*) and other statements expressed attitudes consistent with the view that they would break up (e.g. *Aunt Rachel's favorable view of the relationship was influenced by the fact that on that day Daniella displayed particular affection towards Eric*). Participants rated the extent to which they agreed with each statement on an 11-point scale ranging from -5 (strongly disagree) to 0 (neutral) to 5 (strongly agree).

The next page presented a set of analogy word games (distracter task) and the following page introduced the experimental manipulations. In the decision condition, participants were asked to imagine that "Mark and Jenny live in your town and are pretty close friends of yours. They have been involved in a relationship for over a year…." Participants were also informed that "after reading some information about the relationship, you will be asked to decide whether you think Jenny and Mark will get engaged or break up, and then to make some evaluations about the relationship…." Participants then read the story of Jenny and Mark, which combined the seven issues that had been raised in separate vignettes in the pretest. For example, the story described one incident when Jenny was too busy at work to join Mark and his parents for dinner and another incident when she brought him to a family party. After reading the story participants indicated whether they thought Jenny and Mark would get engaged or break up (by clicking on one of two radio buttons) and rated their confidence that they had made the best possible decision (5-point scale). The posttest appeared next; participants were asked to give their "impressions of the issues in Jenny and Mark's relationship" and then there was a list of statements interpreting the facts in the story or expressing related beliefs. Some of the statements expressed attitudes consistent with the view that Jenny and Mark would stay together (e.g. *Aunt Rose's favorable view of the relationship was correct*) and other statements expressed attitudes consistent with the view that they would break up (e.g. *Aunt Rose's favorable view of the relationship was influenced by the fact that on that day Jenny displayed particular affection towards Mark*). Participants rated the extent to which they agreed with each statement on an 11-point scale. The order of items in the posttest was counterbalanced between participants. The posttest was followed by self-report measures of NFC (Cacioppo & Petty, 1982) and NFS (Neuberg & Newsom, 1993), and demographic questions.

The outline condition was the same as the decision condition, except that after reading the story participants were asked to imagine that, "you have talked to two other mutual friends and found that they have very different views on Jenny and Mark's relationship." The instructions continued, "Caitlin doesn't think Jenny and Mark will get engaged. When you talked to Caitlin, she explained why she thinks Jenny and Mark are headed for a breakup" and then there was a bullet-list of statements interpreting the seven issues in a manner suggesting that the couple would break up. The instructions then continued, "Unlike Caitlin, Brian thinks Jenny and Mark will get engaged. When you talked to Brian, he explained why he thinks they will get engaged," and then there was a list of statements interpreting the seven issues as suggesting that the couple would stay together. After reading the lists, participants made their decisions, confidence ratings, posttest ratings, and personality ratings.

The endowment conditions were the same as the decision condition, except that before reading the story participants were told that "we're going to ask you to get more 'involved' with the story, by imagining that you know the people and playing a small role yourself." Participants in the endowment-engage condition were told that "although they have had some problems (as most couples do), you think they are good for each other and you hope they work things

out," while participants in the endowment-breakup condition were told that "although they have been fairy happy together until now, you think that they may be growing apart and it wouldn't be a good idea for them to rush into a commitment." After reading the story participants in the endowment-engaged condition were reminded, "you don't know whether Jenny and Mark will get engaged or break up, but you think they are good for each other and you'd like to see their relationship work out" while participants in the endowment-breakup condition were reminded, "you don't know whether Jenny and Mark will get engaged or break up, but you think that they are growing apart and shouldn't rush into a commitment." Participants in both conditions then went on to make their decisions, confidence ratings, posttest ratings, and personality ratings.

In the gift condition, participants were told that Jenny loved *Zapotec* jewelry, which was available in Cancun, where the participant was vacationing, so the decision was framed in terms of whether to buy a $150 Zapotec necklace as an engagement present for Jenny; since "the necklace cannot be returned and you cannot think of anything else to do with it," the participant should only buy the necklace if an engagement seemed likely. After reading the story, participants were reminded of their dilemma – whether the probability of an engagement was high enough to justify buying an expensive necklace as an engagement present for Jenny – and then they indicated whether they would buy the necklace. After that, participants made their confidence ratings, posttest ratings, and personality ratings.

The assigned decision conditions were the same as the gift condition, except that before reading the story, participants were asked to "get more 'involved' with the story, by imagining that you know the people and playing a small role yourself" and were instructed, "while you're reading, think about your character's dilemma – whether to buy the necklace for Jenny. At the end of the story we'll tell you what your character decided." After reading the story, participants were reminded of their character's dilemma and those in the assigned-buy condition were asked to imagine that "you finally decided to go ahead and buy the necklace…. your sense was that they probably will get engaged, so it was worth it to buy Jenny an engagement gift you know she'll love" while those in the assigned-not-buy condition were asked to imagine that "you finally decided against buying the necklace… your sense was that they probably won't get engaged, so it wasn't worth it to spend so much money on a necklace you have no use for." Participants in both conditions then completed the posttest and personality measures.

Results and Discussion

Collapsing across all conditions (excluding the assigned decision conditions), participants were evenly split between deciding that Jenny and Mark would get engaged (49.5%) or that they would break up (49.8%), suggesting that the story was highly ambiguous. Average confidence ratings were fairly high among participants who decided that the couple would get engaged (3.52) and those who decided they would break up (3.77). High confidence in a decision about an ambiguous situation is consistent with constraint satisfaction models, in which activation spreads until the system reaches a stable state of coherence.

In the control condition participants were evenly split between the alternatives (50.7% chose the *engaged* alternative). In the outline condition there was a tendency to favor *engaged* (61.0%) but the change from the control condition was not significant. Participants in the gift condition were significantly less likely to choose *engaged* (36.1%) than participants in the control condition, X^2 (1, 239) = 4.90, p < .03, suggesting that people become more cautious when a decision has financial implications. Participants who were given an endowment favoring engagement were more likely to choose *engaged* (65.4%) than participants who were given an endowment favoring breakup (40.8%), X^2 (1, 207) = 4.72, p < .04, suggesting that the endowment manipulation was effective, though the probability of choosing *engaged* was not significantly different in the endowment conditions than in the control condition. Scores on the NFC, p = .55, and PNS, p = .98, were not significantly correlated with decisions.

We first tested for an overall coherence shift by running a 2(pretest, posttest, within Ss) by 2(engagement, breakup items, within Ss) by 2(engage, breakup decision, between Ss) ANOVA, collapsing across the experimental conditions. We found a significant three-way interaction, $F(1, 799) = 84.51$, $p < .001$, suggesting that after participants decided whether they thought Jenny and Mark would get engaged or break up, their attitudes shifted to be more consistent with their decision. As shown in Figure 1, among participants who decided that the couple would get engaged, agreement with statements suggesting that the couple would get engaged increased from pretest to posttest, $t(385) = 3.94$, $p < .001$, and agreement with statements suggesting they would break up decreased, $t(385) = 4.46$, $p < .001$. In contrast, among participants who decided that the couple would break up, agreement with engagement statements decreased, $t(414) = 6.47$, $p < .001$, and agreement with breakup statements increased, $t(414) = 5.86$, $p < .001$.

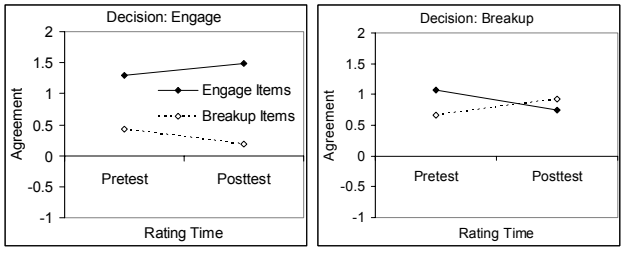


Figure 1: Agreement with Engage and Breakup items at Pretest and Posttest among participants who decided Engage (left) and participants who decided Breakup (right).

This pattern of results suggests that as participants thought about the story and made their decisions, their attitudes about the issues shifted to become more consistent

with their emerging decision and with each other. This would not have been predicted by algebraic information integration models (e.g. Anderson, 1962; Bayes' theorem), which assume that the value of an element is not affected by the values of other elements and does not change when it is integrated with other elements.

We then ran the analysis separately for items relating to facts and items relating to beliefs. For facts the three-way interaction was significant, $F(1, 799) = 73.75$, $p < .001$, and the pattern was similar to the overall analysis (Figure 2). Agreement with engagement items was initially higher than agreement with breakup items among participants who chose *engage*, $t(385) = 15.80$, $p < .001$, and among participants who chose *breakup*, $t(414) = 9.75$, $p < .001$, but among participants who chose *engage* agreement with the two types of items spread apart from pretest to posttest, while among participants who chose *breakup* agreement with the two types of items converged.

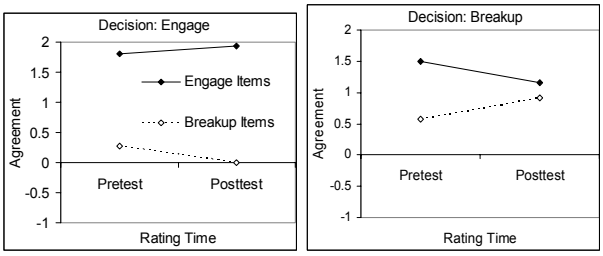


Figure 2: Agreement with Engage and Breakup fact items at Pretest and Posttest among participants who decided Engage (left) and participants who decided Breakup (right).

For beliefs the three-way interaction was significant, $F(1, 797) = 30.97$, $p < .001$, indicating the predicted coherence shift, but the pattern was somewhat different than for the facts (Figure 3). Agreement with engagement items was initially lower than agreement with breakup items among participants who chose engage, $t(385) = 6.60$, $p < .001$, and among participants who chose breakup, $t(414) = 6.72$, $p < .001$. Among participants who chose *engage* agreement with the two types of items converged, while among participants who chose *breakup* agreement with the two types of items spread apart.

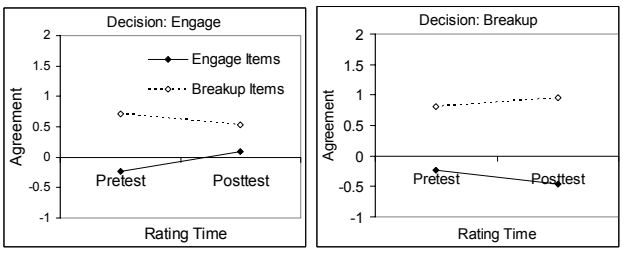


Figure 3: Agreement with Engage and Breakup belief items at Pretest and Posttest among participants who decided Engage (left) and participants who decided Breakup (right).

The unexpected difference in the patterns of results for facts and beliefs reveals that when people consider the facts of a specific couple's relationship they are initially optimistic about their future together, and those who go on to decide that the couple will get engaged become increasingly optimistic, but those who decide that the couple will break up become less optimistic. However, when it comes to abstract beliefs about relationships in general, people are initially pessimistic, and if they decide that a specific couple is likely to break up their beliefs become even more pessimistic, but if they decide that they are likely to get engaged their pessimism decreases. These data suggest that people are initially pessimistic about relationships in general but initially assume an optimistic outlook on specific cases. The data also suggest that regardless of initial attitude, when people reach a decision about a specific case their attitudes toward facts related to the case as well as their general beliefs about relationships shift in a coherent manner.

To test whether the experimental manipulations affected the strength of the coherence shift, we added a variable representing experimental condition, but did not find that it moderated the three-way interaction. We found three-way interactions of similar magnitude within the control, $F(1, 131) = 47.44$, $p < .001$, gift, $F(1, 104) = 23.98$, $p < .001$, and outline conditions, $F(1, 116) = 18.83$, $p < .001$, suggesting that increasing the material importance of a decision or outlining the coherent perspectives does not increase the strength of the coherence shift. The three-way interaction was significant in the endowment-engage, $F(1, 108) = 9.29$, $p > .01$, and endowment-breakup, $F(1, 95) = 14.45$, $p < .001$ conditions, suggesting that giving participants a prior preference within the pre-decision phase did not increase the strength of the coherence shift. The finding that introducing a prior preference within the pre-decision phase did not increase coherence shifts suggests that coherence seeking may generally operate within the pre-decision phase, even when there is no prior preference. Indeed, Simon and colleagues have found coherence shifts among participants who had not yet reached decisions (Holyoak & Simon, 1999; Simon et al., 2001; Simon, Snow, & Read, in press).

When we analyzed data from the two assigned decision conditions the three-way interaction was not significant, $p = .45$. The task in the assigned decision conditions was intended to minimize participants' involvement in the task, since they were simply waiting to be told what their character had decided. Thus, unlike Simon and colleagues' previous research on coherence shifts, in which participants expected to make a decision or had a memorization or communication goal (Simon et al., 2001), participants in the assigned decision conditions of the present research had no active processing goal. Our finding that participants in the assigned decision conditions were less likely to show coherence shifts, then, suggests that under minimal conditions, where participants think about a complex situation without any active processing goal, they may not achieve coherence.

Considering the results of the present research with those of previous research using the same type of paradigm in a legal context (Holyoak & Simon, 1999; Simon et al., 2001; Simon, Snow, & Read, in press) suggests that coherence shifts may occur in an all or nothing fashion. When participants think about a situation without an active processing goal (as in the forced decision conditions of the present research) they may not achieve coherence. However, if they do have an active processing goal (involving memorization, communication, or decision making) they are likely to report coherence shifts, and increasing the importance of the decision, outlining coherent perspectives, giving them a prior preference (as in the present research) or manipulating other aspects of the context (as in previous research) does not substantially affect the strength of the coherence shift. It appears, then, that whenever there is an active processing goal coherence shifts occur and that it may not be possible to further adjust their strength.

To test whether personality moderated the strength of the coherence shift, we added a variable representing personality measure. We found that PNS did not moderate the three-way interaction, $p < .5$, but NFC had a marginal effect overall, $F(1, 794) = 3.41$, $p < .07$, and a significant effect on fact items, $F(1, 794) = 4.39$, $p < .04$. Inspecting the means revealed that among participants who chose *engage*, those with higher NFC scores rated engagement facts higher at posttest than at pretest, $t(188) = 2.19$, $p < .04$, but those with lower NFC did not, $p < .3$. This finding suggests that people who have a greater "tendency to engage in and enjoy thinking" (Cacioppo & Petty, 1982, p. 116) may achieve a greater degree of coherence, though increased thought may only affect coherence among the facts of the issue currently being considered and may not affect more general beliefs.

Summary and Conclusions

We extended Simon et al.'s (2001; Simon, Snow, & Read, in press) paradigm to test for coherence shifts in perceptions of a romantic relationship. When participants thought about a couple's future and decided whether they would get engaged or break up, their attitudes about facts in the relationship and their general beliefs about relationships shifted to become coherent with their decision. When participants were not able to make their own decisions coherence shifts did not occur, suggesting that an active processing goal may be necessary to activate coherence mechanisms. Increasing the importance of the decision, increasing the ease of perceiving coherent perspectives, and introducing a prior preference within the pre-decision phase did not increase the strength of coherence shifts, suggesting that once there is an active processing goal and coherence mechanisms are activated it may not be possible to alter their intensity. Together with previous research in legal contexts (Holyoak & Simon, 1999; Simon et al.; Simon, Snow, & Read, in press), these findings suggest that coherence seeking mechanisms operate within the pre-decision phase in an all or nothing fashion, being activated any time there is an active processing goal. Finally, we found that individuals with higher need for cognition, who chronically engage in more cognitive processing, reported stronger coherence shifts in attitudes about facts of the current case, suggesting that individuals with higher NFC may achieve a greater degree of coherence across a broad range of situations in their daily lives.

Acknowledgments

This research was funded by NSF grant SES-0080424 to Dan Simon and Stephen J. Read.

References

Anderson, N. H. (1962). Application of an additive model to impression formation. *Science*, 138, 817-818.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology, 21*, 107-111.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*, 3–31.

Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruity in the prediction of attitude change. *Psychological Review, 62*, 42-55.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simple structure. *Journal of Personality and Social Psychology, 65*(1), 113-131.

Newcomb, T. M. (1953). An approach to the study of communicative acts. *Psychological Review, 60*, 393-404.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997). Connectionism, parallel constraint satisfaction processes, and gestalt principles: (Re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review*, *1*, 26-53.

Russo, J. E., Medvec, V. H., & Meloy, M. G. (1996). The distortion of information during decisions. *Organizational Behavior and Human Decision Processes, 66,* 102–110.

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The emergence of coherence over the course of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1250–1260.

Simon, D., Snow, C. J., & Read, S. J. (in press). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology*.

Spatial Language and Reference Frame Assignment; The Role of the Located Object

Michele Burigo (mburigo@plymouth.ac.uk)
School of Psychology, Faculty of Science, Drake Circus,
University of Plymouth, PL4 8AA, UK

Kenny Coventry (kcoventry@plymouth.ac.uk)
School of Psychology, Faculty of Science, Drake Circus,
University of Plymouth, PL4 8AA, UK

Abstract

Spatial prepositions work as pointers to localize objects in space. For instance "The book is over the table" indicates that the located object (LO) is somewhere "over" the reference object (RO). To understand where the LO is people need to assign direction to space (selecting a reference frame). Three experiments are reported which investigated the reference frame conflict between LO and RO. We found that when the LO was not vertically aligned, the appropriateness for a given spatial preposition (*above, below, over* and *under*) changes. In general scenes with the LO pointing at the RO were judged less acceptable than scenes with the LO vertically oriented. These results suggest that reference frames for both LO and RO are accessed before direction can be assigned for spatial prepositions. Modifications to Multiple Frame Activation theory (Carlson-Radvansky & Irwin, 1994) are discussed.

Introduction

Spatial language forms an essential part of the lexicon for a competent speaker of a language. In English, spatial prepositions work as pointers to localize objects in space. For instance "The book is over the table" indicates that the located object ("the book") is somewhere "over" the reference object ("the table"). Prepositions like "over" and "behind" (the so-called projective prepositions) are particularly interesting as they require the selection of a reference frame before the assignment of a direction to space specified by the preposition can be established. Levinson (1996) distinguishes between the intrinsic (object-centred), relative (or viewer-centred/deictic), or absolute (environment-centred/extrinsic) reference frames. For example, "the car is behind the house" used intrinsically locates the car in relation to the opposite wall from where the salient front of the house is (which is where the back door is). The relative use of the same expression locates the car directly behind the opposite wall to where the speaker and hearer are standing. The absolute frame locates an object with respect to a salient feature of the environment, such as the gravitational plane or cardinal directions (e.g., North, South, etc.).

Carlson-Radvansky and Logan (1997) have argued that spatial apprehension occurs in a series of stages as follows; (1) identify the reference object (e.g., the house), (2) superimpose multiple reference frames (relative and intrinsic), (3) construct spatial templates and align them to the relevant reference frames, (4) select a reference frame, (5) combine templates into a composite template, (6) search the composite template that fits best with the located object for each position within the template, (7) calculate whether the goodness of fit measure for the located object is high (good or acceptable region) or low (bad region). In this paper we focus on the process underlying the orientation of space and the consequent reference frames selection.

Experimental evidence has demonstrated that by rotating the reference object by 90º (noncanonical orientation), acceptability ratings for *above* mirror the new spatial template that is the sum of all the reference frames active in that moment (Carlson-Radvansky & Irwin, 1994). The acceptability for the given spatial preposition varies as a function of the reference frame activated. Consider the scenes in Figure 1. In the canonical orientation the absolute, relative and intrinsic reference frames overlap. In the noncanonical orientation the absolute reference frame is dissociated from the intrinsic. This produces a lower acceptability for the given spatial preposition because a conflict emerges between all the reference frames activated in that moment (Carlson-Radvansky & Irwin, 1994; Carlson, 1999).

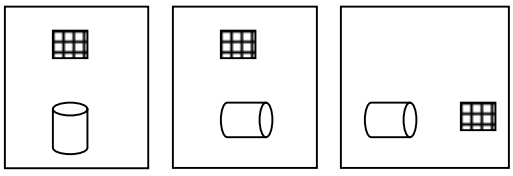


Figure 1: Canonical absolute/intrinsic "above" (left picture), noncanonical absolute "above" (middle picture) and noncanonical intrinsic "above" (right picture).

Although there is evidence that reference frame activation is important, to date studies have only focused on the reference frame generated from the reference object (Carlson & Logan, 2001; Carlson, 1999; Carlson-Radvansky & Logan, 1997). Furthermore, theories of spatial

language largely assume that the assignment of direction is generated from the reference object to the located object. For example, in the Attentional Vector-Sum model (Regier & Carlson, 2001) the direction indicated by a spatial relation is defined as a sum over a population of vectors that are weighted by attention. An attentional beam is focused on the reference object (at the point that is vertically aligned with the closest part of the located object) and separated in a population of vectors pointing toward the located object. But other experiments suggest that both objects (even distractors or those not relevant for the task) require allocation of attention to be processed (Lavie, 1995; Lavie & Cox, 1997). This suggests that both objects could play a role in the spatial apprehension process.

There is much evidence indicating that the LO is important in establishing the acceptability of a range of spatial prepositions (see Coventry & Garrod, 2004 for a review). For example, Coventry, Prat-Sala and Richards (2001) found that the appropriateness of a spatial preposition is correlated with the functional relation between located and reference object. For example, an umbrella is regarded as being more *over* a person if it is shown to protect that person from rain than when the rain is shown to hit the person. Furthermore, Coventry et al. found that the acceptability ratings for *over* and *under* were more influenced by the function of the object than by the relative positions of LO and RO, while conversely *above* and *below* were more influenced by geometry than function. Additionally, in a study which manipulated reference frame conflicts with function present (e.g., the man holding the umbrella in the gravitational plane was either upright, lying down, or upside down), Coventry et al. found that reference frame conflicts influenced the acceptability of *above* and *below* more than *over* and *under*.

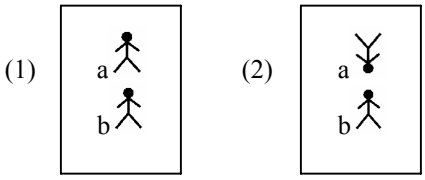


Figure 2: Reference frame conflicts between LO and RO.

However, although there is much evidence that the located object does influence the acceptability of a range of prepositions, no studies to date have examined whether the located object contributes to reference frame assignment, and hence the assignment of direction to space. This paper reports three experiments employing an acceptability rating task where possible reference frame conflicts for both the located object and reference object are manipulated. For example, consider the scene in Figure 2, and the acceptability of man [a] is above man [b]. In (1), the reference frame of man in location [a] is aligned with the reference frame of man [b] (the reference object), but not in (2) where their intrinsic reference frames are in conflict. We predicted that rotating the LO in this way would influence the appropriateness of *over, under, above* and *below*

Experiment 1

In this experiment we tested the hypothesis that the reference frame(s) associated with the located object would affect acceptability of *over, under, above* and *below* to describe the position of the LO in relation to a RO.

Method

Participants & Procedure

Twenty-three undergraduate students from the University of Plymouth participated in this investigation for course credit. All the participants were English native speakers. Participants had to judge the appropriateness of a spatial preposition (*above, below, over* or *under*) to describe pictures using a scale from 1 to 9 (where 1 = not at all acceptable and 9 = perfectly acceptable). All trials showed the located object in a "good" or "acceptable" location, never in a "bad" location (following Carlson-Radvansky and Logan's definitions, 1997).

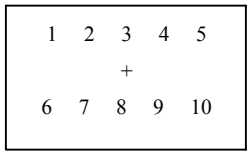


Figure 3: Location for the located object with respect to the reference object (indicated here with a "plus").

The located object could appear in 10 different locations around the reference object (see Figure 3). The sentences were shown before the scene and in this form; <The "located object" is PREPOSITION the "reference object">. The prepositions tested were *above, below, under* or *over*. Two orientations for the located object were used: "vertical" and "pointing at". In the "pointing at" condition the axis of the located object was pointing exactly towards the center-of-mass of the reference object.

Materials

The materials consisted of three stimuli; a circle, an hourglass and a stickman. These objects were selected as the circle does not have an oriented axis, while the hourglass has a salient axis but not an intrinsic top and bottom, and the stickman has a salient axis and an intrinsic top and bottom. We will use the following labels to classify the objects; "no axis" (circle), "ambiguous axis" (hourglass) and "intrinsic axis" (stickman). All the objects employed were presented at the same size and distance from the reference object regardless of the orientation. This is because it has been found that proximity, center-of-mass orientation and distance affect the appropriateness of spatial preposition (Regier & Carlson, 2001).The objects could appear as reference objects or as located objects, but the same object was never shown as LO and RO at the same time.

Design
The experiment consisted of 480 trials constructed from the following variables: 4 spatial prepositions X 10 locations X 3 objects X 2 orientations ("vertical" and "pointing at"). The locations were collapsed in two factors; high vs. low location (2 levels) and proximity (3 levels) as follows; far misaligned (locations 1, 6 and 5, 10) versus near misaligned (locations 2 and 4) versus aligned (central location). All the trials were presented in a randomized order.

Results
A 4-way within subjects ANOVA was performed on the rating data. The variables included in the analysis were; 2 located objects (hourglass versus stickman) x 2 preposition set (above-below vs. over-under) x 2 superior versus inferior prepositions (above-over vs. below- under) x 2 orientations of LO (vertical and pointing at). The division between spatial prepositions has been employed following the Coventry et al. findings summarized above (Coventry, Prat-Sala and Richards, 2001).

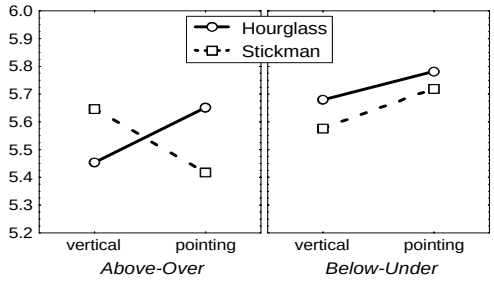


Figure 4: 3-way interaction between superior versus inferior prepositions (above/below vs. over/under), located object and orientation of LO (collapsed over locations).

Trials with the circle as the located object were excluded from the analysis since this kind of object does not have an axis. Furthermore we analyzed only the trials with a circle as the reference object because it has no axis. A main effect of preposition set (above-below vs. over-under) was found, [$F_{(1, 22)} = 7.21$, $p < .05$]. Higher ratings were given for Above-Below ($M = 6.526$) than for Over-Under ($M = 5.192$). This is unsurprising as it known that these spatial prepositions have larger areas of acceptability. No other significant main effects were found. There was a significant 3-way interaction between superior versus inferior spatial prepositions, located object and orientation of LO [$F_{(1, 22)} = 6.694$, $p < .05$], displayed in Figure 4. It is interesting to note that objects with a top/bottom orientation such as a stickman are rated less acceptable when pointing ($M = 5.42$) than when vertical ($M = 5.65$) for trials with above-over, although this was not the case for below-under ($M_{vertical} = 5.58$; $M_{pointing} = 5.72$). None of the other interactions were significant.

Discussion

An interesting difference was found between trials with the stickman and trials with the hourglass as LOs. The stickman trials generate a reference frame conflict in the pointing condition but the hourglass did not. This could be explained by a preferential assignation of a top/bottom orientation based on the vertical plane. In other words an hourglass could not be seen as upside down but always as pointing away from the reference object.

Acceptability rating showed that for inferior spatial prepositions (below-under) the pointing condition was more acceptable than the vertical one. All these results can be explained by the activation of an intrinsic reference frame on the located object that in the case of under-below produces facilitation and with above-below produces conflict. Therefore the results seem to suggest that the orientation of the located object is important in establishing the appropriateness of projective prepositions. However, this experiment only used two located objects (an hourglass and a stickman), so there is an issue regarding the extent to which the results can be generalized. For this reason the aim of the next experiment is to try to replicate the effect of the orientation of the LO using a wider range of LOs and orientations of LO.

Experiment 2

The second experiment utilized the same design and procedure as the first experiment, except that more materials and orientations of LO were included.

Method

Participants & Procedure
Twenty-nine undergraduate students from the University of Plymouth participated in this investigation for course credit. All the participants were English native speakers and none of them took part in the previous experiment. The procedure was the same procedure used for the previous experiment based on the acceptability rating task of the given spatial prepositions; *above, below, over* and *under*.

Materials
This experiment involved a wider number of located objects and two more orientations; "pointing away" from the reference object and "upside down". The reference object in this experiment was always a picture of a football. The located objects were picked from two sets; the first consisted of objects with a distinctive top-bottom (8 new objects "with an intrinsic axis") and the second one of objects with "an ambiguous axis" (7 new objects plus the hourglass). All the stimuli were hand-drawn and transformed to electronic format by a computer scanner.

Design
There were 384 trials constructed from the following variables: 8 located objects X 3 locations (collapsed over side) X 4 spatial preposition X 4 orientations ("vertical", "upside down", "pointing at" and "pointing away"). All the trials were presented in a randomized order. We added 192 distractors where the LOs were objects without salient axes, meaning that a total of 576 trials were presented.

Results

A full factorial ANOVA was chosen to analyze the data. In this analysis we focus on trials where the LO had an intrinsic axis (following the results of Experiment 1). A significant main effect was found for preposition type (above-below vs. over-under), [$F_{(1,28)} = 15.44$, $p < .001$], for superior versus inferior prepositions, [$F_{(1,28)} = 10.72$, $p < .005$], for location [$F_{(1,28)} = 80.17$ $p < .0001$] and for direction [$F_{(1,84)} = 3.35$, $p < .05$]. Objects vertically oriented ($M = 5.75$) were judged more acceptable than the other levels of orientation. In particular the "upside down" ($M = 5.6$) and "pointing at" ($M = 5.55$) orientations produced the lowest ratings (and indeed generated the highest reference frame conflict). The analysis also revealed significant 2-way interactions between preposition set and location [$F_{(1,28)} = 10.96$, $p < .005$], between preposition set and direction [$F_{(1,84)} = 3.23$. $p < .05$] and between superior versus inferior prepositions and direction [$F_{(1,84)} = 2.82$, $p < .05$].

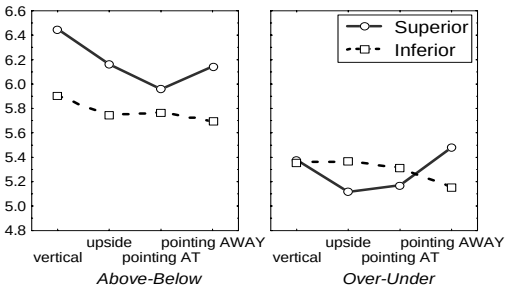


Figure 5. 3-way interaction between orientation of LO, superior-inferior prepositions and preposition set.

Finally, there was also a significant 3-way interaction between superior-inferior prepositions, preposition set and location, [$F_{(1,28)} = 5.45$, $p < .05$], and between preposition set, superior-inferior preposition and direction [$F_{(1,84)} = 3.99$, $p < .01$]. This interaction is displayed in Figure 5. As can be seen in Figure 5, the results of the orientation of LO are clearest for *above*, which exhibited a reliable difference between the vertical orientation of LO and all the other levels of LO. For *over*, pointing away from the RO is also associated with higher acceptability ratings. The results are less clear for inferior prepositions.

Discussion

The pattern of results in this second experiment confirms the hypothesis that the orientation of the located object influences acceptability ratings, although there are clear differences between prepositions. However, in the first two experiments the reference objects were objects without a salient axis. It is therefore possible that the activation of the located object reference frame could depend on the features of the reference object. The next experiment tested whether the effects of the orientation of LO were present across a wider range of ROs.

Experiment 3

This experiment used the same basic methodology as before, but this time with a range of reference objects including ROs without a salient axis, with an ambiguous axis, and with an intrinsic axis.

Method

Participants& procedure

Twenty-three undergraduate students from the University of Plymouth participated in this investigation for course credit. All the participants were English native speakers and they did not take part in any of the previous experiments. The procedure was the same as that used in Experiments 1 and 2.

Materials

For this experiment we used a set of 24 objects (8 "without a salient axis", 8 "with an ambiguous axis" and 8 "with an intrinsic axis"). The objects "with an ambiguous axis" and "with an intrinsic axis" were the same as those used in Experiment 2. We drew 8 new objects "without a salient axis". Thus we were able to study the effect of the reference frame activation on the located object in scenes with different kinds of reference object.

Design

The experiment was composed of 576 trials with the following factors: 8 located objects with an intrinsic axis (treated as random factor), X 3 reference objects (picked up from a set of 24 objects, 8 with no axis, 8 with an ambiguous axis, and 8 with an intrinsic axis; within subjects factor), X 2 prepositions set (between subjects factor), X 2 superior-inferior preposition (within subjects), X 3 locations for the probe (within subjects) and 4 directions for the located object (within subjects). This time preposition set was between subjects; half the participants received *above* and *below* and the other half received *over* and *under*.

Results

We performed two analyses; one by subjects (F^1) and one by materials (F^2). The results were similar for both analyses, so here we report the F^1 analyses alone. The means for all the conditions can be found in Table 1. Significant main effects were found for superior-inferior prepositions [$F_{(1,22)} = 18.74$, $p < .001$], for location [$F_{(1,22)} = 69.14$, $p < .0001$] and for orientation of LO [$F_{(1,44)} = 5.25$, $p < .005$]. Furthermore we found several significant 2-way interactions; between preposition set and RO [$F_{(1,44)} = 3.61$, $p < .05$], between location and RO [$F_{(1,44)} = 4.45$, $p < .05$], between superior-inferior prepositions and orientation of LO [$F_{(1,66)} = 4.93$, $p < .005$] and between location and orientation of LO [$F_{(1,66)} = 3.12$, $p < .05$].

Located Object (intrinsic)		*Reference Object*		
		No axis	Ambig.	Intrin.
Above	vertical	6.281	6.307	6.375
	inverted	5.560	5.542	5.490
	point at	5.524	5.670	5.644
	point away	6.047	6.026	5.974
Below	vertical	5.797	6.036	6.167
	inverted	5.411	5.604	5.453
	point at	5.786	5.823	5.754
	point away	5.387	5.536	5.578
Over	vertical	5.419	5.084	5.479
	inverted	5.047	4.823	5.220
	point at	5.182	5.115	5.188
	point away	5.785	5.366	5.335
Under	vertical	5.131	5.058	5.162
	inverted	4.691	4.901	4.889
	point at	5.335	5.073	5.156
	point away	4.698	4.693	4.901

Table 1. Means for conditions across the four spatial prepositions (*above, below, over and under*).

A 3-way interaction was also significant between superior-inferior prepositions, location and orientation of LO [$F_{(1,66)} = 3.93$, $p < .05$] and a 4-way interaction between superior-inferior prepositions, location, RO and orientation of LO [$F_{(1,132)} = 2.74$, $p < .05$]. Follow-up analysis revealed significant differences in orientation between prepositions and locations, but the effects of orientation were present at all levels of RO.

Discussion

The outcome from this experiment supports the idea that the orientation of the located object affects acceptability ratings even when the reference object has an intrinsic orientation. The results for this experiment mirror the results of the previous experiment, but extend the results to show that the activation of reference frame for the LO is not restricted to cases where the RO does not provide sufficient information to cue a reference frame.

General Discussion

The series of experiments explored the hypothesis that the spatial apprehension process computes a composite template for a given spatial preposition making use of the located object reference frame as well as the reference object reference frame. The results of the experiments confirmed this hypothesis showing that the orientation of the located object affects acceptability ratings for projective prepositions.

The results suggest necessary extensions to the idea of Multiple Frame Activation (Carlson-Radvansky & Irwin, 1994) where it is suggested that in comprehending a scene multiple frames are available. However, we found that an additional reference frame is generated from the located object as well as from the reference object and the final template generated is influenced by its orientation.

In addition to the reliable effects of the orientation of LO for intrinsic objects, the results of Experiment 1 showed some interesting differences between intrinsic objects and objects such as an hourglass with a salient axis, but without an intrinsic axis. For the objects like an hourglass, the "pointing at" condition was considered more acceptable than the vertical condition. A possible explanation is that people assign a subjective top/bottom orientation to "ambiguous" objects. Thus the hourglass in trials with above-over should be seen as pointing away from the reference object instead of pointing at the RO.

The last experiment provided evidence that the conflict among reference frames emerges across a range of reference objects, including those that are more "real" with a top/bottom orientation. So the effect of the located object is not exclusive for circle-like reference objects but it is part of a more general process.

But why should we activate the located object reference frame when the reference frame of the RO should be sufficient to localize the objects in the scene? An explanation is that objects not vertically oriented suggest that there is something implausible in the scene. A cat upside down is not a "plausible" stereotypical mental representation. Thus the knowledge revision function (Holland, Holyoak, Nisbett & Thagard, 1986, Wason, 1960) should look for an explanation; this activates the located object reference frame to process every possible orientation that fits with the whole scene. Another possible explanation is based on the concept of direction of potential motion (Regier, 1996). People perceive objects rotated away from the gravitational plane as falling. So a located object oriented at 90º may be perceived as moving downwards on a path to the left of/right of and away from the reference object.

Implication for existing models

The results found suggest a review of the key characteristics of the spatial apprehension process Carlson-Radvansky & Irwin, 1994; Carlson & Logan, 2001; Hayward & Tarr, 1995; Logan & Sadler, 1996). We found evidence of an involvement of the located object reference frame in the process of assigning direction to space. Therefore evaluating the process of goodness of fit of the spatial preposition involves the located object as well and future studies should take this into account. The finding that the located object interacts with the spatial apprehension process has some repercussions for models of spatial language as well. Models such as the Attentional-Vector-Sum model (Regier & Carlson, 2001) simulate attentional processes, but thus far does not deal with attentional processing of the LO (but see Regier, Carlson & Corrigan, 2004, for a modification of AVS to deal with processing of function). It may be possible to develop the AVS model to deal with the projection of vectors from the LO to the RO as well as the

other way round (see Coventry and Garrod, 2004, for a discussion).

Limitations and future developments

This investigation brings experimental evidence in support of the hypothesis that the located object, in a scene with two objects, takes part in the spatial apprehension process. Future investigations should attempt to ascertain the degree to which features of the LO influence the spatial apprehension process further. For example, in some contexts the LO may be more important than the RO, and vice versa for other contexts. In addition, the present experiments do not tell us anything about the time course of processing of LO reference frames. Studies underway are testing the conflict among reference frames using a reaction time paradigm. Finally, we should consider how these findings can be implemented within frameworks such as the AVS model.

Acknowledgments

We would like to thank Lau Thiam Kok for his drawing skills.

References

Carlson-Radvansky, L. A. & Irwin, D. E. (1994). Reference frame activation during spatial term assignment. *Journal of Memory and Language*, 33, 646 – 671.

Carlson-Radvansky, L. A. & Logan, G. D. (1997). The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37, 411 – 437.

Carlson, L. A. (1999). Selecting a reference frame. *Spatial Cognition and Computation*, 1, 365 – 379.

Carlson, L. A. & Logan, G. D. (2001). Using spatial terms to select an object. *Memory and Cognition*, 29 (6), 883 – 892.

Coventry, K. R., Prat-Sala, M. and Richards, L. (2001). The interplay between geometry and function in the comprehension of Over, Under, Above and Below. *Journal of Memory and Language. 44(3),* 376 – 398.

Coventry, K. R., & Garrod, S. C. (2004). *Saying, seeing and acting. The psychological semantics of spatial prepositions.* Psychology Press; Hove and New York.

Hayward, W. G. & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55, 39 – 84.

Holland, J. H. , Holyoak, K. J., Nisbett, R. E. & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT press.

Lavie, N. (1995). Perceptual load as necessary condition for selective attention. *Journal of Epxerimental Psychology: Human, Perception and Performance*. 21(3), 451 – 468.

Lavie, N. & Cox, S. (1997). On the efficiency of visual attention: efficient visual search leads to inefficient distractor rejection. *Psychological Science*. 8(5), 395 - 398

Levinson, S. C. (1996). Frames of reference and Molyneux's question. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (pp. 109-169). Cambridge, MA: MIT Press.

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (pp. 493-529). Cambridge, MA: MIT Press.

Regier, T (1996). *The Human Semantic Potential. Spatial language and constrained connectionism*. Cambridge, The MIT Press.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130 (2), 273 – 298.

Regier, T., Carlson, L. A., & Corrigan, B. (2004). *Attention in spatial language: Bridging geometry and function*. In L. Carlson & E. van der Zee (Eds.). Functional features in language and space: Insights from perception, categorization and development. Oxford University Press.

Wason, P. C. (1960). On the failure to eliminate hypothesis in conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129 – 140.

Multiple Session Masked Priming: Individual differences in orthographic neighbourhood effects

Claire J. Byrne (Claire.Byrne@med.monash.edu.au)
Department of Psychology, School of Psychology, Psychiatry and Psychological Medicine
Building 17, Monash University, VIC 3800, Australia

Gregory W. Yelland (Greg.Yelland@med.monash.edu.au)
Department of Psychology, School of Psychology, Psychiatry and Psychological Medicine
Building 17, Monash University, VIC 3800, Australia

Abstract

Multiple Session Masked Priming was used to investigate differences between individuals in the fine-tuning of their lexical representations. The form-priming effects were determined separately for each of the 50 participants, from the patterning of masked priming effects over three neighbourhood (N) levels. At each N level, primes varied in their orthographic similarity to the target - Identity, One-Letter-Different and All-Letters-Different. An analysis of the pooled data showed patterns of masked priming consistent with other group studies, but considered separately, an array of individual differences in the tuning of lexical representations was observed.

Inconsistent findings and conflicting evidence characterise much of the word recognition literature, and have lead to a plethora of theories on many aspects of word recognition. One explanation for the conflicting evidence regarding theories of word recognition, is the possibility that it reflects individual differences. The problem here lies in the fact that support for one model or another has generally been drawn from standard group studies, where outcomes are averaged over the sample of participants. Averaged outcomes may show what is generally true for that sample, but fail to capture the critical variations between individuals.

In order to study individual differences a task must be found that will provide reliable data concerning the processing capacities of each participant. There are two issues here. First, the task used must reflect the automatic processes of lexical access, and second, the task must generate sufficient data on each participant to enable stable conclusions to be drawn about their individual performance; that is, we need to be able to generate reliable individual profiles of lexical processing.

Although there is some debate (e.g. Bodner & Masson, 1997; Bodner & Masson, 2001), Forster and Davis' (1984) Masked Priming paradigm is widely considered to be an ideal tool for examining automatic lexical processing. In masked priming, a prime word/nonword is presented very briefly (50-60ms), forward masked (for 500ms) by a row of hash marks (######) and backward masked by an upper case target (500ms), to which a timed lexical decision response is made. The relationship between prime and target is manipulated in some way and if performance on the target is enhanced by a particular prime-target relationship, relative to a control condition, this relationship is taken to reflect the properties of written words important for lexical access. Masked priming has been argued to enable the investigation of lexical access processes free from any influence of more central and strategic processes (e.g. Forster, 1998). This is achieved in the way the prime is presented, making it typically unavailable for conscious report by participants. Critically, while the masked prime is not available for conscious report, a variety of priming effects are consistently observed. Such facilitatory priming effects include repetition-priming (e.g. Forster & Davis, 1984) where prime and target are the same word (e.g. farm – FARM) and form-priming (e.g. Forster, Davis, Schocknect & Carter, 1987; Forster & Veres, 1998) where the prime is of similar orthographic form (e.g. firm – FARM). The absence of both expectancy effects (e.g. Forster, 1998) and priming for nonwords (e.g. Forster & Davis, 1984; Forster et al., 1987; cf. Bodner & Masson, 1997), in addition to the existence of semantic- (e.g. Perea & Gotor, 1997) and cross-language translation-priming (e.g. Kim & Davis, 2003) indicate that masked priming effects reflect lexical level processing.

While the masked priming task provides a means of examining the automatic processes of lexical access, a new variant of this task was required to enable the collection of sufficient data from each participant that they could stand as an experiment in their own right. The Multiple Session Masked Priming paradigm (e.g. Byrne, Yelland, Johnston & Pratt, 2000; Yelland & Byrne, 2001) achieves this by repeatedly testing each participant on the same experiment. Since participants are unaware of the primes, the prime-target relationship cannot be realised, even over repeated test sessions. Thus, in the Multiple Session Masked Priming paradigm, the participant's experience is simply one of repeated exposure to the target items.

Used previously to reveal marked individual variation in the use of orthographic and phonological input codes for lexical access (e.g. Byrne, Yelland, Johnston & Pratt, 2000; Yelland & Byrne, 2001), the Multiple Session Masked Priming paradigm has proved useful for investigating individual differences in lexical processing. This study aims to look at the use of this new technique in other areas of the word recognition literature, namely orthographic neighbourhood effects.

The neighbourhood effect was first investigated by Coltheart, Davelaar, Jonasson and Besner (1977), who coined the term neighbourhood (N) density. A word's N density is simply the number of words that can be made from it by changing one letter. For instance, a high N word like *torn* has many neighbours – *corn, born, turn, town, tore* etc. In this way, N provides a rough index of a word's similarity to other words. The effects of such orthographic similarity have been used by researchers to model constraints on, and organisation of, the internal word-recognition system. Previous experiments looking at the effects of N density have shown form-priming effects when using word or nonword primes that are one-letter-different from their word targets, however, such effects are present for low N density words only, not high N (e.g. Forster et al., 1987). The form-priming effects seen at the low N density are reported consistently, albeit are smaller in magnitude than those observed in repetition-priming (e.g. Andrews, 1997). The lack of form-priming effects in high N words is usually described as an increased tuning of these lexical representations. Forster et al. (1987) first proposed a now commonly used explanation for this change in tuning, whereby a word's lexical representation becomes more finely tuned in response to increasing N density. A word with many neighbours needs to ensure only exact matches can activate its representation, otherwise a large number of these neighbours will need to be considered and errors in activation are more likely. In contrast, a word with only few neighbours can afford to be less stringent (i.e. it can be activated by a one-letter-different neighbour) as there are only a small number of possible lexical candidates. A range of word recognition theories and models have been provided for these results. Such accounts, however, will only prove useful if either, all readers show the same pattern of priming effects as a function of N density, or they are able to account for individual variation in sensitivity to N levels . The question addressed in this study is whether there are individual differences in the priming effects for similar form, indicating differences in the fine-tuning of lexical representations.

Method

Participants

Seventy-one undergraduates, graduates and staff from Monash University, Australia, volunteered to complete the experiment. All were aged between 18 and 33 years with normal or corrected-to-normal vision and were paid a token amount for their participation.

Seventeen participants chose not to complete all 12 sessions, and were therefore excluded from the analyses. In addition, a further four participants were excluded (three on the basis of overall error rates exceeding 20%, and one who was found to be a non-native speaker of English). In all, 50 participants contribute to the final analyses.

Materials and Design

One hundred and eight 4-letter words were used as targets. All were monosyllabic content words with a mean frequency of occurrence of 85.2 (SD=162.9) tokens per million in text (Kucera & Francis, 1967). The target words were selected from three (n=36) N levels (i.e. Low, Medium and High), with written frequency matched across these conditions. In addition, targets were chosen to ensure they did not belong in another target's neighbourhood. For each target word, primes of three types were constructed: (a) an All-Letters-Different (ALD) prime, a nonword which differed at each letter position from the target, to be used as a baseline control; (b) an Identity (ID) prime, which was the exact same word as the target, to measure the maximum priming effect; and (c) a One-Letter-Different (1-LD) prime, a nonword which differed from the target in one letter position, to examine any form-priming effects. The letter position at which the 1-LD prime differed from the target was varied evenly within the three N levels. Primes were matched to targets on N and consonant-vowel patterns within neighbourhood conditions.

An equivalent set of nonword items was constructed to act as foils in the lexical decision task. All matching procedures undertaken for the selection of word targets and primes was repeated for the nonword items, with the exception of written frequency matching. Some characteristics of the target items are shown in Table 1, while example items are shown in Table 2.

Table 1: Characteristics of the Masked Priming Target Items

	N		Frequency
	Range	Mean (SD)	Mean (SD)
Words			
Low N	1 - 3	2.1 (0.8)	84.3 (153.7)
Med N	6 - 7	6.5 (0.5)	85.3 (173.0)
High N	12 -15	13.9 (1.1)	85.9 (165.9)
Nonwords			
Low N	1 - 3	2.3 (0.8)	N/A
Med N	6 - 7	6.5 (0.5)	N/A
High N	10 -15	11.9 (1.5)	N/A

Note. N = Neighbourhood size

Table 2: Example Targets and Primes

	Target	ID Prime	1-LD Prime	ALD Prime
Words				
Low N	FREE	Free	frue	merp
Med N	HELP	Help	hulp	vand
High N	WIDE	Wide	kide	barp
Nonwords				
Low N	NORL	Norl	norf	vube
Med N	PALD	Pald	peld	bist
High N	BAFE	Bafe	bape	fook

Note. N = Neighbourhood size; ID = Identity; 1-LD = One-Letter-Different; ALD = All-Letters-Different

All targets were presented as UPPER CASE letter strings, while primes were presented in lower case. This reduced the possibility that any observed priming effects could be the result of priming by virtue of visual features rather than orthographic form. Three different, counterbalanced versions of the experiment were constructed so that all targets appeared with each of their primes, but only once in each version. Practice items (one exemplar of each word and nonword condition) were constructed and placed at the beginning of each version.

Procedure

For each of the 12 masked priming sessions, participants were seated in a sound attenuated booth where items were displayed on a 15" flat screen, fast decay monitor, controlled by an IBM compatible computer operating DMDX presentation software (developed at Monash University and at the University of Arizona by K.I. Forster and J.C. Forster). Participants began the experiment by pressing a foot pedal, which initiated the presentation of the instructions in the centre of the screen. The practice items were then shown, followed by the experimental items. Progress through the experiment was self-paced with a press of the foot pedal initiating each item. Participants took approximately 10 minutes to complete each session.

Each item was presented in the centre of the screen in the following sequence. First, the forward mask, a row of six hash marks (e.g. ######) was presented for 500ms. This was replaced by the 54ms presentation of the lower case prime, which was in turn replaced by the upper case target for 500ms. Participants made a timed lexical decision response to each target, pressing a blue button with their preferred hand if the target was a word of English and a red button with their other hand if it was not. Participants received response time (RT) and accuracy feedback after each response. RT was measured from the onset of the target display to the button-press response. Both response latencies and errors were recorded for each item by the computer.

Participants were asked to complete 12 masked priming sessions to ensure that each prime-target pair were tested enough to gain sufficient data for every item. Therefore, the three counterbalanced versions were completed four times each. The participants completed the sessions on a roughly daily basis, according to their availability.

Results and Discussion

In order to determine whether the group study approach was masking quite disparate patterns of individual difference amongst participants, analysis of the patterning of the group data was needed to provide a baseline against which the individual differences could be examined.

Group Analyses

The data entered into the group analyses were the condition means for both session and item based data for each of the 50 participants. Analyses of the data from word targets comprised three planned comparison ANOVAs at each of the three N levels, conducted over both participant (F_1) and item (F_2) based data. The first compared the ID and ALD prime conditions, to confirm the existence of a repetition-priming effect. The subsequent analyses looked for orthographic form-priming (1-LD vs ALD) and whether form-priming was equal in magnitude to the repetition-priming effect (ID vs 1-LD)[1]. In the interests of brevity, only the significance level of F_1 and F_2 will be reported. The group mean RTs and error data obtained from participant-based data for word targets are shown in Table 3.

Table 3: Group mean response time (RT), percentage error rate (+/- SE) and priming effects (Δ), as a function of prime type and neighbourhood (N) level for word targets.

N level	Prime	RT (ms)	Δ	% Error	Δ
Low	ID	418 (7.1)	56	3.0 (0.02)	6.8
	1-LD	448 (8.1)	26	7.4 (0.52)	2.4
	ALD	474 (7.3)	-	9.8 (0.40)	-
Medium	ID	419 (7.5)	52	3.9 (0.35)	5.6
	1-LD	449 (8.3)	22	7.7 (0.46)	1.8
	ALD	471 (7.8)	-	9.5 (0.55)	-
High	ID	432 (7.4)	41	6.2 (0.66)	5.7
	1-LD	469 (8.9)	4	12.6 (1.02)	-0.7
	ALD	473 (7.6)	-	11.9 (1.00)	-

Note. ID = Identity; 1-LD = One-Letter-Different; ALD = All-Letters-Different

As can be seen in the summary of RT data (Table 3), participants responded more rapidly to word targets preceded by their ID prime than their ALD prime, yielding significant repetition-priming effects at each of the N levels [F_1 & F_2, $p < 0.01$]. Repetition effects were found in the error data also [F_1 & F_2, $p < 0.01$]. These results indicate that the ID primes facilitate faster and more accurate lexical access for their targets at each N level.

A comparison between the mean RT for the 1-LD and ALD prime conditions showed significant orthographic form-priming effects at the Low and Medium N levels [F_1 & F_2, $p < 0.01$], while at the High N level this comparison was not significant [F_1 & F_2, $p > 0.05$]. This pattern was mirrored in the error data, with a significant decrease in errors for 1-LD primes over ALD primes for Low and Medium N targets [F_1 & F_2, $p < 0.01$], but not for High N targets [F_1 & F_2, $p > 0.05$].

As expected for masked priming studies with group data, the orthographic form-priming effects, seen in the Medium and Low N conditions, were significantly reduced in magnitude relative to those seen with repetition-priming, for both RT [F_1 & F_2, $p < 0.01$] and accuracy [F_1 & F_2, $p < 0.01$] (i.e. ID vs 1-LD). For the High N targets, this

[1] While not orthogonal, these comparisons are meaningful, an attribute Keppel (1991) believes is more important than a set of purely orthogonal comparisons without such meaning.

comparison was also significant [F_1 & F_2, $p < 0.01$], however, this was in the absence of a 1-LD prime advantage.

Overall, the group analyses for the word targets provide results which are comparable with other masked priming studies in this area. It would seem from these results that, on average, only words with High N density become finely tuned, while words without such neighbourhoods (i.e. Low and Medium) are open to activation by close relatives.

In order to complete the traditional group analyses, an investigation of nonword effects was undertaken. An omnibus analysis of the group nonword data revealed significant main effects of N Level [F_1 & F_2, $p < 0.01$] and Prime Type [F_1 & F_2, $p < 0.01$] but no interaction between them [F_1 & F_2, $p > 0.05$]. Subanalyses were then conducted to examine the within factor effects, separately for N level and Prime Type. The group mean RTs and error data, obtained from participant-based analyses of the nonwords, are shown in Table 4 for the three N levels (collapsed over prime type) and the three prime types (collapsed over N level).

Table 4: Group mean response time (RT) and percentage error rate (+/- SE) by prime type and neighbourhood (N) level for nonword targets.

	RT (ms)	% Error
N Level		
Low	471 (8.4)	4.2 (0.42)
Medium	493 (10.1)	9.5 (0.52)
High	500 (11.7)	9.7 (0.78)
Prime Type		
ID	483 (10.9)	7.9 (0.95)
1-LD	486 (10.4)	7.5 (0.93)
ALD	495 (10.7)	8.0 (1.00)

Note. ID = Identity; 1-LD = One-Letter-Different; ALD = All-Letters-Different

As seen in Table 4, the Low N nonword targets were responded to significantly faster and more accurately than both the High N [RT: F_1 & F_2, $p < 0.01$; % Error: F_1 & F_2, $p < 0.01$] and Medium N targets [RT: F_1 & F_2, $p < 0.01$; % Error: F_1 & F_2, $p < 0.01$]. This difference was also evident between the Low and Medium N targets [RT F_1 & F_2, $p < 0.01$; % Error: F_1 & F_2, $p < 0.01$]. However, no such differences were found between the Medium and High N targets [RT: F_1, $p < 0.05$, F_2, $p > 0.05$; % Error: F_1 & F_2, $p > 0.05$]. These results show that only the Low N words had an RT and accuracy advantage, an expected effect in nonwords, as a nonword with more neighbours has a larger number of words to eliminate before concluding that the target is in fact not a word.

Somewhat surprising were nonword priming effects in the group RT data. From Table 4 it can be seen that the ID primed target has a small, but secure, 12ms facilitation relative to the ALD baseline [F_1 & F_2, $p < 0.01$]. A similar sized (9ms) significant facilitation effect was afforded the 1-LD over ALD primes [F_1 & F_2, $p < 0.01$]. Indeed, the 3ms advantage ID primes have over 1-LD primes proved to be significant [F_1 & F_2, $p < 0.01$]. Nonword effects of these magnitudes are not uncommon in group studies; however, they rarely prove significant (Forster, 1998). They do so here due to the power afforded the group analyses with 7,200 data points per condition. Although the absence of nonword effects is usually taken to indicate that masked priming effects for words are showing lexical level processes, the nonword priming effects found here do not contradict this conclusion. As only a lexical representation can be tuned by its N density, the modulation of priming effects by N density shown for words, but not nonwords, would indicate that the word priming effects are lexical in nature. In additon, the large reduction in magnitude of the nonword priming, compared to the word priming, would also indicate different mechanisms were involved in priming for these targets. Taken together, these findings suggest that the priming effects seen for word targets can still be considered lexical in nature, but may contain a small pre-lexical, graphemic priming component.

The critical issue, however, is whether these patterns of outcomes for the group data are an accurate characterisation of the lexical processes of all skilled readers, or the average of individual differences in the nature of lexical tuning across skilled readers.

Individual Analyses

The Multiple Session Masked Priming procedure yields sufficient data for each participant to be considered an experiment in their own right. That is, each of the 50 participants were the equivalent of an experiment, where test sessions replace participants. This gives a fully repeated measures design that provides 2,592 points of data for every individual (12 sessions over 18 conditions of 12 items each). Critically, this provides sufficient data for any one person, that the analyses of the type performed on the group data can be carried out for each participant separately. From this, reliable statements can be made for *each* participant about changes in orthographic form-priming as a function of N level.

Words

Each participant's data were analysed in the same way as the group data. A true appreciation of the subtleties of individual differences found within the data would require consideration of the priming outcomes at each N level, separately for each of the 50 participants. Unfortunately, space does not permit this. Instead, participants were classified according to their pattern of significant priming effects, as a function of N level. Priming effects fell into four categories. Three represent different patterns of priming effects between the ID, 1-LD and ALD conditions that are statistically distinguishable, (a) ALD>1-LD=ID, significant orthographic form-priming of the same magnitude as the repetition-priming effect; (b) ALD>1-LD>ID, orthographic form-priming which was significantly

greater than the baseline, but significantly less than the repetition-priming effect; and (c) ALD=1-LD>ID, no orthographic form-priming, in the presence of secure repetition-priming. These patterns of secure priming effects are illustrated in Figure 1. The fourth category contained cases where the pattern of priming effects was not secure. Each was of the form of (a), (b) or (c) but variability in the participants was such that the pattern of effects could not be statistically distinguished. This fourth group will not be reported on any further here.

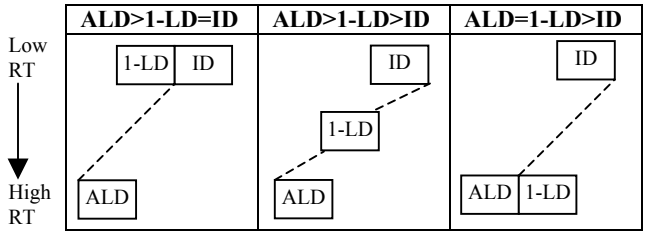


Figure 1: Depiction of the three secure priming patterns found within the data. Dashed lines indicate significant differences in response time (RT) between prime conditions.

Table 5 reports the number of participants displaying each particular priming pattern as a function of N level. The obvious feature of this data is that at Low and Medium N, participants are showing distinctly different patterns of performance; that is, the individual data differs from the grouped data.

Table 5: Number of participants as a function of priming pattern and neighbourhood (N) level.

Priming Pattern	Low N	Med N	High N
ALD>1-LD=ID	3	4	0
ALD>1-LD>ID	28	21	1
ALD=1-LD>ID	14	22	39

Note. ID = Identity; 1-LD = One-Letter-Different; ALD = All-Letters-different

At Low N, a small number of participants displayed a pattern of priming that would indicate their lexical entries for these words were poorly tuned (i.e. ALD>1-LD=ID). For these participants, the priming effects of the 1-LD and ID primes were of the same magnitude, indicating that these words were able to be accessed equally as well by a neighbour as the word itself. The majority of participants showed the priming pattern commonly found in group studies at similar N densities (i.e. ALD>1-LD>ID) (e.g. Andrews, 1997). These participants have slightly better lexical tuning than the former group, with lexical entries showing sharper but not perfect tuning for the target word. In contrast, 14 of the participants at this N level showed a finely tuned processor, which would only activate entries for an exact word match. Here the pattern shows no form-priming, even though repetition-priming was still present (i.e. ALD=1-LD>ID).

For Medium N level words there was a small shift towards increased specificity of lexical tuning, with an increase in the number of participants showing the highest level of tuning (i.e. ALD=1-LD>ID). More importantly though, there were still considerable differences between participants in the specificity of their lexical processes. This changed for High N words, with almost all the participants displaying finely tuned lexical entries for these targets (i.e. ALD=1-LD>ID).

Overall, the results shown in Table 5 indicate a general shift towards more finely tuned lexical processing with increasing N densities; a result which ties in with previous research (e.g. Forster, 1987). More importantly, however, these results demonstrate the variability between subjects in their tuning level, especially at the Low and Medium N densities; illustrated by the variation in priming patterns at these levels.

In addition to demonstrating an overall shift to fine tuning as N increases, the data obtained in this study allowed for investigation into the nature of this shift within each participant, by looking at their individual patterns of effects across N levels. Although space does not permit a full account of this data, three groups accounted for the majority of participants. The largest of these (n=16) followed the priming pattern shown in the overall group analyses. That is, these participants showed finely tuned entries for High N level targets (i.e. ALD=1-LD>ID) but slightly less tuned entries for Medium and Low N words (i.e. ALD>1-LD>ID). The second largest group (n=10) showed fine tuning (i.e. ALD=1-LD>ID) which was unaffected by changes in N level, while the third (n=6) differed from this only by having less tuning for entries at Low N densities (i.e. ALD>1-LD>ID). The remaining participants showed variations upon these themes.

Nonwords

Each participant's nonword data was analysed in the same manner as the group data, beginning with a single omnibus analysis. As with the group analyses; none of the 50 participants showed an interaction between the N Level and Prime Type factors, which was expected given that N density only tunes *lexical* representations. Forty-six participants showed a main effect of N Level and 30 showed a main effect of Prime Type. Again these effects were investigated with separate subsequent analyses.

Forty of the participants' N level effects followed the same pattern as seen in the group analyses, that is, RTs to Low N targets were significantly faster than those of the Medium or High N targets. However, there seemed to be at least some variation being masked by the group study approach. One participant's Low *and* Medium N nonword targets showed a response time advantage over High N targets, while two further participants showed RT differences between each N level (i.e. RTs increasing with N level). In addition, two participants had non-secure N level effects. Each of these variations of N level effects could be explained by the fact that with increasing N density, lexical processes have more competing entries to

eliminate before concluding a nonword. Perhaps the individual pattern variation could be accounted for by variations in sensitivity to these competitors.

Half of the participants with Prime Type effects showed a similar pattern to that seen with the group analyses; that is, an equivalent RT advantage for both the 1-LD and ID conditions. Three participants showed an RT advantage for the ID condition only, while a further 12 had non-secure effects. For these 30 participants, differences between the repetition-priming effects for nonwords overall and words at each N level were calculated (Low: M=41, SE=3.3; Medium: M=36, SE=3.2; High: M=27, SE=3.4). Consistent with the group analyses, these mean differences indicate large reductions in the magnitude of nonword priming effects, compared to word priming, for these participants. It would seem that individual differences in performance are not restricted to lexical processing, but also extend to the level of graphemic priming induced by orthographically related prime-target pairs.

Conclusions

Although space constraints precluded the reporting and interpretation of the full data from this study, one major conclusion can be drawn. Although the group data were consistent with previous studies of neighbourhood density, the Multiple Session Masked Priming results revealed the group approach was masking an array of individual differences in lexical tuning. Indeed, individual variation was also evident in the existence of pre-lexical, graphemic processing.

A simple explanation of the source of such variation may be that the internal coding of representations are handled differently between the lexical processors of individuals. This may be due to some predisposition and/or individualised set of experiences, but at this stage far more would need to be known about individual differences in other aspects of written word recognition in order to make a more specific claim.

The results of this study suggest that future research should consider the Multiple Session Masked Priming paradigm and look more carefully at other fundamental claims about lexical processing drawing solely on group outcomes, in the hope of developing more sensitive models of lexical access and written word recognition.

Acknowledgments

The authors wish to thank the reviewers of this paper for their comments. Special thanks also to Ken Forster for his invaluable suggestions.

References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval – Resolving neighbourhood conflicts. *Psychonomic Bulletin and Review, 4,* 439-461.

Bodner, G. E., & Masson, M. E. J. (1997). Masked repetition priming of words and nonwords: Evidence for a nonlexical basis for priming. *Journal of Memory and Language, 37,* 268-293.

Bodner, G. E., & Masson, M. E. J. (2001). Prime validity affects masked repetition priming: Evidence for an episodic resource account for priming. *Journal of Memory and Language, 45,* 616-647.

Byrne, C., Yelland, G. W., Johnston, M. B., & Pratt, C. (2000). Individual variation in the reliance on orthographic and phonological processes in the word recognition of skilled adult readers. 27th Australian Experimental Psychology Conference, Queensland. *Australian Journal of Psychology, 49* (Suppl.).

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI.* Hillsdale, NJ: Erlbaum.

Forster, K. I. (1987). Form-priming with masked primes: The best match hypothesis. In M. Coltheart (Ed.), *Attention and Performance XII.* Hillsdale, NJ: Erlbaum.

Forster, K. I. (1998). The pros and cons of masked priming. *Journal of Psycholinguistic Research, 27,* 203-233.

Forster, K. I., & Davis, C. (1984). Masked priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 680-698.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *Quarterly Journal of Experimental Psychology, 39,* 211-251.

Forster, K. I., & Veres, C. (1998). The prime lexicality effect: Form-priming as a function of prime awareness. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24,* 498-314.

Keppel, G. (1991). *Design and Analysis: A researchers handbook.* NJ: Prentice Hall.

Kim, J., & Davis, C. (2003). Task effects in masked cross-script translation and phonological priming. *Journal of Memory and Language, 49,* 484-499.

Kucera, H., & Francis, W. N. (1967). *Frequency analysis of English useage.* Boston: Houghton Mifflin.

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short SOAs in lexical decision and naming. *Cognition, 17,* 459-477.

Yelland, G. W., & Byrne, C. (2001, April). *Multiple session masked priming and individual differences in lexical processing.* Paper presented at the Inaugural Masked Priming Symposium, "Masked Priming: State of the Art". Maquarie University, Sydney, Australia.

NLS: A Non-Latent Similarity Algorithm

Zhiqiang Cai (zcai@memphis.edu)

Danielle S. McNamara (dsmcnamr@memphis.edu)

Max Louwerse (mlouwers@memphis.edu)

Xiangen Hu (xhu@memphis.edu)

Mike Rowe (mprowe@memphis.edu)

Arthur C. Graesser (a-graesser@memphis.edu)

Department of Psychology/Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152 USA

Abstract

This paper introduces a new algorithm for calculating semantic similarity within and between texts. We refer to this algorithm as NLS, for Non-Latent Similarity. This algorithm makes use of a second-order similarity matrix (SOM) based on the cosine of the vectors from a first-order (non-latent) matrix. This first-order matrix (FOM) could be generated in any number of ways; here we used a method modified from Lin (1998). Our question regarded the ability of NLS to predict word associations. We compared NLS to both Latent Semantic Analysis (LSA) and the FOM. Across two sets of norms, we found that LSA, NLS, and FOM were equally predictive of associates to modifiers and verbs. However, the NLS and FOM algorithms better predicted associates to nouns than did LSA.

Introduction

Computationally determining the semantic similarity between textual units (words, sentences, chapters, etc.) has become essential in a variety of applications, including web searches and question answering systems. One specific example is AutoTutor, an intelligent tutoring system in which the meaning of a student answer is compared with the meaning of an expert answer (Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, Harter, Person, & the TRG, 2000). In another application, called Coh-Metrix, semantic similarity is used to calculate the cohesion in text by determining the extent of overlap between sentences and paragraphs (Graesser, McNamara, Louwerse & Cai, in press; McNamara, Louwerse, & Graesser, 2002).

Semantic similarity measures can be classified into Boolean systems, vector space models, and probabilistic models (Baeza-Yates & Ribeiro-Neto, 1999; Manning & Schütze, 2002). This paper focuses on vector space models. Our specific goal is to compare Latent Semantic Analysis (LSA, Landauer & Dumais, 1997) to an alternative algorithm called Non-Latent Similarity (NLS). This NLS algorithm makes use of a second-order similarity matrix (SOM). Essentially, a SOM is created using the cosine of the vectors from a first-order (non-latent) matrix. This first-order matrix (FOM) could be generated in any number of ways. However, here we used a method modified from Lin (1998). In the following sections, we describe the general concept behind vector space models, describe the differences between the metrics examined, and present an evaluation of these metrics' ability to predict word associates.

Vector Space Models

The basic assumption behind vector space models is that words that share similar contexts will have similar vector representations. Since texts consist of words, similar words will form similar texts. Therefore, the meaning of a text is represented by the sum of the vectors corresponding to the words that form the text. Furthermore, the similarity of two texts can be measured by the cosine of the angle between two vectors representing the two texts (see Figure 1).

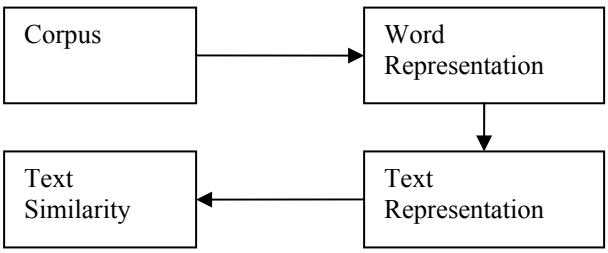


Figure 1. From Corpus to Text Similarity.

The four items of Figure 1 can be described as follows. First, the corpus is the collection of words comprising the target texts. Second, word representation is a matrix G used to represent all words. Each word is represented by a row vector g of the matrix G. Each column of G is considered a "feature". However, it is not always clear what these features are. Third, text representation is the vector $v = G^{T}a$ representing a given text, where each entry of a is the number of occurrences of the corresponding word in the text. Fourth, text similarity is represented by a cosine value between two vectors.

More specifically, Equation 1 can be used to measure the similarity between two texts represented by a and b,

respectively. For reasons of clarity, we do not include word weighting in this formula.

$$sim(a,b) = \frac{a^T GG^T b}{\sqrt{a^T GG^T a}\sqrt{b^T GG^T b}} \quad (1)$$

Latent Semantic Analysis (LSA)

LSA is one type of vector-space model that is used to represent world knowledge (Landauer & Dumais, 1997). LSA extracts quantitative information about the co-occurrences of words in documents (paragraphs and sentences) and translates this into an *N*-dimensional space. The input of LSA is a large co-occurrence matrix that specifies the frequency of each word in a document. Using singular value decomposition (SVD), LSA maps each document and word into a lower dimensional space. In this way, the extremely large co-occurrence matrix is typically reduced to about 300 dimensions. Each word then becomes a weighted vector on *K* dimensions. The semantic relationship between words can be estimated by taking the cosine between two vectors. This algorithm can be briefly described as follows.

(1) Find the word-document occurrence matrix *A* from a corpus[1].
(2) Apply SVD: $A = U\Sigma V^T$.
(3) Take the row vectors of the matrix *U* as the vector representations of words.

Non-Latent Similarity (NLS) Model

NLS is proposed here as an alternative to latent similarity models such as LSA. NLS relies on a first order, non-latent matrix that represents the non-latent associations between words. The similarity between words (and documents) is calculated based on a second-order matrix. The second order matrix is created from the cosines between the vectors for each word drawn from the FOM. Hence, for NLS, the cosines are calculated based on the non-latent similarities between the words, whereas for LSA, the similarities are based on the cosines between the latent vector representations of the words. The following section describes the components and algorithms used in NLS.

Lin's (1998) Algorithm Our starting point for NLS is Lin's (1998) algorithm for extracting the similarity of words. Similarity is based upon the syntactic roles words play in the corpus. A syntactic role is designated here as a feature. For example, "the Modifier of the NP *man*" is a feature. A word has this feature if and only if it is used as the modifier of *man* when *man* is part of an NP in the corpus. For example, if the corpus contains the phrase *the rich man*, then *rich* has the (adjectival) feature of modifying *man.* Each feature is assigned a weight to indicate the feature's importance. This algorithm is briefly described as follows.

(1) For each word base, form a feature vector.
(2) For each pair of word bases, find the similarity of two word bases from the corresponding two feature vectors.

In Lin's algorithm, the similarity is calculated according to Equation 2.

$$sim(w_1, w_2) = \frac{2 \times I(F(w_1) \cap F(w_2))}{I(F(w_1)) + I(F(w_2))} \quad (2)$$

$F(w)$ is the set of features possessed by the word w and $I(F)$ is the "information" contained in the feature set F: $I(F) = \sum_{f \in F} u(f)$. u is the weight function of the feature f.

First-Order Matrix LSA is referred to as latent because the content is not explicit or extractable after SVD. Thus, the features that two similar words share are "latent." In contrast, every feature is explicit and directly extractable from the matrix using Lin's (1998) algorithm. Hence, it is non-latent, and can be used as a first-order similarity matrix (FOM).

We created the FOM using a modification of Lin's algorithm with cosines rather than proportions. First, we parsed all of the sentences (about 2 million) in the TASA corpus using Lin's MINIPAR parser (Lin, 1998). This provided about 9 million word-relation-word triplets. Table 1 shows the triplets extracted for the sentence *People did live in Asia, however*.

Table 1: An example with word-relation-word triplets.

Word1	Relation	Word2
Live	V:s:N	people
Live	V:aux:Aux	do
Live	V:mod:Prep	in
In	Prep:pcomp-n:N	Asia
Live	V:mod:A	however

A "feature" consists of a word (e.g., Word1 or Word2) and a relation that contains a verb (V), noun (N), or modifier (A). For example, the association between the word *live* and its relation to *people*, which is "V:s:N", comprises two features (*live* - V:s:N; *people* - V:s:N). About 400,000 such features were obtained. Each feature was assigned a weight, using Lin's formula. We adopted an occurrence frequency threshold, which yielded 10363 nouns (occurrence > 50), 5687 verbs (occurrence > 5), and 6890 modifiers (occurrence > 10). For each of the selected words, a feature vector was formed according to the features it involved.

We modified Lin's method in the last step. Specifically, rather than applying Equation 2 to the feature vectors, the cosine between any two feature vectors was calculated. This provided a FOM containing the similarity between all word pairs. In addition, the FOM guarantees a property called "decomposability", which will be addressed in the next section.

[1] Hu et al. (2003, theorem 2) proved that the LSA similarity measure is a special case of (1)

Non-Latent Similarity (NLS) Algorithm The logic behind the use of a second order matrix to represent textual similarity relies on a reformulation of the algorithm used in general vector models. Specifically, Equation 1 can be rewritten as Equation 3.

$$sim(a,b) = \frac{a^T Sb}{\sqrt{a^T Sa}\sqrt{b^T Sb}} \quad (3)$$

When the columns of G are normalized to be unit vectors, S becomes a word-similarity matrix [2]. In other words, each entry of $S, s_{ij} = g_i g_j^T$, is the similarity of two words represented by row vectors g_i and g_j, respectively. Essentially, a word-similarity matrix (S) is used rather than word representation vectors (G).

From Equation 3 we can see that the similarity of two texts is determined by two factors: the word occurrences in each text and the similarity between words. Since we can do little to the occurrence vectors (other than applying word weighting), the word similarity matrix will determine the validity of the measure of text similarity. In other words, Equation 3 provides a good measure if and only if similar words have similar vector representations. If similar words have dissimilar vector representations or dissimilar words have similar representations, then the measure provided by Equation 3 is unreliable. Therefore, the verification of the validity of the word representation, at least in terms of text similarity comparison, is equivalent to the verification of the validity of the word similarity matrix (or FOM in this case).

While it is not possible to directly judge the quality of a vector representation, it is possible to judge the validity of word similarity. Provisions for such a judgment will be made in the next section of this paper.

Equation 3 raises an important question: Instead of creating the similarity matrix S by the word representation matrix G, can we find the similarity matrix by any other method that provides a better word similarity measure? One of the conditions under which this question may be answered is that the similarity matrix S, no matter how it is created, must be decomposable. That is, there exists a matrix G (we do not have to find it) such that $S = GG^T$. This condition is necessary to guarantee that the value calculated from Equation 3 ranges from -1 to 1.

The FOM that we generated by the modified Lin's method is decomposable and can therefore be used in Equation 3 for text comparison. However, that matrix is high-dimensional (N by N, where N is the total number of words). This will cause some computational complexity. To reduce the number of dimensions, we kept only the 400 largest similarity values for each word and set the other smaller values to be zero. Thus, the similarity matrix became sparse and the computational complexity was reduced. However, this made the similarity matrix un-decomposable and invalid for Equation 3.

[2] The normalization guarantees that the similarity between any two words will not exceed the similarity of a word to itself and that the values are in a known range [-1,1].

The decomposability therefore raises a new question: Is there a straightforward way to guarantee both decomposability and validity of the similarity matrix S? An easy way of guaranteeing these criteria is by using a word similarity matrix to act as a word representation matrix.

Suppose S is a word similarity matrix regardless of its creation method. Then each column vector in S contains the similarities of a particular word to all other words. Therefore, each column vector can represent the corresponding word.

Table 2. A small section of a first order matrix.

	chair	table	strength
desk	0.16	0.17	0
bed	0.14	0.13	0
speed	0	0	0.14
success	0	0	0.11

Table 2 is a small section of our FOM. It can be seen that the column vectors for *chair* and *table* are very similar to each other, but quite different from that of "strength". In the complete matrix, *desk* is the 4th nearest neighbor of (i.e., most similar to) *chair* and the 1st nearest neighbor of *table*. In addition, *bed* is the 2nd nearest neighbor of *chair* and the 5th nearest neighbor of *table* (see http://cohmetrix.memphis.edu/wordsim/wf1.aspx).

If we believe that similar words should share most nearest neighbors (a group of words that are most similar to a given word), then similar words should have similar column vectors in S. Therefore, we can create a new word similarity matrix by the cosine between the column vectors of S, $\tilde{S} = D^T S^T SD$, where D is a diagonal matrix formed by the reciprocal of the norms of the column vectors of S. We call $\tilde{S}$ the second-order word similarity matrix (SOM) and S the first-order similarity matrix (FOM). This new matrix $\tilde{S}$ is obviously decomposable and should maintain the validity of the original word similarity matrix.

If the SOM is valid, then we can form a measure based on the FOM:

$$sim(a,b) = \frac{a^T D^T S^T SDb}{\sqrt{a^T D^T S^T SDa}\sqrt{b^T D^T S^T SDb}} \quad (4)$$

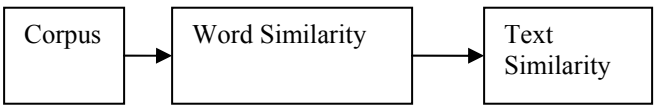


Figure 2. From Corpus to Text Similarity (SOM).

Equation 4 provides a new algorithm for text comparison, which relies solely on the similarity matrix. We call this algorithm the Non-Latent Similarity (NLS) algorithm, assuming that the FOM is non-latent. Figure 2 shows the difference between NLS and the general vector-space model. When compared with Figure 1, we can see that the "representations" are replaced by the similarity matrix.

Evaluation

In this section, we compare NLS to LSA to examine the differences between the latent analytic method exemplified by LSA and the non-latent method of NLS. We examine the validity of these two methods by examining their ability to predict word associates obtained from two sources of free association norms. We also examine the ability of the FOM to predict these word associates. The ability of FOM and NLS to predict word associates should be reflective of the overall validity of NLS to predict similarity of text corpora, which is crucial to our new algorithm shown in Equation 4.

We have two concerns. First, is our FOM valid? Second, if our FOM is valid, then will the second order similarity matrix (SOM) be valid as well? To answer these questions, we compared the validity of the following three similarity matrices generated by three different methods.

- *LSA*: The similarity matrix created from TASA corpus by LSA.
- *FOM*: The similarity matrix created from TASA corpus using the modified version of Lin's method.
- *NLS*: The second order similarity matrix based on the above *FOM*.

Our overall question addressed the ability of the three similarity metrics (LSA, FOM, and NLS) to correctly list word associates. We were also interested in examining how this ability varied as a function of several variables. First, we were interested in whether the results remained stable across norming databases. We chose to use two sets of free association norms: the Edinburgh Associative Thesaurus (EAT; Kiss, Armstrong, Milroy, & Piper, 1973) and the University of South Florida Free Association Norms (USFFAN; Nelson, McEvoy, & Schreiber, 1998).

We were also interested in how the results differed across word types (i.e., nouns, verbs, vs. adjectival/adverbial modifiers). One difference between the three classes of words is the amount of semantic contextualization. Specifically, the meaning of verbs and modifiers is usually context dependent, whereas the meaning of nouns is less dependent on the context (e. g., Graesser, Hopkinson & Schmid, 1987). For example, in the phrase *a big house*, the size of the adjectival modifier *big* depends on the noun *house*. It could be argued, moreover, that words that are more concrete are less context-dependent. Adjectives are less concrete than nouns so they would be more context-dependent. A similar argument could be made for verbs, which are more context dependent than nouns.

We expected the context-dependency factor to most affect the performance of LSA, because the success of LSA relies heavily on the occurrence of words in similar contexts, and essentially taps into that factor to assess word similarity. The basic assumption behind LSA is that words used in similar context have similar representations. Thus, if a word is less context-dependent, LSA may be less able to tap into associations.

While NLS similarly uses semantic context to compute similarity, it also uses syntactic context. The word similarities are extracted not only from the similar semantic context but also from the similar syntactic roles that the words play. That is, the FOM includes syntactic relations as features, whereas word order and the relations between words are ignored in LSA. Thus, we expected LSA to be less successful in identifying the associates of nouns as compared to modifiers and verbs. We did not expect this factor to affect the performance of NLS. We expected that FOM and NLS would be sensitive to both context based and non-context based associations.

To examine these factors, we randomly selected 135 common words, composed of 45 modifiers (including adjectives and adverbs), 45 nouns, and 45 verbs. We then determined the first most commonly listed and the second most commonly listed associate to those words, based on the association norms provided by EAT and the USFFAN. Finally, we determined whether each of the three similarity metrics listed the first and second most commonly listed associate from the respective norm database. A criterion was set in the following analyses: A metric identifies an associate of a word if, according to the metric, the associate is among the top five nearest neighbors of the word. While not extremely strict, the cutoff was intended to be relatively conservative compared to setting the cutoff at 20 words.

Results

Table 3 shows the proportion associates identified by each metric. A 3 x 2 x 2 analysis of variance (ANOVA) was conducted that included the between-words variable of word type (noun, verb, adjectival/adverbial modifier) and the within-words variables of associate (first, second) and database (EAT, USFFAN).

Table 3: Proportion of correctly identified associates listed in the top five nearest neighbors provided by LSA, FOM, and NLS as a function of the free association norms and word types.

	EAT			USFFAN		
	Mod	Noun	Verb	Mod	Noun	Verb
Associate 1						
LSA	0.40	0.11	0.16	0.31	0.07	0.13
FOM	0.40	0.36	0.13	0.31	0.31	0.16
NLS	0.38	0.36	0.11	0.27	0.36	0.13
Associate 2						
LSA	0.07	0.04	0.09	0.13	0.04	0.18
FOM	0.18	0.11	0.11	0.16	0.16	0.11
NLS	0.16	0.11	0.11	0.13	0.13	0.16

There was a main effect of word type, $F(2, 132) = 3.4$, $MSE = .471$, $p < .05$. Bonferroni Means tests indicated that the proportion of associates identified for modifiers ($M = .243$) was significantly greater than for verbs ($M = .122$), but not significantly greater than for nouns ($M = .187$). There was an effect of associate, $F(1, 131) = 19.5$, $MSE = .330$, $p<.001$, reflecting a greater proportion of first

associates identified (M = .250) than second associates (M = 0.120). There was also an interaction between word type and associate, $F(2, 131) = 4.2$, $p < .05$. This interaction reflected an effect of word type for first associates, $F(2,132) = 5.5$, $MSE = .533$, $p<.01$ ($M_{modifier} = .34$ $M_{noun} = .26$ $M_{verb} = .14$), compared to no differences between word types for second associates, $F<1$, ($M_{modifier} = .14$, $M_{noun} = .11$, $M_{verb} = .12$). Thus, the metrics were unable to identify the second associates, regardless of word type.

Finally, there was significant effect of similarity metric, $F(2,264) = 4.6$, $MSE = .139$, $p < .05$, and an interaction between metric and word type, $F(4,264) = 4.1$, $p<.01$. This interaction is depicted in Figure 3. The interaction reflects the finding that the three metrics were equally successful in identifying the associates to modifiers and verbs, whereas FOM and NLS were significantly more successful in identifying the associates to nouns than was LSA, $F(2,88)= 4.1$, $MSE = .052$, $p < .05$.

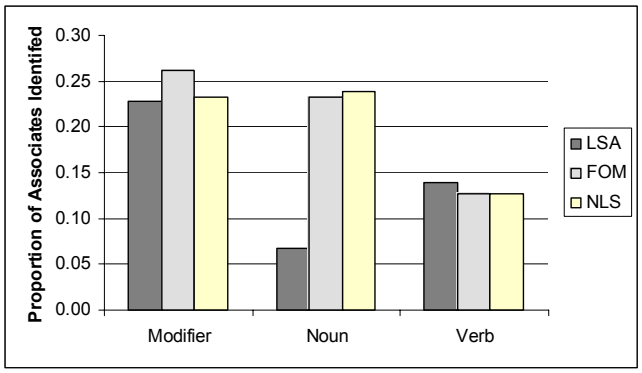


Figure 3. Proportion of associates identified (in the top 5 of the list) by the three similarity metrics.

These results did not depend on where the cutoff was drawn, (e.g., top 5 vs. top 20). Of course, the means increased with a more lax cutoff. For example, the overall accuracy of associate identification for LSA increased from 20% to 28% when the cutoff was set at 20 (i.e., when 20 of the words output by LSA were considered). Similarly, the overall accuracy for NLS increased from 27% to 42% when the cutoff was set at 20 words. Thus, there was a 140% and 157% increase respectively for LSA and NLS. The results also remained the same when word frequency was entered as a covariate. Essentially, these trends emerged regardless of how we examined the data.

There were no differences as a function of norming database. This indicates that the results we have reported should remain stable across norming databases.

Conclusions

In summary, we have provided an alternative algorithm, NLS, which makes it possible to use any non-latent similarity matrix to compare text similarity. This algorithm uses a second-order similarity matrix (SOM) that is created using the cosine of the vectors from a first-order (non-latent) matrix. This FOM could be generated in any number of ways. We used a modified form of Lin's (1998) algorithm to extract non-latent word similarity from corpora. Our evaluation of NLS compared its ability to predict word associates to the predictions made by the FOM and LSA. The critical difference between the algorithms addressed the latency of the word representations. The use of SVD results in latent word representations in LSA, whereas the use of the syntax parser in NLS results in a non-latent representation. We found that NLS, using the similarity matrix that we generated, identified the associates to modifiers and nouns relatively well. Both LSA and NLS were equally able to identify the associations to the modifiers. In contrast, none of the metrics successfully identified the associates to the verbs.

FOM versus NLS

There were two motivations for examining the results from the FOM as well as NLS. The first was to examine the validity of using a FOM. The second was to examine the correspondence in results between FOM and NLS. That is, if the FOM is valid, is the SOM valid as well? We found that NLS and FOM were equally successful in identifying all types of associations. This result indicates that SOM maintains the validity of FOM. The result supports the validity of using the NLS algorithm.

One consideration is that the second order similarity matrix may reveal new similarity relations which do not exist or are weak in the FOM. It is not hard to imagine that two words that have weak similarity in FOM may share some nearest neighbors and thus reveal a stronger relation between the two words in SOM. Nonetheless, we found here that the second-order matrix maintains the validity of FOM as much as possible, assuming the FOM is valid. When the FOM is decomposable, it can be directly used in NLS. The SOM is used when FOM is computationally heavy or is not decomposable. Our future investigations will work toward a better understanding of the situations that require a SOM as opposed to a FOM, or vice versa.

LSA versus NLS and FOM

We confirmed our predicted results that LSA would be less accurate in identifying the associates to context-independent nouns than to adjectival or adverbial modifiers, which have greater context dependency. We further predicted that this difference would not occur for NLS and the FOM. Indeed, NLS and FOM were equally predictive of noun and modifier associates. Thus, one advantage of NLS is that it makes use of both semantic and syntactic information within the text corpora. Specifically, the FOM includes both syntactic and semantic relations as features. Here, we have documented this advantage solely with respect to word similarities. However, we expect that this advantage will also improve the detection of similarity across larger bodies of text.

Verbs versus Nouns and Modifiers

One result that has baffled us is why NLS and LSA are both unable to pick up on the associates to the verbs. We

considered several explanations. First, one might think that the number of forms of the word would be a factor to consider. Since verbs tend to have more forms than do modifiers (e.g., *add* has four forms: *add, added, adding, adds*), a typical vector space model would contain relatively less information about any one form of the verb. This factor may explain the inability of LSA to identify the associates to verbs. However, it cannot do so for NLS because we used the word base, not the word itself, when forming the matrix.

We further considered that humans may have produced a greater variety of associates to verbs than to nouns or modifiers. If so, then across the two databases (i.e., EAT and USFFAN), the match between the associates in one database to another should vary as a function of word type. However, this was not the case. The two databases matched the first associate for 69% of the words, with no differences across word types. There was lower agreement (40%) and greater variance for the second associate, but not in the expected direction.

An alternative explanation regards the contextualization of verbs as compared to nouns. As we stated earlier, the meaning of verbs is more dependent on semantic context than are nouns. In addition, verbs seem to be used in a wider variety of contexts. Whereas a person can do only so much with a *chair*, the person can *sit* just about anywhere and anyhow. One can imagine eating, walking, and thinking in any number of contexts, whereas the contexts for chairs and cars are more constrained. Hence, semantic context is more variable for verbs than for nouns. This variability may render models such as NLS or LSA unable to determine the 'meaning' of verbs.

This idea is in line with notions of how verbs are represented with semantic representations. Generally, verbs are treated as the links between the concepts. Verbs constitute the relations or links between nodes. Essentially, we see here that vector space models are less able to abstract meanings of relations than the meanings of concepts.

This notion gains clarity when we examine the associates to verbs that were provided by LSA and NLS. The EAT associates to *try* are *attempt* and *again*. LSA's top five predictions were *do*, *if, you*, *can*, and *wa*y. FOM's predictions were *think*, *say*, *go*, *know*, and *ask*. We can provide many examples such as these where the associates produced by the metric make little sense. The associations predicted for nouns and modifiers, in contrast, showed obvious relationships to the target word. This observation leads us to conclude that these metrics are not able to use contextual information of verbs, perhaps because that information is not available.

Acknowledgments

The research was supported by grants from DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research (N00014-00-1-0600), National Science Foundation (SBR 9720314, REC0106965, ITR0325428), and the Institute of Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DOD, NSF or IES. The TASA corpus used was generously provided by Touchtone Applied Science Associates, Newburg, NY, who developed it for data on which to base their Educators Word Frequency Guide.

References

Baeza-Yates, R., Ribeiro-Neto, B. (Eds.) (1999). *Modern Information Retrieval*. New York , ACM Press.

Graesser, A. C., Hopkinson, P. L. & Schmid, C. (1987). Differences in interconcept organization between nouns and verbs. *Journal of Memory and Language*, *26*, 242-253.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (in press). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers.*

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8,* 129-148.

Hu, X., Cai, Z., Franceschetti, D., Penumatsa, P., Graesser, A.C., Louwerse, M.M., McNamara, D.S., & TRG (2003). LSA: The first dimension and dimensional weighting. *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 587-592). Boston, MA: Cognitive Science Society.

Kiss, G.R., Armstrong, C., Milroy, R. & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A.J. Aitkin, R.W. Bailey & N. Hamilton-Smith (Eds.), *The computer and literary studies***.** Edinburgh : University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis. *Psychological Review*, *104*, 211-240.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of International Conference on Machine Learning* (pp. 296-304), Madison, Wisconsin.

Manning, C.D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT press.

McNamara, D.S., Louwerse, M.M. & Graesser, A.C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/.

How deep are effects of language on thought? Time estimation in speakers of English, Indonesian, Greek, and Spanish

Daniel Casasanto[†] **Lera Boroditsky** **Webb Phillips** **Jesse Greene**
Shima Goswami **Simon Bocanegra-Thiel** **Ilia Santiago-Diaz**
MIT Department of Brain & Cognitive Sciences, 77 Massachusetts Avenue NE20-457
Cambridge, MA 02139 USA

Olga Fotokopoulu **Ria Pita**
Aristotle University of Thessaloniki, Greece

David Gil
Max Planck Center for Evolutionary Anthropology
Jakarta Field Station, Indonesia

Abstract

Do the languages that we speak affect how we experience the world? This question was taken up in a linguistic survey and two non-linguistic psychophysical experiments conducted in native speakers of English, Indonesian, Greek, and Spanish. All four of these languages use spatial metaphors to talk about time, but the particular metaphoric mappings between time and space vary across languages. A linguistic corpus study revealed that English and Indonesian tend to map duration onto linear distance (e.g., a *long* time), whereas Greek and Spanish preferentially map duration onto quantity (e.g., *much* time). Two psychophysical time estimation experiments were conducted to determine whether this cross-linguistic difference has implications for speakers' temporal thinking. Performance on the psychophysical tasks reflected the relative frequencies of the 'time as distance' and 'time as quantity' metaphors in English, Indonesian, Greek, and Spanish. This was true despite the fact that the tasks used entirely non-linguistic stimuli and responses. Results suggest that: (1.) The spatial metaphors in our native language may profoundly influence the way we mentally represent time. (2.) Language can shape even primitive, low-level mental processes such as estimating brief durations – an ability we share with babies and non-human animals.

Introduction

"Are our own concepts of 'time,' 'space,' and 'matter' given in substantially the same form by experience to all men, or are they in part conditioned by the structure of particular languages?" (Whorf, 1939/2000, pg. 138.) This question, posed by Benjamin Whorf over half a century ago, is currently the subject of renewed interest and debate. Does language shape thought? The answer *yes* would call for a reexamination of some foundational theories that have guided Cognitive Science for decades, which assume both the universality and the primacy of non-linguistic concepts (Chomsky, 1975; Fodor, 1975). Yet despite unreserved belief among the general public that people who talk differently also think differently (ask anyone about the Eskimos' words for snow), it has remained widely agreed among linguists and psychologists that they do not.

Skepticism about some Whorfian claims has been well founded. Two crucial kinds of evidence have been missing from many previous inquiries into relations between language and thought: objectively evaluable linguistic data, and language-independent psychological data. A notorious fallacy, attributable in part to Whorf, illustrates the need for methodological rigor. Whorf (1939) argued that Eskimos must conceive of snow differently than English speakers because the Eskimo lexicon contains multiple words that distinguish different types of snow, whereas English has only one word to describe all types. The exact number of snow words the Eskimos were purported to have is not clear. (This number has now been inflated by the popular press to as many as two-hundred.) According to a Western Greenlandic Eskimo dictionary published in Whorf's time, however, Eskimos may have had as few as two distinct words for snow (Pullum, 1991).

Setting aside Whorf's imprecision and the media's exaggeration, there remain two problems with Whorf's argument, which are evident in much subsequent 'Language and Thought' research, as well. First, although Whorf asserted an objective difference between Eskimo and English snow vocabularies, his comparative linguistic data were subjective and unfalsifiable: it is a matter of opinion whether any cross-linguistic difference in the number of snow words existed. As Geoffrey Pullum (1991) points out, English could also be argued to have multiple terms for snow in its various manifestations: *slush, sleet, powder, granular*, *blizzard, drift, etc*. The problem of unfalsifiability would be addressed if cross-linguistic differences could be demonstrated empirically, and ideally, if the magnitude of the differences could be quantified.

A second problem with Whorf's argument (and others like it in the contemporary Cognitive Linguistics literature) is that it uses purely linguistic data to motivate inferences about non-linguistic thinking. Steven Pinker illustrates the resulting circularity of Whorf's claim in this parody of his reasoning: "[Eskimos] speak differently so they must think differently. How do we know that they think differently? Just listen to the way they speak!" (Pinker, 1994, pg. 61). This circularity would be escaped if non-linguistic evidence could be produced to show that two groups of speakers who talk differently also think differently in corresponding ways.

[†]Corresponding author: Daniel Casasanto (djc@mit.edu)

But what counts as 'non-linguistic' evidence? Recent studies have tested predictions derived from cross-linguistic differences using behavioral measures such as accuracy and reaction time. Oh (2003) investigated whether Korean and English speakers would remember motion events differently, consistent with the way motion is habitually encoded in their native languages (i.e., in terms of 'path' or 'manner' of motion). Participants described videos of motion events, and then took a surprise memory test probing small details of the videos. Oh found that, as expected, English speakers used more manner-of-motion verbs in their video descriptions than Korean speakers. English speakers also performed better than Koreans on the manner-relevant portion of the memory test. Oh refers to the memory task as 'non-linguistic,' yet the questions were posed using motion language, and participants may have recalled their own verbal descriptions of the videos while responding.

Boroditsky (2001) investigated whether speakers of English and Mandarin think differently about time. English typically describes time as horizontal, while Mandarin commonly uses vertical spatiotemporal metaphors. Boroditsky found that English speakers were faster to judge sentences about temporal succession (e.g., *March comes earlier than April*) when primed with a horizontal spatial event, but Mandarin speakers were faster to judge the same sentences when primed with a vertical spatial stimulus. This was true despite the fact that all of the sentences were presented in English.

These studies by Oh and Boroditsky support a version of the Whorfian hypothesis which Slobin (1986) has termed *thinking for speaking*: language can affect thought when we are thinking with the intent to use language, plausibly by directing attention to elements of our experience that are ordinarily encoded in the language we speak. For instance, because English tends to encode information about manner of motion more often than Korean does, Oh's English subjects may have automatically attended to the manner information in the videos more than her Korean subjects did. Some researchers have characterized the effects of thinking for speaking as uninterestingly weak (Pinker, 1994; Papafragou, Massey, & Gleitman, 2002). Results such as Oh's and Boroditsky's suggest otherwise: at minimum, thinking for speaking appears to influence ubiquitous cognitive processes such as attention and memory, and is capable of changing the nature of our abstract mental representations. Furthermore, habits formed while thinking for speaking are likely to be practiced even when people are not explicitly encoding information for language. We never know when we might want to talk about an event at some later point, so it is in our best interest to encode language-relevant details as a matter of policy.

Can the influence of language on thought go beyond thinking for speaking? Much of our mental life is unutterable: what words can capture the sound of a clarinet, or explain the color *red* to a blind person (Locke, 1689/1995; Wittgenstein, 1953)? Can peculiarities of our native language shape even the deep, primitive kinds of representations that we share with pre-linguistic infants and non-human animals? Previous research suggests that language can affect our high-level linguistic and symbolic representations in the abstract domain of time (Boroditsky, 2001). The goal of the present study was to find out whether language can also shape our low-level, non-linguistic, non-symbolic temporal representations. A linguistic study was conducted to investigate a previously unexplored pattern in spatiotemporal metaphors, and to quantify cross-linguistic differences in the way these metaphors are used by speakers of English, Indonesian, Greek, and Spanish (Experiment 1). To determine whether these cross-linguistic differences have consequences for speakers' non-linguistic time representations, the results of the linguistic study were used to predict performance on a pair of psychophysical time estimation tasks, with entirely non-linguistic stimuli and responses (Experiments 2 and 3).

Experiment 1:
Time in a Bottle or Time on the Line?

Linguists have noted that spatial metaphors are often used to talk about non-spatial phenomena -- in particular abstract phenomena such as social rank (e.g., a *high* position), mathematics (e.g., a *low* number), and time (e.g., a *long* vacation) (Clark, 1973; Gibbs, 1994; Jackendoff, 1983; Lakoff & Johnson, 1980). Recently, psychologists have begun to explore the proposal that these metaphors in language provide a window on our underlying mental representations in abstract domains, using the domain of time as a testbed (Boroditsky, 2000, 2001; Boroditsky & Ramscar, 2002; Casasanto & Boroditsky, 2003; Gentner, 2001). In general, this work has focused on how time can be expressed (and by hypothesis conceptualized) in terms of linear space. Linear spatiotemporal metaphors are pervasive in English, and are used to talk about various aspects of time, including succession (e.g., Monday comes *before* Tuesday), motion through time (e.g., Let's move the meeting *forward*), and duration (e.g., a *short* intermission). But is time necessarily conceptualized in terms of uni-dimensional space? English speakers also talk about *oceans of time, saving time in a bottle*, and compare epochs to *sand through the hourglass*, apparently mapping time onto volume.

Experiment 1 compared the use of 'time as distance' and 'time as quantity' metaphors across languages. Every language examined so far uses both distance and quantity metaphors, but their relative prevalence and productivity appear to vary markedly. In English, it is natural to talk about *a long time*, borrowing the structure and vocabulary of a spatial expression like *a long rope.* Yet in Spanish, the direct translation of 'long time', *largo tiempo*, sounds awkward to speakers of most dialects. *Mucho tiempo*, which means 'much time', is preferred.

In Greek, the words *makris* and *kontos* are the literal equivalents of the English spatial terms *long* and *short*. They can be used in spatial contexts much the way *long* and *short* are used in English (e.g., *ena <u>makry</u> skoini* means 'a

long rope'). In temporal contexts, however, *makris* and *kontos* are dispreferred in instances where *long* and *short* would be used naturally in English. It would be unnatural to translate *a long meeting* literally as *mia makria synantisi*. Rather than using distance terms, Greek speakers typically indicate that an event lasted a long time using *megalos*, which in spatial contexts means physically 'large' (e.g., a big building), or using *poli*, which in spatial contexts means 'much' in physical quantity (e.g., much water). Compare how English and Greek typically modify the duration of the following the events (literal translations in parentheses):

1e. long night
1g. megali nychta (*big night*)
2e. long relationship
2g. megali schesi (*big relationship*)
3e. long party
3g. parti pou kratise poli (*party that lasts much*)
4e. long meeting
4g. synantisi pou diekese poli (*meeting that lasts much*)

In examples 1g. and 2g., the literal translations might surprise an English speaker, for whom *big night* is likely to mean 'an exciting night', and *big relationship* 'an important relationship'. For Greek speakers, however, these phrases communicate duration, expressing time not in terms of uni-dimensional space, but rather in terms of physical quantity (i.e., three-dimensional space).

For Experiment 1, the most natural phrases expressing the ideas 'a long time' and 'much time' were elicited from native speakers of English, Indonesian, Greek, and Spanish (see table 1). The frequencies of these expressions were compared in a very large multilingual text corpus: www.google.com. Each expression in table 1 was entered as a search term. Google's language tools were used to find exact matches for each expression, and to restrict the search to web pages written only in the appropriate languages.

Table 1: Distance and quantity metaphors for duration.

	Distance	Quantity
English	long time	much time
Indonesian	waktu panjang	waktu banyak
Greek	makry kroniko diatstima	poli ora
Spanish	largo tiempo	mucho tiempo

Results The number of google 'hits' for each expression was tabulated, and the percentage of distance hits and quantity hits was calculated for each pair of expressions, as a measure of their relative frequency (see figure 1). Results showed that in English and Indonesian, distance metaphors were more frequent than quantity metaphors. The opposite pattern was found in Greek and Spanish. A Chi-Square test showed that the relationship between distance and quantity metaphors varied significantly across languages (X^2=8.5x10^5,df=3, p<0.0001). These findings corroborate native speakers' intuitions for each language. Additional quantitative studies are in progress to validate these results.

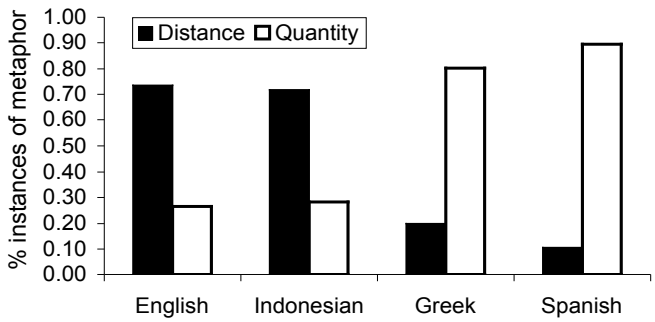


Figure 1: Corpus search results. Black bars indicate the percentage of distance metaphors and white bars the percentage of quantity metaphors found for each language.

Experiments 2 & 3:

Do People Who Talk Differently Think Differently?

How might this difference in the way English, Indonesian, Greek, and Spanish speakers talk about time affect the way they think about it? Linguists and psychologists have argued that our conception of time is intimately dependent on our knowledge of space, noting that in many languages, people can hardly avoid using spatial words when they talk about time (Clark, 1973; Gibbs, 1994; Jackendoff, 1983; Lakoff & Johnson, 1980). Behavioral studies show that changing someone's immediate spatial environment or frame of reference can dramatically change the way they interpret temporal language (Boroditsky, 2000; Boroditsky & Ramscar, 2002). But does space influence our temporal thinking even when we are not *thinking for speaking*?

A recent study by Casasanto & Boroditsky (2003) shows that space influences even our low-level, non-linguistic, non-symbolic representations of time. English speakers watched lines 'growing' across a computer screen, one pixel at a time, and estimated either how far they grew or how much time they remained on the screen. Estimates were made by clicking the mouse to indicate the beginning and end of each spatial or temporal interval. Line distances and durations were varied orthogonally, so there was no correlation between the spatial and temporal components of the stimuli. As such, one stimulus dimension served as a distr actor for the other: an irrelevant piece of information that could potentially interfere with task performance. Patterns of cross-dimensional interference were analyzed to reveal relationships between subjects' representations of space and time. Results showed that subjects were unable to ignore irrelevant spatial information when estimating time (even when they were encouraged to do so). Line stimuli of the same average duration were judged to take a longer time when they grew a longer distance, and a shorter time when they grew a shorter distance. In contrast, line duration did not affect subjects' distance estimates. This asymmetric relation between space and time was predicted based on patterns in language: we talk about time in terms of space more than we talk about space in terms of time (Lakoff & Johnson, 1980).

These findings suggest that the metaphoric relationship between time and space is not just linguistic, it is also conceptual. Not only do people talk about time in terms of space, they also think about time using spatial representations. However, the experiments reported in Casasanto & Boroditsky (2003) leave the Whorfian question unanswered: do metaphors in language merely reflect underlying conceptual structures, or might the metaphors we use also play some role in constructing concepts, or establishing their interrelations?

In the present study speakers of four different languages performed the same pair of non-linguistic psychophysical tasks, which required them to estimate time while overcoming spatial interference. It was reasoned that if people's concepts of time and space are substantially the same universally, irrespective of the languages they speak, then performance on these tasks should not differ between language groups. If, on the other hand, the spatiotemporal metaphors people use affect how they represent time and space non-linguistically, then performance should vary in ways predicted by participants' language-particular metaphors.

For Experiment 2, subjects performed a 'growing line' task similar to the task described above. It was reasoned that the English participants in our previous study may have suffered interference of distance on duration estimation, in part, because these notions are conflated in the English language. It is hard to imagine expressing the idea 'a long time' in English without using an adjective that can also indicate spatial extent. Piaget (1927) made a similar suggestion when he observed that young French speaking children often mistook distance for duration, noting that both of these concepts are commonly described in French using the adjective *longue*. We predicted that speakers of 'Distance Languages' (i.e., English and Indonesian) would show a considerable effect of distance on time estimation when performing the growing line task.

If conflations in language contribute to confusions in thought, can distinctions in language help speakers distinguish closely related concepts? We reasoned that speakers of languages that do not ordinarily express duration in terms of distance might have an easier time distinguishing the spatial and temporal information conveyed in our growing line stimuli. We predicted that speakers of 'Quantity Languages' (i.e., Greek and Spanish) would show only a mild effect of distance on time estimation when performing the growing line task.

For Experiment 3, a task complementary to the growing line task was developed. Subjects watched a schematically drawn container of water filling up, one row of pixels at a time, and estimated either how full it became or how much time it remained on the computer screen, using mouse clicks. We predicted the converse pattern of behavioral results for the filling container task as for the growing line task: speakers of Quantity Languages would show a considerable influence of 'fullness' on time estimation, whereas speakers of Distance Languages would show a milder effect.

Methods for Experiment 2: Growing Lines

Subjects A total of 65 subjects participated in exchange for payment. Native English and Spanish speaking participants were recruited from the Greater Boston community, and were tested on MIT campus. Native Indonesian speakers were recruited from the Jakarta community, and were tested at the Cognation Outpost in the Jakarta Field Station of the Max Planck Center for Evolutionary Anthropology. Native Greek speakers were recruited from the Thessaloniki community, and tested at the Aristotle University of Thessaloniki.

Materials Lines of varying lengths were presented on a computer monitor (resolution=1024x768 pixels, dpi=72), for varying durations. Durations ranged from 1000 milliseconds to 5000 milliseconds in 500 millisecond increments. Displacements ranged from 100 to 500 pixels in 50 pixel increments. Nine durations were fully crossed with nine displacements to produce 81 distinct line types. Lines started as a single point and 'grew' horizontally across the screen one pixel at a time, from left to right along the vertical midline. Each line remained on the screen until its maximum displacement was reached.

Written instructions were given prior to the start of the task, in the native language of the participant. Care was taken to avoid using distance metaphors for time. The task itself was entirely non-linguistic, consisting of lines (stimuli) and mouse clicks (responses).

Procedure Participants viewed 162 growing lines, one line at a time. Immediately before each trial, a prompt appeared indicating that the subject should attend either to the line's duration or to its spatial displacement. Space trials and time trials were randomly intermixed.

To estimate displacement, subjects clicked the mouse once on the center of an 'X' icon, moved the mouse to the right in a straight line, and clicked the mouse a second time to indicate they had moved a distance equal to the maximum displacement of the stimulus. To estimate duration, subjects clicked the mouse once on the center of an 'hourglass' icon, waited the appropriate amount of time, and clicked again in the same spot, to indicate the time it took for the stimulus to reach its maximum displacement.

All responses were self-paced. Importantly, for a given trial, subjects reproduced either the displacement or the duration of the stimulus, never both. Response data were collected for both the trial-relevant and the trial-irrelevant stimulus dimension, to ensure that subjects were following instructions.

Methods for Experiment 3: Filling Containers

Subjects A total of 74 subjects participated in exchange for payment. Subjects were recruited at the same time as those who participated in Experiment 2, from the same populations.

Materials and procedure The filling container task was closely analogous to the growing line task (Experiment 2). Participants viewed 162 containers, and were asked to imagine that each was a tank filling with water. Containers were simple line drawings, 600 pixels high and 500 pixels wide. Empty containers filled gradually, one row of pixels at a time, for varying durations and 'volumes,' and they disappeared when they reached their maximum fullness. Nine durations were fully crossed with nine volumes to produce 81 distinct trial types. For each trial, participants estimated either the amount of water in the container (by clicking the mouse once at the bottom of the container and a second time at the appropriate 'water level'), or they estimated the amount of time that the container took to fill (by clicking the hourglass icon, waiting the appropriate time, and clicking it again, as in Experiment 2). Durations ranged from 1000 milliseconds to 5000 milliseconds in 500 millisecond increments. Water levels ranged from 100 to 500 pixels, in 50 pixel increments.

As before, written instructions were given prior to the start of the task, in the native language of the participant. Care was taken to avoid using quantity metaphors for time. The task itself was entirely non-linguistic, consisting of containers (stimuli) and mouse clicks (responses).

Results for Experiments 2 and 3

Both time estimates and space estimates were collected for each subject, but since our present hypothesis concerns effects of language and space on time estimation, only data for the time estimation trials are reported here. (See Casasanto & Boroditsky, 2003 for a discussion of subjects' space estimation in a related task.)

Time estimation, within-domain effects Overall, for both Experiments 2 and 3 subjects' time estimates were highly accurate across all language groups, as indicated by the strong correlations between the actual stimulus duration and subjects' estimated stimulus duration. These correlations did not differ significantly between groups or between tasks (see table 2).

Table 2: Time estimation results for Experiments 2 and 3. Slope and r-square values for the correlation between actual stimulus duration and subjects' grand averaged estimates of stimulus duration. (Perfect performance would be indicated by a slope of 1.00 and r^2 of 1.00.)

	Growing Lines		Filling Containers	
	Slope	r2	Slope	r2
English	0.75	0.999	0.76	0.999
Indonesian	0.71	0.996	0.68	0.997
Greek	0.74	0.995	0.77	0.996
Spanish	0.72	0.996	0.71	0.995

Time estimation, cross-domain effects The within-domain results reported in table 2 are important to establish that subjects were able to estimate time well, and importantly, that subjects estimated time about equally well in all groups, and on both tasks. Of principal interest, however, are the cross-domain effects (i.e., effects of actual distance and actual quantity on estimated time), which are summarized in figure 2.

To investigate the effect of spatial interference on time estimation, grand averaged time estimates in milliseconds were plotted as a function of actual stimulus displacement in pixels (i.e., line length or water level). A line of best fit was computed, and the slope was used as an index of effect strength. In our previous experience with similar tasks, we found the strongest linear effects on the dependent variable (i.e., estimated time) near the middle of the range of the independent variable (i.e., actual stimulus displacement), possibly due to 'endpoint effects' commonly observed in magnitude estimation tasks. For the analyses reported here, the outer points were removed, and the middle five points of the correlations were analyzed.

Cross-domain effects varied markedly across language groups. For the growing line task, English and Indonesian speakers showed a strong effect of distance on time estimation (English: Slope=1.49, r^2=0.98; t=8.5; df=3; $p<0.001$; Indonesian: Slope=1.40, r^2=0.80; t=3.4; df=3; $p<0.01$). By contrast, Greek and Spanish speakers showed weak, non-significant effects of distance on time estimation (Greek: Slope=0.47, r^2=0.33; t=1.2; df=3; *ns*; Spanish: Slope=0.20, r^2=0.13; t=0.7; df=3; *ns*).

For the filling container task, the opposite pattern of results was found. English and Indonesian speakers showed a weak, non-significant effect of volume on time estimation (English: Slope=0.18, r^2=0.12; t=0.6; df=3; *ns*; Indonesian: Slope=0.13, r^2=0.51; t=1.7; df=3; *ns*), whereas Greek and Spanish speakers showed strong effects of volume on time estimation (Greek: Slope=1.24, r^2=0.95; t=6.9; df=3; $p<0.001$; Spanish: Slope=1.16, r^2=0.97; t=8.5; df=3; $p<0.001$).

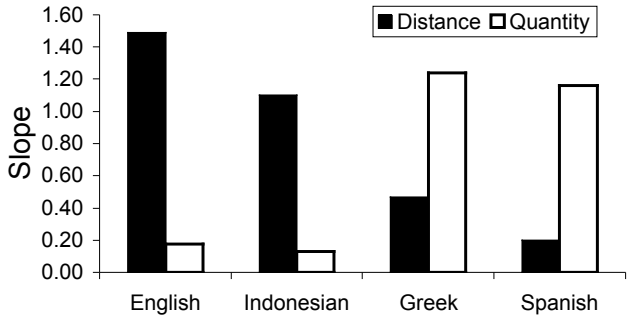


Figure 2: Effects of distance and quantity interference on time estimation.

A 4 x 2 factorial ANOVA with Language and Task as between-subject factors revealed a highly significant Language by Task interaction ($F(3,139) = 5.25$, $p<0.002$), with no main effects, signaling a true crossover interaction.

Linear regression analysis revealed a highly significant positive relation between the frequency of Distance and Quantity metaphors in each language (as measured in Experiment 1) and the amount of Distance and Quantity interference on time estimation (as measured in Experiments 2 and 3) (Slope=1.62, r^2=0.84; t=5.6; df=6; $p<0.001$).

General Discussion and Conclusions

Do people who talk differently also think differently? Performance on a pair of psychophysical time estimation tasks differed dramatically for speakers of different languages, in ways predicted by their language-particular spatiotemporal metaphors. The effects of distance interference and quantity interference on time estimation in speakers of English, Indonesian, Greek, and Spanish corresponded strikingly to the relative prevalence of distance metaphors and quantity metaphors found in these languages (compare figures 1 and 2). This was true despite the fact that the behavioral tasks comprised entirely non-linguistic stimuli and responses.

Returning to the question of Whorf's posed in the introduction, it is possible that our concepts of time and space are "given in substantially the same form by experience" to all of us, *and also* that they are "in part conditioned by the structure of particular languages." Perhaps people learn associations between time and space via physical experience (e.g., by observing moving objects and changing quantities). Since presumably the laws of physics are the same in all language communities, pre-linguistic children's conceptual mappings between time, distance, and quantity could be the same universally. When children acquire language, these mappings could be adjusted, plausibly by a process analogous to Hebbian learning: each time we use a linguistic metaphor, we may invoke the corresponding conceptual mapping. Speakers of Distance Languages then would invoke the time-distance mapping frequently, eventually strengthening it at the expense of the time-quantity mapping (and vice-versa for speakers of Quantity Languages). Alternatively, experience may not teach us to map time onto space. It could be language that causes us to notice structural parallels between these domains, in the first place. On this possibility, language would be responsible for establishing the time-distance and time-quantity conceptual mappings evident in our adult subjects, not just for modifying these mappings. Studies are in progress on young learners of Distance and Quantity languages to explore these possibilities.

The findings we present here are difficult to reconcile with a 'universalist' view of language-thought relations according to which language calls upon pre-formed, antecedently available non-linguistic concepts, which are presumed to be "universal" (Pinker, 1994, pg. 82) and "immutable" (Papafragou, Massey, & Gleitman, 2002, pg. 216). Rather, these results support what we might call a *deep* version of the linguistic relativity hypothesis (to distinguish it from the so-called *weak* version which posits that language affects 'thinking for speaking,' and from *strong* linguistic determinism). The particular languages that we speak can influence not only the representations we build for the purpose of speaking, but also the non-linguistic representations we build for remembering, acting on, and perhaps even perceiving the world around us.

Acknowledgments

Thanks to the citizens of Cognation. Research supported in part by fellowships from the NSF and the Vivian Smith Advanced Studies Institute to DC, and an NSF grant to LB.

References

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition, 75*(1), 1-28.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognit Psychol, 43*(1), 1-22.

Boroditsky, L., & Ramscar, M. (2002). The Roles of Body and Mind in Abstract Thought. *Psychological Science, 13*(2), 185-189.

Casasanto, D., & Boroditsky, L. (2003). *Do we think about time in terms of space?* Paper presented at the 25th Annual conference of the Cognitive Science Society, Boston, MA.

Chomsky, N. (1975). *Reflections on Language*. New York: Pantheon.

Clark, H. H. (1973). Space, Time, Semantics and the Child. In T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.

Gentner, D. (2001). Spatial Metaphors in Temporal Reasoning. In M. Gattis (Ed.), *Spatial Schemas and Abstract Thought*. Cambridge: MIT Press.

Gibbs, R. W., jr. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge: Cambridge University Press.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Locke, J. (1689/1995). *An Essay Concerning Human Understanding*. Amherst: Promethius Books.

Oh, K. (2003). *Habitual patterns of language use and thinking for speaking: a Whorfian effect on motion events.* University of California, Berkeley, Berkeley.

Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle, 'n' roll: the representation of motion in language and cognition. *Cognition, 84*, 189-219.

Piaget, J. (1927/1969). *The Child's Conception of Time*. New York: Ballantine Books.

Pinker, S. (1994). *The Language Instinct*. New York, US: William Morrow and Company.

Pullum, G. K. (1991). *The Great Eskmo Vocabulary Hoax and other irreverent essays on the study of language*. Chicago: University of Chicago Press.

Slobin, D. (1986). From "THOUGHT AND LANGUAGE" to "THINKING FOR SPEAKING" in *Rethinking Linguistic Relativity*, Grumperz & Levinson, eds.

Whorf, B. (2000). *Language, Thought and Reality*. Cambridge, MA: MIT Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell

Grammatical Processing Using the Mechanisms of Physical Inference

Nicholas L. Cassimatis (cassimatis@itd.nrl.navy.mil)
Naval Research Laboratory, Code 5513
4555 Overlook Ave. SW
Washington, DC 20375

Abstract

Although there is considerable evidence that humans use the same mechanisms for linguistic and nonlinguistic cognition, the thesis of linguistic modularity will remain plausible so long as well-established formal properties of syntax remain unexplained in terms of domain-general cognitive mechanisms. This paper presents several dualities between the formal structure of syntax and cognitive structures used to represent the physical world. These dualities are used to construct a cognitive model of syntactic parsing that uses only the mechanisms required for infant physical reasoning. The model demonstrates how a formal syntactic constraint, the c-command condition on binding, can be explained by a cognitive process used in physical reasoning. Several consequences for language development and the doctrine of linguistic modularity are considered.

Introduction

Although there is extensive evidence that humans use the same or similar mechanisms for linguistic and nonlinguistic cognition, the precise manner in which nonlinguistic cognitive processes are related to the formal properties of human grammar have yet to be determined.

In the field of linguistic semantics, several researchers have noticed extensive parallels between physical and abstract semantic fields. For example, Jackendoff (1990) has formalized the semantics of many verbs with primitive conceptual structures such as *GO*, *TO*, *FROM*, *PATH*, etc. Leonard Talmy (1985) has shown that semantic fields for psychological, social, argumentative and many other domains involve notions of force dynamics that underlie semantic fields for physical domains. Cognitive psychologists (e.g, Boroditsky, 2001; Spelke & Tsivkin, 2001) have found that the way in which language represents a concept can influence cognition using that concept. Clark's (1996) work culminates a long tradition beginning in the philosophy of language that analyzes language use as a species of social interaction. Bloom (2000) presents evidence that children use cognitive abilities that exist for nonlinguistic purposes to learn the meaning of words.

Some researchers (e.g., Langacker, 1999) have explored the interaction of grammar and nonlinguistic cognition by advancing a "cognitive grammar" research program that views the grammatical structure of sentences as the result of the process which maps linear sequences of words into nonlinear cognitive structures. Although the research has explained many linguistic phenomena, it has not shown in detail how this transformation explains specific syntactic constraints such as the empty-category principle, subjacency and the anaphoric binding principles that occur in some form in most mature formal theories of human syntax. Until these apparently peculiar formal properties are accounted for using general cognitive mechanisms, the thesis that humans use different mechanisms for syntactic and nonlinguistic processing remain plausible.

This paper outlines a mapping of structures in formal grammar to cognitive structures used to represent physical events, shows how to use this mapping to construct a model of human syntactic parsing that uses only the mechanisms of a model of infant physical reasoning and demonstrates how this model explains a universal, putatively innate and language-specific grammatical constraint in terms of domain-general cognitive mechanisms.

Grammatical Structure	Cognitive structure
Word, phrase, sentence	Event
Constituency	Meronomy
Phrase structure constraints	Constraints among (parts of) events
Word/phrase category	Event category
Word/phrase order	Temporal order
Phrase attachment	Event identity
Coreference/binding	Object identity
Traces	Object permanence
Short- and long-distance dependencies	Apparent motion and long paths.

Table 1. Dualities between elements of physical and grammatical structure.

Structural Dualities

The structures of grammar and of naïve physics appear more similar when a verbal utterance is conceived as an event that is composed of a sequence of word utterance subevents. Like physical events, verbal events belong to categories, combine to form larger verbal events and are ordered in relation to other verbal events according to lawful regularities. This section examines these dualities in detail, and shows that many grammatical structures have analogues to nonlinguistic cognitive structures. These dualities are summarized in Table 1.

Notation

In order to explain the mapping between syntactic structure and cognitive structures used to represent the physical world, it will be helpful to use a formal notation for representing physical events. This paper uses a notion based on the notation Cassimatis (2002) uses to present

problems to his model of physical reasoning. Although there is no claim that the notation resembles the mind's representations for syntactic or physical structure, the next section will show how to use this formalism to present sentences to a model physical reasoning so that the model can use its own representations and processes to infer the syntactic structure of sentences.

In this formalism, events, objects and places have names. Predicates describe attributes on and relations among named entities. For example, an event in which an object, `x`, moves from `p1` to `p2` during the temporal interval, `t`, is indicated with the following propositions: `Category(e, MotionEvent)`, `Agent(e, x)`, `Origin(e, p1)`, `Destination(e, p2)`, `Occurs(e, t)`. Intervals are ordered using Allan's (1983) temporal relations. For example, `Before(t1,t2)` indicates that `t1` finishes before `t2` begins and `Meets(t1, t2)` indicates that `t1` ends precisely when `t2` begins. Category hierarchies are described using subcategory relationships, e.g., `Subcateogry(Fly, MotionEvent)`. `PartOf(e1,e2)` indicates that event `e1` is part of event `e2`. That two names for events, objects or places refer to the same object is indicated using an identity relationship. For example, `Same(o1,o2)` indicates that "`o1`" and "`o2`" name the same object. Finally, regularities between physical events can be expressed using material implication. For example, that an unsupported object falls is indicated:

```
Location(o,p1,t1) + Below(p2,p1) +
Empty(p2,t1)
→
Category(e,MotionEvent) + Origin(e,p1)
+ Destination(e,p2) + Occurs(e,t2) +
Meets(t1,t2).
```

With this background, it is now possible to describe several dualities between syntactic and physical structure.

Physical and verbal event perception both have a linear order.

Although human vision has two-dimensional access to a three-dimensional physical world, there is a linear order to human perception. People can attend to only one region of space at a time. Large, complicated and/or spatially distributed visual scenes must be perceived through a series of attentional foci. For example, a person standing between two houses, A and B, can perceive that A is to the left of B by turning to the left, focusing on house A, turning to the right and focusing on house B. Likewise we perceive verbal utterances as a linear sequence of word utterance events.

Further, such multidimensionality as there is in the visual system is not unique to it. Spoken phonemes have a multidimensional character. In most phonological theories phonemes are points in a multi-dimensional vector space with dimensions such as "voiced" or "nasal".

Thus, both perceiving physical scenarios and perceiving spoken utterances involve a linear sequence of foci that each integrate multiple dimensions of information.

Utterances are events.

The philosophical tradition of "speech act theory", (which is psycholinguistically implemented by Clark (1996)), holds that linguistic utterances are actions used to achieve goals. In this way, words are similar to other nonlinguistic actions such as gesturing or tool use. Other people's actions are events we must perceive in order to interpret their intent. Both verbal and nonverbal events occur over temporal intervals. Like nonverbal events, verbal utterances can be executed with various manners (hastily, carefully, loudly, softly).

Thus, the same concepts used to describe physical events can be used to describe verbal utterances. For example, using the present notation, the utterance of the word "dog" at time, `t`, may be represented, `Category(e, dog-utterance)`, `Occurs(e, t)`.

Word order is temporal order.

The temporal order of a set of physical events has important consequences for their ultimate result. For example, pulling a gun's trigger *before* loading it results in a much different event from the pulling its trigger *after* loading it. This is also a fundamental feature of grammar: the result (in terms of its effect on the listener) of uttering "The dog", uttering "bit" and then uttering "John" is much different from the result of uttering "John", "bit" and then "the dog". In our notation, "John bit the dog" is represented as sequence of utterance events:

```
1. Category(e1, JohnUtterance)
   Occurs(e1, t1)
2. Category(e2, BitUtterance)
   Occurs(e2, t2)
   Meets(t1, t2)
```

Etc.

Physical and linguistic events both belong to categories, which exist in hierarchies.

Word and phrase categories are an important component of almost every serious syntactic theory and especially important in some (e.g., Pollard & Sag, 1994). Categories are also an essential part of most every other domain of cognition. The previous subsection demonstrated that the same `Category` predicate that represents the category of a physical event can represent the category of a word or phrase utterance. Likewise, just as physical categories exist in hierarchies (e.g., `Subcategory( RunningEvent, MotionEvent)`, so do verbal and phrasal categories (e.g., `Subcategory(CommonNoun, Noun)` and `Subcategory( TransitiveVerbPhrase, VerbPhrase)`).

Just as the category of a physical event determines which other events it occurs with (e.g., a gun-firing event tends to be preceded by a trigger-pulling event), so does the category of a word or phrase determine the distribution of words and phrases (e.g., transitive verbs are often followed by noun phrases).

Constituency is a meronomic relationship.

Physical events combine into larger events, which themselves can combine into even larger events. Word utterance events combine into phrase utterance events which combine into larger phrase utterance events. Meronomy[1] is thus a feature of both physical and verbal events. Predicates for representing physical event meronomy can capture phrasal constituency. For example, the noun phrase "the dog" can be represented thus:, `Category(e, CommonNounPhrase)`, `Category(e1, Determiner)`, `Occurs(e1, t1)`, `Category(e2, CommonNoun)`, `Occurs(e2, t2)`, `PartOf(e1,e)`, `PartOf(e2,e)`, `Meets(e1,e2)`.

The same notation for expressing physical regularities can be used to represent phrase structure rules and constraints. For example, a rule for a transitive verb's arguments can be expressed thus:

```
Category(verb, TransitiveVerb) +
Occurs(verb, t-verb)
→
Exists(object) + Category(object,
NounPhrase) + Occurs(object, t-object)
+ Before(t-verb, t-object).
```

Coreference and binding are object-identity relationships.

Coreference and binding are perhaps the most obvious identity relationships in language. Consider the following sentence, where "the dog" refers to an object, `d`, "the cat" refers to an object, `c`, and "it" refers to an object, `i`:

The dog chased the cat through the park where it lives.

"it"'s reference is ambiguous. It can refer to the dog (`Same(i,d)`), to the cat (`Same(i,c)`) or to some other object in the conversation or the environment (`Same(i,?)`). In each case, the coreference is just a special kind of identity relationship.

Identity is an extremely widespread and important relationship in everyday physical reasoning. When we lose visual contact with an object because we turn our gaze or because it is occluded and then see a similar object, we must decide whether the sightings are of the same object. Many infant reasoning experiments test for sensitivity to a physical constraint (e.g., continuity (Kestenbaum et al., 1987) or category persistence (Xu & Carey, 1999)) by testing whether infants are surprised by identities that violate those constraints.

Phrase attachment is an event identity relationship.

The occurrence of a physical event often implies the occurrence of another physical event. For example, when an object resting on shelf falls to the floor (event `f`), there must have been an event (`p`) which pushed the object off the shelf. One can infer the pushing event from the falling event even if the pushing event is not visible. Later, after observing marks left by a cat's claws on the shelf, we can infer a cat walking event (`w`). If this cat walking event occurred near the original location of the object that fell, then the cat walking event might be *identical* to the pushing event, i.e., `Same(p,w)`.

Event identity is an important feature of grammar as well. For example, the existence of a prepositional phrase utterance within a sentence utterance implies the existence of a noun or verb utterance that the prepositional phrase is an argument or adjunct of. For example, in the sentence "John saw the man with the telescope", the "with the telescope" utterance event implies the existence of an utterance event, `pp-head`, which takes "with the telescope" as an argument or adjunct. In this case, `pp-head` might be the "John" or "man" utterance event. More formally, either one of the following propositions might be true: `Same("John",pp-head)` or `Same("the man", pp-head)`. Thus phrase attachment and attachment ambiguity are instances of event identity and uncertainty about event identity.

Traces and Object Permanence

In many clauses, the arguments of a word are spoken. For example, in (1), the subject and object phrases of "eats" are spoken directly before and after it:

(1) [The man] eats [steak].

In some cases, however, the argument of a word is not spoken near it. For example, in (2), the subject noun phrase of "eats" is not adjacent to it. Sentence (3) show this distance is not an obstacle to "eats" requiring its subject to be third-person singular.

(2) The man John said [_ eats [steak]] was wearing a hat.
(3) *The men John said [_ eats [steak]] was wearing a hat.

Thus, even though the subject of "eats" is a "long distance" from it, that subject's character is constrained by or "depends" on "eats". Such relationships are often called "long-distance dependencies". Some grammatical theories posit the existence of a "trace" that is the subject of "eats"

[1] Meronomy is the study of the relationships of an entity with its parts.

and is left when "the man" is "moved" out of that subject position to another part of the sentence.

Invisible events such as traces and long-distance dependencies are common features of physical inference. For example, if a ball is rolled behind a screen on a flat table and fails to roll out from the other end of the screen, one can posit the existence of a second object behind the screen blocking the first object and make inferences about it (for example, that it is large and massive enough to stop the rolling ball) without ever perceiving the obstacle itself. Thus just as understanding language (depending on one's favorite syntactic theory) requires reasoning about phonetically unrealized phrases, physical event understanding often requires one to reason about events that are not "visually realized", i.e., perceived.

Long-distance dependencies and apparent motion

It was just noted that in (2), the number of the phrase "The man" is constrained by "eats", which is grammatically distant from it. Characterizing and inferring these long-distance dependencies has been a difficult problem for linguistic theorists and designers of sentence parsers for most of modern linguistic history. This contrasts with "short-distance" dependencies which are much simpler. Notice that the subject of "eats" is much closer to and more obvious in (4) than it is in (2).

(2) The man John said [_ eats [steak]] was wearing a hat.
(4) The steak John said [the man eats _] was tender.

A rough way of characterizing the difference between the two sentences is that the immediate proximity of "eats" to its subject makes their relationship much more obvious.

In the case of physical inference, the phenomenon of object permanence is an example of proximity (in space and time) making an identity relationship much more obvious than the identity of two objects perceived over a long-distance. When people see an object at a particular place and time and then in a fraction of a second see an object with a similar appearance, the two object sightings are perceived as the "apparent motion" of a single object. When the distance between the two objects in space and time is much larger, the identity is no longer obvious. For example, when a red Toyota drives into a crowded parking deck and a red Toyota emerges an hour later, the two car sightings might or might not be of the same car. The identity is not so obvious. Thus, long-distance dependencies and the problems they pose are a common feature of linguistic as well as physical events.

Infant physical reasoning mechanisms are sufficient to infer grammatical structure.

The dualities between physical and grammatical structure suggest that mechanisms for inferring the physical structure of the world might be useful for inferring the grammatical structure of an utterance. This section presents a model of syntactic understanding that is based on Cassimatis' (2002) model of infant physical reasoning. The model accounts for a wide variety of syntactic phenomena, including gapped constructions, long-distance dependencies and binding principles, using only the mechanisms included in the physical reasoning model.

Since the central argument of this paper is that human physical reasoning mechanisms, whatever they are ultimately found to be, are sufficient to parse syntactic structure, this paper has only discussed how to formulate parsing problems as physical reasoning problems and does not discuss in any detail the mechanisms of the physical reasoning model.

Polyscheme contains several modules, called *specialists*, for representing aspects of the physical world. Grammatical knowledge was added to Polyscheme using the representations of these specialists. For example, the category hierarchy (strictly speaking, a multi-tree) of Polyscheme's category specialist was used to represent lexical and phrasal category relationships; its temporal specialist was used to represent word order; its meronomic specialist was used to represent phrasal constituency and Polyscheme's physical constraint specialist was used to represent phrase structure constraints. No modifications of Polyscheme representations were needed to represent grammatical knowledge.

Physical problems are presented to Polyscheme using the formal language outlined in the last section. The structural dualities described in that section enable sentences to be presented to Polyscheme so that it can use its physical reasoning mechanisms to parse them. For example, the sentence, "The dog John bought bit him", is represented by a series of propositions, `Category(w1, TheUtterance) Occurs(w1, t1), Category(w2, DogUtterance) Occurs(w2, t2), Meets(t1,t2),` etc.

Upon receiving sentences in this format Polyscheme (using its mechanisms for resolving uncertainties) infers that the event `w2` is a `NounUtteranceEvent` and that it should be preceded by an event, `dogDeterminer`, that is a `DeterminerUtterance` event. Polyscheme's inference that "the" is the determiner of "dog" is represented by the proposition: `Same(w1,dogDeterminer)`, which indicates that the determiner event implied by "dog" is "the". Polyscheme's parse of an utterance is represented by a set of such propositions. They represent the identity of heads, arguments and adjuncts implied by words and phrases in the utterance (e.g., `dogDeterminer`) to the actual spoken words or phrases perceived (e.g., "the"). These propositions constitute a complete description of the syntactic structure of a sentence. Figure 1 illustrates the parse of the sentence, "The dog John bought bit him".

The crucial point is that once a sentence is represented in Polyscheme's input format, *only the mechanisms needed for physical inference are necessary to infer the grammatical structure of the sentence.*

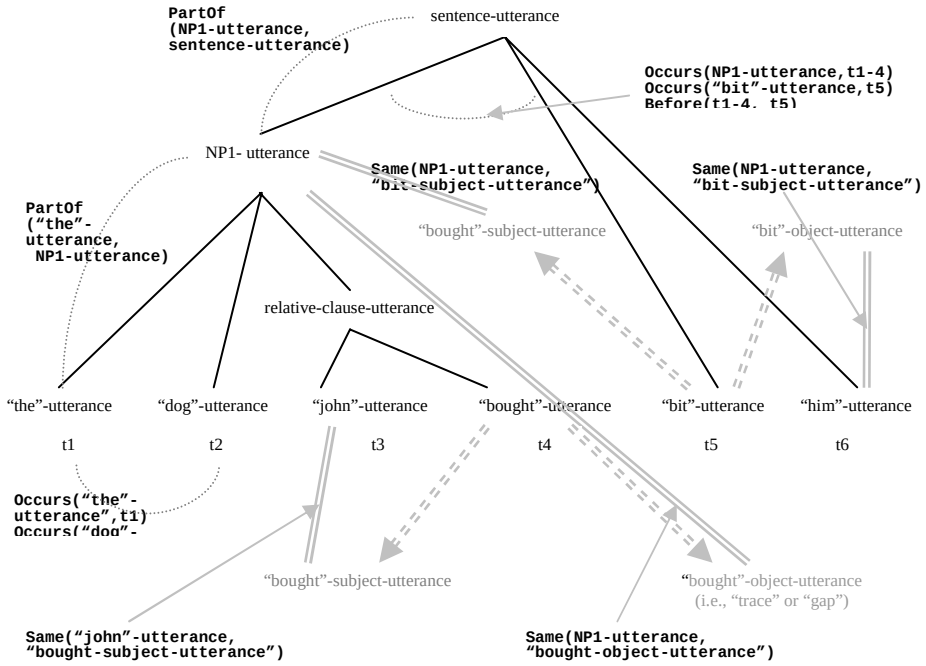


Figure 1. The syntactic structure of a sentence represented using concepts from infant physical reasoning.

Domain-general cognitive processes can enforce grammatical constraints.

The existence of several language universal constraints on syntactic structure is perhaps the most apparently unique feature of syntactic theory. Since they appear so peculiar to language, these constraints lend creditability to the thesis of linguistic modularity. This section argues that a supposed language-specific constraint, the c-command condition on binding, can be represented using the same cognitive structures used to represent physical events and that the cognitive processes used in the Polyscheme physical reasoning model from the last section explain how parsing obeys such constraints.

Radford (1997) formulates the c-command condition on binding thus: *A bound constituent must be c-commanded by an appropriate antecedent*. He defines c-command by stating that *A node X c-commands Y if the mother of X dominates Y, X ≠ Y and neither dominates the other*.

C-command is a constituency relationship and can therefore be reformulated using the notation of section 2:

> *X* ***c-commands*** *Y if* `PartOf(X,Z)` *and there is no* `X'` *such that* `PartOf(X,X')` *and* `PartOf(X',Z)` (i.e.,"Z is the mother of X")*;* `PartOf(Y,Z)` ("Z dominates Y")*;* `Same(X,Y)` *is false* ("X≠Y")*; and* `PartOf(X,Y)` *and* `PartOf(Y,X)` *are both false* ("neither dominates the other").

Having thus reformulated c-command as a meronomic relationship, it is possible explain how a cognitive process called *part inhibition*, which Polyscheme uses for physical reasoning, forces Polyscheme to observe the c-command condition on binding when parsing sentences.

In most physical interactions, a moving object "stays together". The parts of the object move together with the rest of the whole object. Spelke (1990) has termed this the "Cohesion Principle". This principle implies that people tracking the motion of an object composed of many smaller objects need only track the compound object. Its component objects will be wherever the whole object is. This makes, for example, the task of tracking one object composed of seven smaller objects generally much less than seven times more difficult than tracking one simple object. Thus, the Cohesion Principle supports the practice of paying more attention to a whole objects than to its parts. Markman (1989) found evidence that children do this when learning words. In Polyscheme, this attention preference can be implemented with "part inhibition":

> *When entities* $e_1, \ldots, e_n$*, are learned to be part of a larger entity*, E*, inhibit each of the* e_i.

When active during syntactic parsing in Polyscheme, part inhibition suppresses the activation of phrases that are forbidden antecedents under the c-command condition of binding. This is illustrated for the sentence, "The doctor Mary met at Bill's house likes herself". Notice that by the time processing reaches "herself", the model will have inferred that several noun phrases are constituents (directly or indirectly) of other noun phrases. In particular:

- `PartOf("house", "Bill's house").`
- `PartOf("Bill's house", "The doctor Mary met at Bill's house").`
- `PartOf("Mary", "The doctor Mary met at Bill's house").`

Part inhibition will therefore inhibit (and hence make them less likely binding targets) "house", "Bill's house" and "Mary" because they are each part of at least one larger utterance event. The following rewrite of the sentence shows these inhibited noun phrases in light gray:

[The doctor [Mary] met at [[Bill]'s house]] likes herself.

This example demonstrates that a single cognitive process (meronomic inhibition) can help syntactic inference conform to a grammatical constraint (on anaphoric binding) and on a physical constraint (on object motion). Although this is only one of many language-universal syntactic constraints, it raises the possibility that other constraints can be so treated.

Conclusions and future directions

Considerable work remains to establish that the mechanisms underlying physical reasoning also support syntactic parsing and that the model this paper presents is on the right track. The model must be extended to account for more languages and more grammatical phenomena, especially accounts of other universal syntactic constraints. The influence of mechanisms such as part inhibition on the observance of syntactic constraints must also be empirically confirmed. To the extent that the proposed dualities between cognitive structures and processes involved in inferring the structure of physical events and the syntactic structure of sentences are real, several important consequences follow.

First, the ability of a cognitive process such as part inhibition on help inference conform both to syntactic binding conditions and to the cohesion constraint on object motion is relevant to arguments for the existence of innate linguistic knowledge. These arguments (e.g., Chomsky, 1975) assert that children's early linguistic experience is too poor for them to learn these grammatical constraints and conclude that the constraints must therefore be part of some innate linguistic knowledge. This paper raises the possibility that these constraints are the linguistic manifestations of cognitive processes involved in cognition generally. These processes themselves may be innate or children may develop them as they learn to interact with their physical environment. In either case, since children's physical experience is so much richer than their early linguistic experience, this and many issues surrounding a putatively innate language faculty cannot therefore be resolved through *a priori* learnability arguments alone.

Finally, this work raises two methodological opportunities. First, since many other arguments in developmental psychology are also of the form, "behavior B implies knowledge or mechanism X", cognitive models which display B without X can potentially falsify those claims. Second, if the same mechanisms underlie physical reasoning and syntactic parsing, then corresponding to each language universal syntactic constraint should be a cognitive mechanism that supports the observance of this constraint in the same way that supports binding constraints. This suggests the potential for the theoretical posits of syntactic theory to be used as clues for the discovery of cognitive processes and *visa versa*.

References

Allen, J. F. (1983). *Maintaining knowledge about temporal intervals*. Communications of the ACM, 26:832--843.

Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychology, 43(1), 1-22.

Cassimatis, N. (2002). Polyscheme: A Cognitive Architecture for Integrating Multiple Representation and Inference Schemes. Ph.D. Dissertation. MIT Media Laboratory.

Chomsky, N. (1975). Reflections on language. New York: Pantheon.

Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, England.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Kestenbaum, R., Termine, N., & Spelke, E. S., (1987). *Perception of objects and object boundaries by 3-month-old infants*. British Journal of Developmental Psychology, 5, 367-383.

Langacker, Ronald W. (1999). *Grammar and Conceptualization.* Berlin: Mouton de Gruyter, 1999.

Markman, Ellen M. (1989). Categorization and Naming in Children: Problems of Induction. Cambridge, MA: MIT Press, 1989.

Pollard, C. & Sag., I. (1994). *Head-driven Phrase Structure Grammar*. CSLI Publications, Stanford, 1994.

Radford, Andrew (1997). *Syntactic theory and the structure of English: A minimalist approach*. Cambridge: Cambridge University Press.

Spelke, E. S. (1990). Principles of Object Perception. *Cognitive Science* 14: 29-56.

Spelke, E. S., & Tsivkin, S. (2001). Language and number: A bilingual training study. *Cognition* , 78, 45–88.

Talmy, L. (1985). *Force Dynamics in Language and Thought*. Papers from the Parasession on Causatives and Agentivity.

Xu, F., Carey, S. & Welch, J. (1999) Infants ability to use object kind information for object individuation. *Cognition*, 70, 137-166.

Reactive Agents Learn to Add Epistemic Structures to the World

Sanjay Chandrasekharan (schandra@sce.carleton.ca)
Terry Stewart (tcstewar@connect.carleton.ca)
Institute of Cognitive Science, Carleton University,
Ottawa, Canada, K1S 5B6

Abstract

We provide a computationally tractable model of how organisms can learn to add structures to the world to reduce cognitive complexity. This model is then implemented using two techniques: first using a genetic algorithm, and then using the Q-learning algorithm. The results clearly show that organisms with only reactive behavior can learn to systematically add structures to the world to reduce their cognitive load. We show that such learning can happen in both evolutionary time and within an agent's lifetime. An extension of this model (currently being implemented) is then illustrated, where organisms with just reactive behavior learn to systematically generate and use internal structures akin to representations.

Many organisms generate stable structures in the world to reduce cognitive complexity (minimize search or inference), for themselves, for others, or both. Wood mice (*Apodemus sylvaticus*) distribute small objects, such as leaves or twigs, as points of reference while foraging. They do this even under laboratory conditions, using plastic discs. Such 'way-marking' diminishes the likelihood of losing interesting locations during foraging (Stopka & MacDonald, 2003). Red foxes (*Vulpes vulpes*) use urine to mark food caches they have emptied. This marking acts as a memory aid and helps them avoid unnecessary search (Henry, 1977, reported in Stopka & MacDonald, 2003). The male bower bird builds colorful bowers (nest-like structures), which are used by females to make mating decisions (Zahavi & Zahavi, 1997). Ants drop pheromones to trace a path to a food source. Many mammals mark their territories.

At the most basic level, cells in the immune system use antibodies that bind to attacking microbes, thereby 'marking' them. Macrophages use this 'marking' to identify and destroy invading microbes. Bacterial colonies use a strategy called 'quorum sensing' to know that they have reached critical mass (to attack, to emit light, etc.). This strategy involves individual bacteria secreting molecules known as auto-inducers into the environment. The auto-inducers accumulate in the environment, and when it reaches a threshold, the colony moves into action (Silberman, 2003).

Such 'doping' of the world is commonly seen in lower animals. Most large animals (large body & brain size) do not exploit this strategy. Humans, however, do so to a tremendous degree. Markers, color-codes, page numbers, credit-ratings, badges, shelf-talkers, speed bugs, road signs, post-it notes, the list of epistemic structures used by humans is almost endless. Humans also add structures to the world to reduce cognitive complexity for artifacts. Examples include bar codes (makes check-out machines' decisions easier), content-based tags in web pages (makes Web agents' decisions easier), sensors on roads (helps the traffic light program's decision-making), etc.

The pervasiveness of such structures across species indicates that adding structure to the world is a fundamental cognitive strategy (Kirsh, 1996). Note that these structures predominantly serve a task-smoothening function – they make tasks easier for agents. Some of these structures have referential properties, but they do not exist for the purpose of reference. From here onwards, we will term such stable structures that provide "cognitive congeniality" (Kirsh, 1996), *epistemic structures*. The term is derived from a distinction between epistemic and pragmatic action made by Kirsh (1994).

How do organisms generate and use such structures? Can this generation of structures be captured computationally? These are the questions we address in this paper.

A Taxonomy and a Property

Most of the literature on epistemic structures is by David Kirsh, and from the field of Distributed Cognition in general. Kirsh's work explores the structural and computational properties of such structures, and how they function. We are interested in the other half of the problem, i.e., how such structures are generated and used. We use Kirsh's model to develop a situated cognition model of how such structures are generated. We then outline two simulations we implemented to test this model. An extension of this model (currently in progress) is then described.

Epistemic structures can be classified into three types, based on whom they are generated for. (examples of each in brackets).

1. Structures generated for oneself (Cache marking, bookmarks)
2. Structures generated for oneself and others (Pheromones, color codes)
3. Structures generated exclusively for others (Warning smells, badges)

A central feature of such structures is their task-specificity (more broadly, function/goal-orientedness). To illustrate this concept, consider the following example. Think of a

major soccer match in a large city, and thousands of fans arriving in the city to watch. The organizers put up large soccer balls on the streets and junctions leading up to the venue. Fans would then simply follow the balls to the game venue. Obviously, the ball reduces the fans' cognitive load, but how? To see how, we have to examine the condition where big soccer balls don't exist to guide the fans.

Imagine a soccer fan walking from his hotel to the game venue. She makes iterated queries to the world to find out her world state (What street is this? Which direction am I going?), and then does some internal processing on the information gained through the queries. After every few set of iterated queries and internal processing, she updates her world state and mental state, and this continues until she reaches her destination.

What changes when the ball is put up? The existence of the big soccer ball cuts out the iterated queries and internal processing. These are replaced by a single query for the ball, and its confirmation. The agent just queries for the ball, and once a confirmation of its presence comes in, she updates her world state and internal state. The ball allows the agent to perform in a reactive, or almost-reactive mode, i.e., move from perception to action directly. The key advantage is that almost no (or significantly less) inference or search is required.

This happens because the ball is a task-specific structure; it exists to direct soccer fans to the game venue. Other structures, like street names and landmarks in a city, are function-neutral or task-neutral structures. The fans have to access these task-neutral structures and synthesize them to get the task-specific output they want. Once the huge ball, a task-specific structure, exists in the world, they can use this structure directly, and cut out all the synthesizing. How the soccer fans manage to discover the ball's task-specificity is a separate and relevant issue, but we will not address it here. Task-specificity is a property of all epistemic structures found in nature, including pheromones and markers.

Kirsh's model of "changing the world instead of oneself" (Kirsh, 1996), postulates that such generation of structures involve task-external actions, and these structures work by deforming the state space, so that paths in a task environment are shortened. Such structures also allow new paths to be formed in the task environment. Kirsh's model tackles only physical structures generated by organisms, like tools. He does not consider structures generated for cognitive congeniality.

The Tiredness Model

How are task-specific structures that lower cognitive complexity generated? In this paper we consider the case of non-human organisms like ants, wood mice and red foxes. We will make two reasonable assumptions here. One, organisms sometimes generate random structures in the environment (pheromones, urine, leaf piles) as part of their everyday activity. Two, organisms can track their physical or cognitive effort (i.e., they get 'tired'), and they have a built-in tendency to reduce tiredness.

Now, some of the randomly generated structures are encountered while executing tasks like foraging and cache retrieval. In some random cases, these structures make the task easier for the organisms (following pheromones reduces travel time, avoiding urine makes cache retrieval faster, avoiding leaf-piles reduce foraging effort). In other words, they shorten paths in the task environment. Given the postulated bias to avoid tiredness, these paths get preference, and they are reinforced. Since more structure generation leads to more of these paths, structure generation behavior is also reinforced.

This theoretical framework gives us the basis for building artificial agents who also display the ability to learn to systematically generate useful structures in their environment.

The Simulation

To test and investigate the above model of epistemic structure generation, we have developed a computational model, where simple agents in a simple world, given feedback only in terms of their 'tiredness' (i.e., the effort required to perform their task), learn to systematically add structures to their environment.

The task we have chosen is analogous to foraging behavior, i.e., navigating from a home location to a target location and back again. Our environment consists of a 30x30 toroidal grid-world, with one 3x3 square patch representing the agent's home, and another representing the target. This 'target' can be thought of as a food source, to fit with our analogy to foraging behavior.

Agent Actions

At any given time, an agent can do one of five possible actions. The first and most basic of these is 'moving randomly'. This consists of going straight forward, or turning to the left or right by 45 degrees and then going forward. The agent does not pick which of these three possibilities occurs (there is a 1/3 chance of each).

In deciding the actions available to the agent, we needed to postulate some basic facilities within each agent. In our case, we felt it was reasonable to assume that the agents could distinguish between their home and their target. To do this, we added two more actions to the agents' repertoire. These are exactly like the first action, but instead of moving randomly, the agent would move towards whichever square is sensed to be the most 'home-like' (or the most 'target-like'). Initially, the only things in the environment that are 'home-like' or 'target-like' are the home and the target themselves.

One way to think about these actions is to consider the pheromone-following ability of ants. Common models of ant foraging (e.g. Bonabeau et al, 1999) consist of the automatic release of two pheromones: a 'home' pheromone and a 'food' pheromone. The ants go towards the 'home' pheromone when they are searching for their home, and they go towards the 'food' pheromone when foraging for food. This exactly matches these two actions in our agents.

The 'home' pheromone would be an example of a 'home-like' structure in the ant environment.

The fourth and fifth possible actions provide for the ability to generate these 'home-like' and 'target-like' structures. In the standard ant models, this could be thought of as the releasing of pheromones. However, our simulation has an important and very key distinction. Here, this ability to modify the environment is something the agents can do *instead* of moving around. That is, this generation process requires time and effort. The best way to envisage this is to think of an action that a creature might do which inadvertently modifies its environment in some way. Examples include standing in one spot and perspiring, or urinating, or rubbing up against a tree. These are all actions which modify the environment in ways that might have some future effect, but do not provide any sort of immediate reward for the agent. Kirsh (1996) terms these 'task-external actions'.

It must be stressed here that we are not presuming any sort of long-term planning on the part of the agents. We are simply specifying a collection of actions available to them, and they will choose these actions in a purely reactive manner (i.e., based entirely on their current sensory state). It may also be noted that our 'actions' are considered at a slightly higher level than is common in agent models. Our agents are not reacting by 'turning left' or 'going forward'; they are reacting by 'following target-like things' or 'moving randomly'. Furthermore, they do not initially have any sort of association between the action of making 'home-like' structures and the action of moving towards 'home-like' things. Any such association must be learned (either via evolution, or via some other learning rule).

Also, our agents are not designed to form structures automatically as they wander around (as is the case in standard ant models). In our simulation, a creature must expend extra effort to systematically generate these structures in the world. An agent that does this will be efficient only if the effort spent in generating these structures is more than compensated for by the effort saved in having them. Moreover, these are not permanent structures. The agents' world is dynamic and the structures do not persist forever. The 'home-likeness' or 'target-likeness' of the grid squares decrease exponentially over time. Furthermore, these structures also spread out over time. A 'home-like' square will make its neighboring squares slightly more 'home-like'. This can be considered similar to ant pheromones dispersing and evaporating, or leaf/twig piles being knocked over and blown around by wind or other passing creatures.

Agent Sensing

Since our agents are reactive creatures and thus do no long-term planning, they require a reasonably rich set of sensors. We have given them four sensors, two external and two internal, to detect their current situation. The two external sensors sense how 'home-like' and how 'target-like' the current location is (digitized to 4 different levels). The internal sensors are two simple bits of memory. One indicates whether the agent has been to the target yet, and the other indicates how long it has been since the agent generated a structure in its environment (up to a maximum of 5 time units). This is all that the agents can use to determine which action to perform.

This configuration gives each agent 192 (4 x 4 x 6 x 2) possible different sensory states.

The Learning Rules

For a purely reactive agent, we need some way of determining which action the agent will perform in each of these 192 states. We investigated two different methods for matching sensory states to actions: a Genetic Algorithm, and Q-Learning.

Stage 1: The Genetic Algorithm

For our first model, we used a genetic algorithm to determine which action to take in each situation. The genome consisted of a simple list of actions, one to perform in each state. To evaluate a particular genome, we started 10 agents in the home location and ran the simulation for 1000 time steps. The evolutionary fitness was the agents' average tiredness (i.e., how long it took each agent to make it back home from the target).

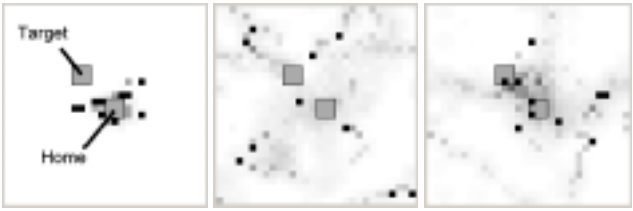


Figure 1: The computer model at 10, 100, and 300 time steps. Black dots are the agents. The shading is darker the more 'home-like' or 'target-like' a particular square is. This run shows typical agent behaviour after 300 generations.

Result: Initially, the agents behaved randomly. Starting at the 'home', they would wander about and might, by chance, find the target and then, if they were very lucky, their home. Indeed, most agents did not find the target and make it back within the 1000 time steps. On average, we found that each agent was completing 0.07 foraging trips every 100 time steps. After a few hundred generations, the agents were soon completing an average of 1.9 trips in that same period of time. In other words, the agents were able to, on an evolutionary time scale, learn to make use of their ability to sense and generate structures in the world. Furthermore, this ability provided a very large advantage over completely random behaviour.

This result confirmed that it is possible for agents to learn to systematically generate and use structures in the world in an evolutionary time scale. It also showed that we had not

chosen an impossible task for the agents to learn. However, for our purposes, we were much more interested in an individual agent learning to generate epistemic structures within that agent's lifetime. To investigate this, we turned to the Q-Learning algorithm.

Stage 2: Q-Learning

The heart of our investigation was to determine whether a simple, general learning algorithm would allow our agents to discover and make use of the strategy of systematically adding structures to the world. In keeping with our 'tiredness' theory, the only feedback the learning mechanism had was an indication of the exertion or effort. The delayed-reinforcement learning rule known as Q-Learning (Watkins, 1989) seemed best suited for this task. (Other similar algorithms will be investigated in future work). The Q-Learning algorithm[1] develops an estimate of the eventual outcome of performing a given action in a given situation. The agent then performs the action with the highest expected payoff.

Using the Q-Learning algorithm, we again ran 10 agents for 1000 time steps. To indicate 'tiredness', we gave them a reinforcement value of -1 all the time (indicating a constant 'punishment' for expending any effort). When they returned home after finding the target, they were given a reinforcement of 0, and they were then sent back out again for another trip. Each agent independently used the Q-Learning algorithm, and there was no communication between the agents.

Result: The dark line in figure 2 shows the results averaged over 100 separate trials. We can clearly see that the agents are improving over time (i.e., they are spending less time to perform their foraging task).

Stage 3: Confirmation

Although we have observed improvement over time, we still need to show that it is the agents' ability to systematically add structures to the world that is causing this effect. To prove this, we re-ran the experiment, this time removing the agents' ability to generate structures in the world. No other changes were made.

Result: We found that when the agents were unable to generate structures in the world, Q-Learning did not provide as much improvement[2]. This result is shown in the lighter line in Figure 2. There is still a small improvement given by Q-Learning, but we are able to conclude that the significant improvement seen in the previous experiment is due to the agents' ability to modify their environment.

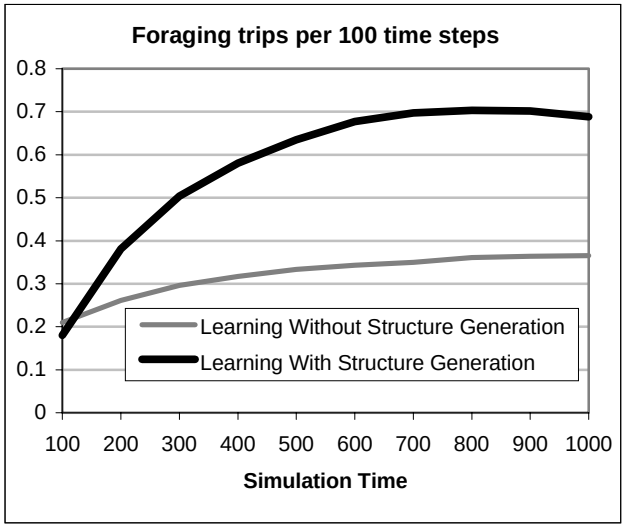


Figure 2: The effect of epistemic structure generation. The foraging rate is measured in trips per 100 time steps. A foraging rate of 0.5 means that trips require an average of 200 time steps to complete.

We can also see from Figure 2 that having these extra actions available does incur some cost in the early stages. Initially, the agents perform slightly worse. However, the advantage of being able to form epistemic structures quickly improves the agents' performance. By the end of the simulation, agents require only around 150 time steps to make a complete trip (a foraging rate of 0.66 trips in 100 time steps). This is twice as quick as agents without the structure-forming ability.

Table 1: Time spent performing various actions.

Action	With Structure Generation	Without Structure Generation
Move randomly	10%	32%
Toward 'home-like'	19%	36%
Toward 'target-like'	13%	32%
Make 'home-like'	35%	
Make 'target-like'	23%	

When we analyzed the actions of the agents, we found that they actually spent 58% of their time generating structures. This is striking, since time spent generating these structures means less time for wandering around trying to find the target or their home. Table 1 gives the breakdown of how time was allocated to different actions. The data indicates that epistemic structure generation allowed the agents to go from spending 300 time steps down to 150 time steps to complete their foraging task, even

[1] The estimated reward for performing action a in state s is $Q(s,a)$. This is increased by $\alpha(r+\gamma\max(Q(s',b))-Q(s,a))$, where r is the immediate reward/punishment, s' is the resulting state, γ is the future discounting rate (set to 0.5), and α in the learning rate (0.2). We used an ε-choice rule with ε set to 0.1, so the agents choose the action with the highest expected reward 90% of the time, and the other 10% they perform an action at random.

[2] Q-Learning also did not provide significant improvement if the agents were only able to generate one type of structure, or if any of the agent's sensors were removed.

though over half of those 150 time steps are spent standing still. There is clearly a large efficiency advantage to making use of these structures.

There are many Reinforcement Learning algorithms available other than Q-Learning, and any one of them could be used in this sort of model. As we investigate other, more complex situations, we will try using these alternatives to Q-Learning, such as actor-critic methods. All of these models learn in a similar way, but with rather different details, and so the resulting high-level behaviour may be different.

Conclusions

The Q-Learning system is a concrete implementation of our model: a simple learning mechanism that allows agents with purely reactive behavior to systematically add structures to the world to lower search.

The 'tiredness'-based learning model implemented in this simulation can explain the generation of task-specific structure in cases 1 and 2 (structures for oneself and structures for oneself & others). Case 2 (structures generated for oneself & others) is explained by appealing to the similarity of systems – if a structure provides congeniality for me, it will provide congeniality for other systems like me. In our computer model, the agents ended up forming structures that were useful for everyone, even though they were just concerned about reducing their own tiredness. This was possible only because the agents were similar to each other. This is similar to how paths are formed in fields: one person cuts across the field to reduce his physical effort, others, sharing the same system and wanting to reduce their effort, find the route optimal. As more people follow the route, a stable path is formed.

For case 3, (structures generated exclusively for others), the 'tiredness' model explains only some cases. For instance, it could explain the generation of warning smells and colors exclusively for others, because the effect of such structures could be formulated in terms of tiredness (the release of some chemical ends up cautioning predators, which reduces the number of fleeing responses the organism makes, thus reducing tiredness, which, when fed back, reinforces the initial action). However, this model, as it stands, cannot explain the generation of structures like the bower or the peacock's tail, which do not seem to provide any tiredness benefit for the generator.

Other Models

It is worth noting that our model presents a novel simulation of ant behaviour. The closest existing models are those in (Bonabeau et al, 1999) which use the 'home-pheromone' and the 'food-pheromone'. This is in contrast to such models as (Nakamura & Kurumatani, 1996), where a land-based and an airborne pheromone are used, or any models of the Cataglyphis species of ant, which uses a complex landmark-navigation scheme which allows it to return directly to the nest (Miller & Wehner, 1988).

That said, all of these other models assume both that pheromones are continually being released while the ant forages, and that there is no learning happening during the foraging behaviour. Our Q-Learning model does not make either of these assumptions.

We were unable to find references indicating that real ants might, in fact, learn to use pheromones, or any research that indicates that the effort required to produce these pheromones might interfere with foraging behaviour. So our model may not be a good one for understanding ants. However, the fact that our agents are able to learn to reflexively generate these cognitively beneficial structures in the absence of any immediate feedback to their benefit, indicates a simpler way to model more complex creatures that exhibit such behaviour.

Future Work

Our current simulation implements a learning process based on the feedback of tiredness. It leads to organisms generating task-specific *external* structures in the world. These are structures that lower cognitive load, accessed by organisms at run-time, while they execute tasks.

Interestingly, the same model can explain generation and tracking of *internal* structures in organisms. The actions which generated structure in our simulation were actions that affected the environment. But this does not have to be the case. Just as we had both internal and external sensors, we can have actions which affect either the state of the world *or* the state of the agent itself. In other words, we can use this model to investigate the generation of *internal* structure (i.e., representations).

As an example, consider foraging bees. Suppose that, just as our agents left traces in the world of their activity via their structure-generating actions, we have the bees leave a sequence of internal memory traces corresponding to landmarks (say a tall tree, a lake, a garden) as a result of their everyday foraging activity. In some foraging trips of some bees, the trace sequences match to some degree the external structures they perceive. Such trips involve less search, because they lead to food more directly, i.e., they form shorter paths in the task environment. Over time, using the exact same learning mechanisms that apply in the external case, the bias against tiredness leads to such paths being used more, and so they are reinforced. This leads to landmark-based navigation, which, in fact, exists in bees (Gould, 1990). As in the case of external structures, the generation of such memory traces is reinforced because more traces lead to more such shorter paths in the task environment. We are currently working on a computational model of this example. Interestingly, recent research shows homing pigeons using human-generated environment structure in a similar fashion to reduce cognitive load. They follow highways and railways systematically to reach their destination (Guilford, 2004).

The above framework presents a situated cognition model of how memory structures come to be used as task-specific structures, and why such internal structures are

systematically generated. If such task-specific memory structures are considered to be representations (that is, they stand for something specific in the world), then the model explains, in a computationally tractable manner, how organisms with just reactive behavior can learn to generate and use representations.

The model also explains what such 'primitive' representations are: they are the internal traces of the world that allow the agent to shorten paths in a task environment. Roughly, they are computation-reducing structures (and equivalently, energy-saving structures). They are internal 'stepping stones' that allow organisms to efficiently negotiate the ocean of stimuli they encounter. This means the traditional cognitive science view, that thinking is computations happening over representations, presents a secondary process – it describes a privileged path in the task environment. In the stepping stone view, representations are crucial for organisms, but they are just useful, incidental entities, not fundamental entities by themselves. We are exploring the philosophical implications of this view.

All source code for the simulations can be found at:
http://www.carleton.ca/iis/TechReports/code/2004-01/

Acknowledgment

Some of the ideas presented in this paper were developed while the first author worked as a pre-doctoral fellow with the Adaptive Behavior and Cognition (ABC) Group of the Max Planck Institute for Human Development, Berlin. He acknowledges the group's support, particularly the strong encouragement and critical feedback from Dr. Peter Todd and Dr. John Hutchinson.

References

Alcock, J. (1998*). Animal Behavior: An evolutionary approach*, Sunderland, Mass., Sinauer Associates.

Bogen, J. (1995). Teleological explanation. In Honderich (Ed.), *The Oxford Companion to Philosophy*. New York: Oxford University Press.

Bonabeau E., Dorigo M. and Theraulaz G. (1999) *Swarm intelligence: From natural to artificial systems.* Santa Fe Institute studies in the sciences of complexity. New York: Oxford University Press.

Clark, A. (1997). *Being There: putting brain, body, and world together again*, Cambridge, Mass., MIT Press.

Gould, J.L. (1990) Honey bee cognition. *Cognition*, 37, 83-103.

Guilford, T., Roberts, S. & Biro, D. Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models. *Journal of Theoretical Biology*, published online, (2004).

Henry, J.D. (1977). The use of urine marking in the scavenging behaviour of the red fox (Vulpes vulpes). *Behaviour*, 62:82-105.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513-549.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, Vol 4, No. 3/4, 415-452.

Miller, M., & R. Wehner (1988). Path integration in desert ants, Cataglyphis fortis. *Proceedings of the National Academy of Sciences* USA 85: 5287-5290.

Nakamura, M., & Kurumatani, K. (1996). Formation mechanism of pheromone pattern and control of foraging behavior in an ant colony model. *Proceedings of the Fifth International Conference on Artificial Life*, 67-74.

Silberman, S. (2003), The Bacteria Whisperer. *Wired*, Issue 11.04, April 2003.

Stopka, P. & Macdonald, D. W. (2003) Way-marking behavior: an aid to spatial navigation in the wood mouse (Apodemus sylvaticus). *BMC Ecology*, published online, http://www.biomedcentral.com/1472-6785/3/3

Watkins, C. (1989). *Learning From Delayed Rewards*, Doctoral dissertation, Department of Psychology, University of Cambridge, Cambridge, UK.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A missing piece of Darwin's puzzle*. Oxford: Oxford University Press.

Context-Driven Construction Learning

Nancy Chang (nchang@icsi.berkeley.edu)
UC Berkeley, Department of Computer Science and
International Computer Science Institute
1947 Center St., Suite 600, Berkeley, CA 94704

Olya Gurevich (olya@socrates.berkeley.edu)
UC Berkeley, Department of Linguistics
1203 Dwinelle Hall, University of California at Berkeley
Berkeley, CA 94720-2650

Abstract

We present a computational model of how partial comprehension of utterances in context may drive the acquisition of children's earliest grammatical constructions. The model aims to satisfy convergent constraints from cognitive linguistics and crosslinguistic developmental evidence within a statistically driven computational framework. We examine how the tight coupling between contextually grounded language comprehension and learning processes can be exploited to improve the model's ability to search the space of possible constructions. In particular, previously learned constructions may not fully account for all contextually perceived mappings between forms and meanings. In the model, these incomplete analyses directly prompt the formation of new relational mappings that bridge the gap. We describe an experiiment applying the model to the acquisition of English verb island constructions and discuss how the model handles more complex examples involving Russian morphological constructions. Together these demonstrate the viability of the overall approach and representational potential of the model.

Beyond single words

How do children make the leap from single words to complex combinations? The simple act of putting one word in front of another to indicate some relation between their meanings is widely considered the defining characteristic of linguistic competence and the key to unlocking the combinatorial and expressive power of language. A viable account of the acquisition of these combinatorial patterns, or *grammatical constructions*, would thus have significant implications for any theory of language that aspires to cognitive plausibility.

As with most issues impinging on the nature of grammar, linguistic and developmental inquiries into the source of combinatorial constructions have bifurcated along theoretical lines. These reflect divergent assumptions about, among other things, what kind of learning bias children bring to the task, how the target linguistic knowledge should be represented, what kind of data should be considered part of the training input, and how (if at all) language learning interacts with other linguistic and cognitive processes. Theoreticians within the formalist "learnability" paradigm, for example, have generally restricted their attention to the form domain, taking the input for learning to be a set of surface strings (each a sequence of surface forms) and positing relatively abstract structures that govern the combination of linguistic units.

This paper takes as starting point the hypothesis that the learning problem at hand may encompass a broader subset of the child's experience, centrally including meaning as it is communicated in context. We assume along with many theories of language that the basic unit of linguistic knowledge, for both lexical items and larger phrasal and clausal units, is a symbolic pairing of form and meaning, or *construction* (Langacker, 1987; Goldberg, 1995; Fillmore and Kay, 1999). Since the target of learning is rooted in both form and meaning domains, the learner should exploit information from both domains during learning.

Most importantly, we view linguistic constructions as inherently dependent on and supportive of dynamic processes of language *use*, anchored in a communicative context. A crucial but often neglected source of bias in learning constructions must therefore be how much they help the child meet her communicative goals.

This paper presents a computational model of construction learning consistent with these principles, focusing on how language understanding drives language learning. We describe a statistically driven machine learning framework that takes as input a sequence of child-directed utterances paired with their associated situational context, along with the current grammar, or set of constructions; this grammar is initially restricted to lexical items. The utterances are passed to a language understanding system (Bryant, 2003) that produces a partial interpretation, which provides the basis for the learning model to form new constructions. We present results showing how the model acquires simple English "verb island" constructions (Tomasello, 1992), and discuss how the same mechanisms handle the more complex constructions involved in Russian nominal case marking. These studies lend support for the larger program of integrating cognitive and constructional approaches to linguistics, crosslinguistic developmental evidence, and machine learning techniques to address the puzzles of language acquisition.

The Construction Learning model

We briefly describe the construction learning model in terms of (1) the target representation of learning, (2) assumptions about the child language learning scenario, and (3) the computational learning framework; see (Chang, 2004; Chang and Maia, 2001) for more details.

Target representation: embodied constructions

Embodied Construction Grammar (Bergen and Chang, in press; Chang et al., 2002) is a computationally explicit formalism for capturing insights from the construction grammar and cognitive linguistics literature. ECG supports an approach to language understanding based on two linked

processes: **analysis** determines what constructions and schematic meanings are present in an utterance, resulting in a *semantic specification* (or *semspec*); the semspec serves to parameterize a **simulation** using active representations (or *embodied schemas*) to produce context-sensitive inferences.

Semantic representations in ECG are richly detailed and cognitively motivated, incorporating image schemas, motor schemas, force-dynamic schemas, and fine-grained representations of event and participant structure. But for ease of exposition, we omit most of this detail in our simple examples below, since it is not crucial for our current focus on the acquisition of the earliest constructions with constituent structure.[1]

We highlight a few aspects of the formalism relevant for the learning model discussion to follow, exemplified by the lexically specific clausal THROW-TRANSITIVE construction shown in Figure 1. The formalism draws from both object-oriented programming languages and constraint-based grammars, including notations for expressing features, inheritance, typing, and unification/coindexation.

construction THROW-TRANSITIVE
constituents
t1 : REFERRING-EXPRESSION
t2 : THROW
t3 : REFERRING-EXPRESSION
form
$t1_f$ `before` $t2_f$
$t2_f$ `before` $t3_f$
meaning
$t2_m$.thrower $\longleftrightarrow$ $t1_m$
$t2_m$.throwee $\longleftrightarrow$ $t3_m$

Figure 1: Representation of a lexically specific THROW-TRANSITIVE construction, licensing expressions like *You throw the ball*, with separate blocks listing constituent constructions (t1, t2, t3), form constraints (e.g., the word order relation `before`) and meaning constraints (e.g., the identification binding $\longleftrightarrow$).

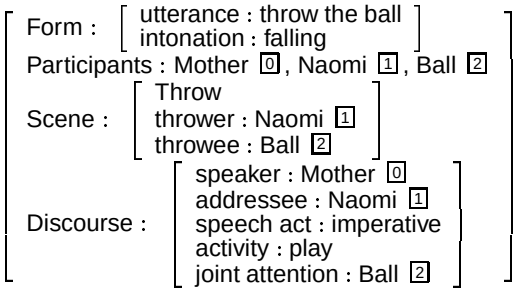
All constructions have **form** and **meaning** blocks, but the **constituents** block appears only in the complex constructions that are the target of the present learning enterprise. These constituents may be typed as instances of particular constructions, and their form and meaning components (or *poles*) may be referred to (shown with a subscripted f or m) by the constraints listed in the form and meaning blocks. Form constraints are used to capture (partial) word order and other relations between form segments. In the meaning domain, the primary relation is *identification*, or unification, between two meaning entities. In particular, we will focus on role-filler bindings, in which a role (or feature) of one constituent is identified with another constituent. The example construction involves three constituents – two referring expressions and the verb THROW. Their form poles are constrained to come in a specified order, and the meaning poles of the two referring expressions fill the specified roles (thrower and throwee) of the verbal constituent's meaning pole.

[1]These features play a key role in the acquisition of argument structure and grammatical markers; we return to this issue later.

Input: modeling the child learning scenario

Children entering the two-word stage (typically toward the end of the second year) are relatively savvy event participants, having developed a wealth of structured knowledge about the participant roles involved in different events and the kinds of entities likely to fill them (Nelson, 1996; Tomasello, 1992). Their single-word vocabularies typically include names for familiar people and objects, as well as some words for actions. They make use of pragmatic knowledge and joint attention to infer both communicative intentions (Tomasello, 1995) and subtle lexical distinctions (Bloom, 2000), and often respond appropriately to multi-word comments and queries from their parents even in the single-word stage (Bloom, 1973). That is, children can robustly interpret utterances beyond their productive abilities, using (incomplete) lingustic knowledge and relatively sophisticated inference abilities.

These findings suggest that grammar learning may, rather than suffer from the poverty of the stimulus, instead capitalize on the opulence of the substrate. Our learning model thus assumes an ontology of known concepts and an initial lexicon of constructions, represented in ECG. Input data reflects the child's ability to perceive an utterance with a particular intonational contour and segment it into a sequence of word forms, and to pragmatically infer the relevant participants and events in the accompanying situation, as shown in the example input below, where boxed index numbers indicate identification links between participants:



The example input represents a discourse event in which the mother says "throw the ball" with falling intonation to the child (Naomi). We assume the child can infer (using pragmatic cues) that the corresponding main scene concerns a throwing event to be performed by the child on a particular ball attended to in context. Note here that the action is the inferred intent of the mother, and may or may not be carried out by the child. But the (intended) role-filler structure is assumed in our model to be inferrable in context and thus available to the learning mechanism.

Besides these assumptions, the learning model also draws on findings about the developmental course of construction learning. Early word combinations appear to be lexically specific, with a gradual transition to more general constructions (Tomasello, 1992); crosslinguistically they tend to relate to a small set of basic scenes (Slobin, 1985); and acquisition phe-

nomena are sensitive to a number of usage-based considerations (Tomasello, 2003; Clark, 2003) such as the frequency with which a construction is encountered, the simplicity of its form and meaning, and how easily a particular utterance can be analyzed into its component constructions.

In sum, the model incorporates strong assumptions about the child's conceptual and lexical knowledge and pragmatic abilities, based on developmental evidence. Relatively weak assumptions are made about innate syntactic biases: the ECG formalism allows word order as a possible form constraint. Thus most of the learning bias comes from the meaning domain, and the constructional assumption that forms and meanings are linked.

Computational learning framework

We now describe a computational model of how constructions can be learned from experience. The input is a sequence of utterances paired with their meanings in context, as described in the last section. The learner has access to a language analysis process like that described earlier, which produces a (partial) interpretation of the input utterances based on the current (potentially incomplete) set of constructions. The learning task is then modeled as an incremental search through the space of possible grammars, where the learner adds new constructions on the basis of encountered data. As in the child learning situation, the goal of learning is to converge on an optimal set of constructions, i.e., a grammar that is both general enough to encompass significant novel data and specific enough to accurately predict previously seen data.

A suitable overarching computational framework for guiding the search is provided by the minimum description length (MDL) heuristic (Rissanen, 1978), which is used to find the optimal analysis of data in terms of (a) a compact representation of the data; and (b) a compact means of describing the original data in terms of the compressed representation. The MDL heuristic exploits a tradeoff between competing preferences for smaller grammars (encouraging generalization) and for simpler analyses of the data (encouraging the retention of specific/frequent constructions). This is an approximation of the same tradeoff exploited in previous work applying Bayesian model merging to learning verbs (Bailey, 1997) and context-free grammars (Stolcke, 1994). We extend these approaches to handle the relational structures of the ECG formalism and the process-based assumptions of the model.

Learning strategies. The model may acquire new constructional mappings in two ways:

relational mapping New relational map(s) are formed to account for form-meaning mappings present in the input but unexplained by the current grammar.

reorganization Regularities across known constructions are exploited, either to merge two similar constructions into a more general construction, or to compose two constructions that cooccur frequently into a single construction.

Each construction is also associated with a weight that is incremented as a result of its successful use in analysis.

Algorithms for these operations are given elsewhere (Chang and Maia, 2001; Chang, 2004); relational mapping plays the most crucial role in proposing new relational constraints among constituents and will be illustrated in more detail in the next section.

Evaluating grammar cost. The strategies above provide means for updating the current grammar; the model must then determine which update is optimal at any point in learning, according to some length-based evaluation criterion. We use an approximation of the Bayesian posterior probability of the grammar G given the data D that we call the *cost* of G:

$$\begin{aligned}
\text{cost}(G|D) &= m \cdot \text{size}(G) + n \cdot \text{cost}(D|G) \\
\text{size}(G) &= \sum_{c \in G} \text{size}(c) \\
\text{size}(c) &= n_c + r_c + \sum_{e \in c} \text{length}(e) \\
\text{cost}(D|G) &= \sum_{d \in D} \text{score}(d) \\
\text{score}(d) &= \sum_{x \in d} (\text{weight}_x + p \cdot \sum_{t \in x} |\text{type}_t|) \\
&\quad + \text{height}_d + \text{semfit}_d
\end{aligned}$$

where m and n are learning parameters that control the relative bias toward model simplicity and data compactness. The size(G) is the sum over the size of each construction c in the grammar (n_c is the number of constituents in c, r_c is the number of constraints in c, and each element reference e in c has a length, measured as slot chain length). The cost (complexity) of the data D given G is the sum of the analysis scores of each input token d using G. This score sums over the constructions x in the analysis of d, where weight_x reflects relative (in)frequency, $|\text{type}_t|$ (where t ranges over the constituents of x) denotes the number of ontology items of type t (i.e., the number of alternative fillers for the constituent), summed over all the constituents in the analysis and discounted by parameter p. The score also includes terms for the height of the derivation graph and the semantic fit provided by the analyzer as a measure of semantic coherence.

These criteria favor constructions that are simply described (relative to the available meaning representations and the current set of constructions), frequently useful in analysis, and specific to the data encountered.

Learning from meaning in context

This section describes in greater detail the integration of the learning model with an implemented construction analyzer (Bryant, 2003). We illustrate the analyzer-learner interaction with an example based on the input data shown earlier.

Constructional analysis. On encountering new data, the learner first calls a construction analyzer designed to perform the analysis process described earlier (Bryant, 2003).

The analyzer consists of a set of *construction recognizers* that recognize the input forms of each construction and check whether the relevant semantic constraints are satisfied. The analyzer draws on partial parsing techniques so that utterances not fully covered by known constructions can nevertheless yield partially filled in semantic specifications. Moreover, unknown forms in the input can be skipped, allowing quite simple constructions to provide at least skeletal interpretations of more complex utterances.

In the example, we assume the current grammar includes lexical constructions for *throw* and *ball*, but no word combinations or construction for the article *the*. The utterance "throw the ball" at this stage produces a semspec containing two schemas, corresponding to the meanings of the two recognized constructions, but no associations between them:

```
SCHEMA13 (Ball)
SCHEMA3 (Throw)
    thrower: SCHEMA4 (Human)
    throwee: SCHEMA8 (Physical-Object)
```

Here, SCHEMA13 corresponds to the meaning pole of the BALL construction, and SCHEMA3 corresponds to the meaning pole of the THROW construction.

Resolution. We extended the existing analyzer with a resolution procedure that matches the output semspec against the input context. Like other resolution (e.g. reference resolution) procedures, it relies on category/type constraints and (provisional) identification bindings. The resolution procedure attempts to unify each schema and constraint appearing in the semspec with some type-compatible entity or relation in the context. In the example, SCHEMA13 resolves by this process to the salient Ball in the input, and SCHEMA3 resolves to the Throw action in context.

Relational mapping. At this point the learner has a partial semspec that through resolution accounts for a subset of the information available in the input context description (namely, the presence of a throwing event and a ball). The learner now searches for a candidate relational mapping present in the input context but not accounted for by the semspec – that is, a form relation that is unused in the current analysis, paired with a meaning relation that is unaccounted for in the semspec. These relations must be structurally isomorphic, that is, their arguments must involve form and meaning poles of the same constituent constructions. In the example, the input includes a number of unexplained meaning relations – for example, the identity of the speaker and addressee, and both Throw schema roles. But only one of these – the binding between the throwee role and the ball – involves meanings that are also accounted for in the input, and for which there is a corresponding form relation over the form poles of the relevant constructions (i.e., an ordering relation between *throw* and *ball*).

The situation is depicted in Figure 2, where the input utterance-context pair are shown as form and meaning schemas and relations on either side of the figure. Constructions found by the analyzer are shown in the center, accounting for the form and meaning schemas drawn with solid heavy lines (i.e., the recognized input and produced semspec). The discovery of structurally isomorphic relations over the form and meaning poles of the two recognized constructions leads to the hypothesis of the new lexically specific THROW-BALL construction shown in the figure (with heavy dotted lines) and formally in Figure 3.

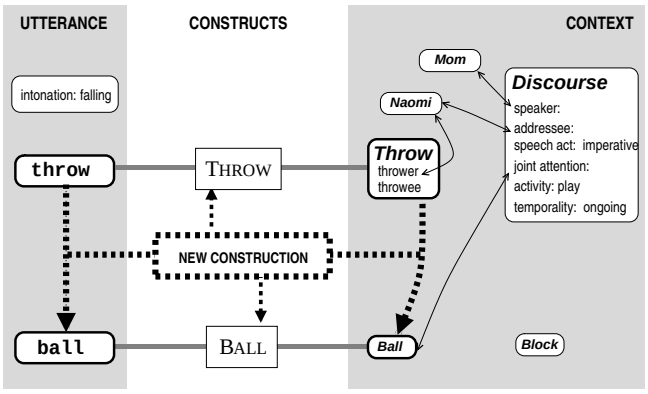


Figure 2: Relational mapping in the learning model for the utterance *throw (the) ball*. Heavy solid lines indicate structures matched during analysis; heavy dotted lines indicate the newly hypothesized mapping.

construction THROW-BALL
constituents
t1 : THROW
t2 : BALL
form
$t1_f$ before $t2_f$
meaning
$t1_m$.throwee $\longleftrightarrow$ $t2_m$

Figure 3: Example learned construction: THROW-BALL learned from the utterance-context pair in Figure 2.

This example illustrates the simplest relational mapping strategy; the requirement of strictly isomorphic form and meaning relations can also be relaxed to allow more complex relational correspondences (expressed using longer constraints). All such mapping strategies are designed to discover how known constructions may fit together in larger structures, thus giving rise to constituent structure.

Once these structured (but lexically specific) constructions are learned, they are subject to reorganization, such that multiple constructions involving *throw* and a specific thrown object may be merged into a generalized *throw*-Object construction (contingent on the MDL learning criteria). We now explore how the model can learn patterns of this kind from a corpus of child-directed utterances.

Experiment: English verb island constructions

The construction learning model was tested in an experiment targeting the acquisition of lexically specific, or item-

based, constructions; we focus on patterns centering on specific verbs. This task is of cognitive interest, since "verb island" constructions appear to be learned on independent trajectories (i.e., each verb forms its own "island" of organization (Tomasello, 1992; Tomasello, 2003)).

Input data. The training corpus for the experiment is a subset of the Sachs corpus of the CHILDES database of parent-child transcripts (Sachs, 1983; MacWhinney, 1991) annotated as part of a study of motion utterances (Dan I. Slobin, p.c.). The transcript data consists of parent and child utterances occurring during a joint background activity (e.g., a meal or play). All motion expressions were annotated with descriptions of the inferred speaker meaning and the surrounding discourse and situational context. We used a subset of this corpus containing 829 labeled motion-related child-directed utterances spanning the child's development from 1;3 through 2;6, during which the child makes the transition from the single-word stage. Parental utterances were extracted into input data of the form shown above.

Evaluation criteria. The goal of language learning in our framework is to improve language understanding. We thus defined a quantitative measure intended to gauge how new constructions incrementally improve the model's comprehensive capacity. We defined a grammar G's *coverage* of data D as the percentage of total bindings b in the data (i.e., role-filler bindings relevant to the verb) included in its interpretation (semspec), and measured coverage at each stage of learning. The *throw* subset, for example, contains 45 bindings to the roles of the Throw schema (thrower, throwee, and goal location). At the start of learning, the model has no combinatorial constructions and can account for none of these, but as learning progresses, the model should learn constructions that allow it to cover increasingly more of these bindings.

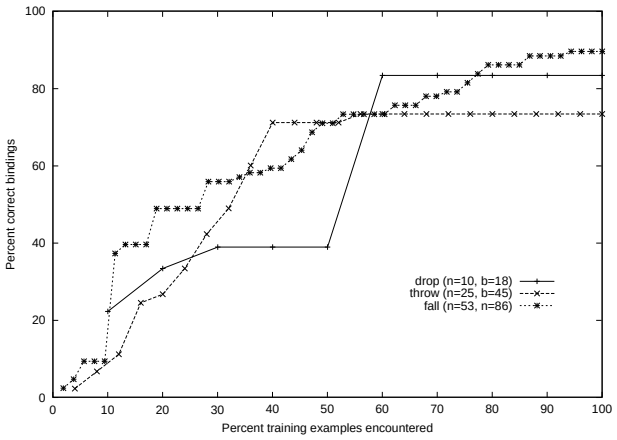


Figure 4: Incremental coverage for three verb islands. (Graphs are scaled relative to subcorpus size.)

Results. Figure 4 shows results for three verb islands: *drop* (n=10 examples), *throw* (n=25), and *fall* (n=50); other verbs followed similar patterns. In all cases coverage gradually improved over the course of learning, as expected, and the model was able to account for a majority of the bindings in the data relatively quickly. But as shown by these examples, the particular learning trajectories were distinct: *throw* constructions show a gradual build-up before plateauing; *fall* has a more fitful climb that seems to converge at an upper bound; and *drop* has an even more jagged rise. A possible explanation for some of these differences may lie in pragmatic differences: *throw* has a much higher percentage of imperative utterances than *fall* (since throwing is pragmatically more likely to be done on command). The relational mapping strategy used in the experiment misses the association of an imperative speech-act with unexpressed agent, which has a more pronounced effect on the learning of *throw* constructions.

Also as expected, the earliest constructions are combinations of specific words (e.g, *throw-ball*, *throw-frisbee*, *you-throw*), giving rise later in learning to more general constructions (e.g., *throw*-Object and Agent-*throw*). Figure 5 shows the number of each type learned.

	lexical	general	total
drop	5	1	6
throw	11	4	15
fall	21	9	30

Figure 5: Number of constructions learned for each verb, including both fully lexically specific constructions and verb island constructions with at least one generalized argument.

Discussion. Despite the small corpus sizes, the results are indicative of the model's ability to acquire useful verb-based constructions. Differences in verb learning lend support to the verb island hypothesis and illustrate how the particular semantic, pragmatic and statistical properties of different verbs can affect their learning course.

Case study: Russian

The verb island experiment demonstrates the model's ability to acquire constituent structure, an essential step in moving beyond lexical items. But the child's learning scenario may be significantly more complicated. We briefly consider some problems that arise for learners of comparable Russian constructions and how the model addresses them.

In Russian, casemarkers suffixed on nouns indicate the participant role played by their associated referents. Word order is thus highly variable: *malchik brosaet devochk-e myach* (boy-NOM throw-3s girl-DAT ball-ACC) and *devochk-e brosaet myach malchik* (girl-DAT throw-3s ball-ACC boy-NOM) have the same participant structure, glossed as 'boy throws ball to girl'. Moreover, the same marker may be ambiguous over multiple class/case combinations (e.g., *-a* indicates either Feminine-I/NOM or Masculine-Animate/ACC).

Flexible word order does not in itself pose an obstacle to the model. Deferring nominal morphology for the moment (see below), the first multi-word constructions learned by the

model (via relational mapping) are, like their English equivalents, both verb-specific and fixed-order (e.g., one for each of the examples above). During construction reorganization, the model seeks candidates for merging that are similar in both meaning and form; separate fixed-order constructions involving the same constituents with equivalent participant structures are prime candidates. Generalizing over these constructions leads to a new construction that contains all the shared structure of the original constructions, omitting in this case the order constraints.

Morphological constructions are similar to word combinations in involving constituency, though word-internal. The main difference is that casemarkers do not occur independently of their nominal contexts, and are first learned as part of an unstructured larger unit. Thus the relational mapping strategy for learning constituent structure cannot apply directly. We assume, however, that over time the child is able to segment words into stems and endings, based on general pattern-detection mechanisms (Peters, 1985). Then the model can merge multiple constructions with the same stem and different endings (e.g., merging *devochk-e* (girl-DAT) and *devochk-a* (girl-NOM) yields a stem *devochk-* with no participant role specified). Similarly, a particular casemarker occurring on different stems (but the same verbal context) can be merged to yield a suffix construction whose meaning pole is associated with a specific participant role (or multiple roles, since polysemous markers are allowed). The resulting stem and casemarker constructions may then serve as constituents for larger morphological constructions.

Conclusion

The work described in this paper are best characterized as first steps toward concrete computational validation of our broad research paradigm. The model is intended to offer a detailed picture of the pivotal role meaning in context plays in the acquisition of grammar. It draws on evidence from across the cognitive spectrum arguing for a construction-based grammar formalism, extensive prior knowledge, and a data-driven, incremental learning course.

We have concentrated on the acquisition of constituent structure, as demonstrated by the verb island learning experiment. Note that we have not addressed how the model learns constructions that depend on more general semantic categories; these include both general argument structure constructions corresponding to basic scenes (caused motion, manipulative activity, etc.), and casemarking constructions that generalize across verbs. These categories are not assumed to be universal, but rather must be learned based on the fine-grained semantic structure available in the ECG representation. In ongoing work we are investigating the conditions and assumptions that allow such constructions to emerge. We are also exploring the relative rates of acquisitions of different classes of verbs and continuing to test the robustness of the model to crosslinguistic data.

Acknowledgments

Thanks to Srini Narayanan, Eva Mok, Shweta Narayan, other members of the ICSI Neural Theory of Language group, and anonymous reviewers for useful feedback.

References

Bailey, D. R. (1997). *When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs*. PhD thesis, University of California at Berkeley.

Bergen, B. K. and Chang, N. (in press). Simulation-based language understanding in Embodied Construction Grammar. In *Construction Grammar(s): Cognitive and Cross-language dimensions*. John Benjamins.

Bloom, L. (1973). *One word at a time: the use of single word utterances before syntax*. Mouton & Co., The Hague.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

Bryant, J. (2003). Constructional analysis. Master's thesis, University of California at Berkeley.

Chang, N. (2004). *Constructing Grammar: A computational model of the emergence of early constructions*. PhD thesis, University of California at Berkeley.

Chang, N., Feldman, J., Porzel, R., and Sanders, K. (2002). Scaling cognitive linguistics: Formalisms for language understanding. In *Proc. 1st International Workshop on Scalable Natural Language Understanding*, Heidelberg, Germany.

Chang, N. C. and Maia, T. V. (2001). Learning grammatical constructions. In *Proc. 23rd Cognitive Science Society Conference*, pages 176–181.

Clark, E. V. (2003). *First Language Acquisition*. Cambridge University Press, Cambridge, UK.

Fillmore, C. and Kay, P. (1999). *Construction grammar*. CSLI, Stanford, CA. To appear.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. 1*. Stanford University Press.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Erlbaum, Hillsdale, NJ.

Nelson, K. (1996). *Language in Cognitive Development: The Emergence of the Mediated Mind*. Cambridge University Press, Cambridge, U.K.

Peters, A. M. (1985). *Language Segmentation: Operating Principles for the Perception and Analysis of Language*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Sachs, J. (1983). *Talking about the there and then: the emergence of displaced reference in parent-child discourse*, pages 1–28. Lawrence Erlbaum Associates.

Slobin, D. I. (1985). Crosslinguistic evidence for the language-making capacity. In Slobin, D. I., editor, *Theoretical Issues*, volume 2 of *The Crosslinguistic Study of Language Acquisition*, chapter 15. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models*. PhD thesis, Computer Science Division, University of California at Berkeley.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press, Cambridge, UK.

Tomasello, M. (1995). Joint attention as social cognition. In P., C. M., editor, *Joint attention: Its origins and role in development*. Lawrence Erlbaum Associates, Hillsdale, NJ. Educ/Psych BF720.A85.J65 1995.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.

Developing a conceptual framework to explain emergent causality: Overcoming ontological beliefs to achieve conceptual change

Elizabeth S. Charles* and Sylvia T. d'Apollonia**
*College of Computing, Georgia Institute of Technology, 801 Atlantic Drive, Atlanta, GA 30332-0280
Tel: 404-385-4035, Fax: 404-894-5041, echarles@cc.gatech.edu
** Dawson College, 3040 Sherbrooke West, Montreal, QC, H3Z 1A7, sapollonia@education.concordia.ca

Abstract

One approach to conceptual change suggests that ontological barriers may impose beliefs that contribute to learners' misconceptions and misunderstanding of many science concepts. Overcoming this hurdle requires ontological training, which we argue may be possible using concepts and behaviors related to the discipline of complexity. We investigated the difficulties related to learning complex systems concepts, specifically systems exhibiting emergent causal processes. Results showed that all students acquired the following three concepts: Multiple Levels of Organization, Local Interactions, and Probabilistic Behavior. However, all but one student remained unable to develop and use a sophisticated understanding of the concepts of Nonlinearity and Randomness. This suggests that these latter concepts may be the most deeply rooted and robust of the ontologically based misconceptions. Further research is required to investigate if this tendency toward "causal determinacy" may be modified using other types of interventions.

Introduction

Beliefs are thought to have substantial affects on how we interact with and interpret the world. Recent studies in fields such as theories of self (Dweck, 1999) and epistemological beliefs (Hofer & Pintrich, 2002) suggest that these ways of thinking also may affect learners' ability to perform certain tasks or construct certain types of knowledge. It is therefore reasonable to propose that ontological beliefs may play a significant role in learners' misunderstanding of concepts whose mechanisms are unfamiliar or completely unknown.

Chi, Slotta and deLeeuw (1994) put forward the argument that robust misconceptions associated with the learning of certain key science concepts[1] may be the result of assigning these concepts to incorrect ontological categories. It is possible also that lacking knowledge of a specific ontological category limits learners' ability to construct explanatory frameworks for a certain class of science concept.

The ontological category at the heart of this inquiry is that of emergent causal processes. It describes the behavior of phenomenon that rely on the interactions of multiple agents, all operating under the same constraints, without centralized control, influenced by flows of information with feedback loops and selection mechanisms, which generate multiple levels of organization within a system. The nonlinear and probabilistic nature of these complex systems is responsible for the seemingly magical transformations that occur between levels of the system. Put simply, emergence is characterized as the higher-level system's behavior, which arises, but cannot be predicted, from the behavior of individual lower-level entities in the system.

Conceptual Challenges of Emergence

Although we know a lot about emergent causal processes, we continue to be challenged by why these concepts pose obstacles to learners. Duit, Roth, Komorek and Wilbers (1998), and Penner (2000), among others, have studied what students learn about complex systems when provided with different types of models. From their work we know that it is possible to learn some aspects of emergent behaviors, but these studies have not articulated the dimensions nor have they looked at the potential for transfer of this explanatory framework to achieve conceptual change.

Although students may be exposed to the behaviors and functioning of complex systems in general course work (e.g., diffusion of gases), it appears that many do not understand the concepts deeply; and they do not transfer these explanations to other instances of emergence (Jacobson, 2000). In fact, Jacobson's work shows that novice learners do not correctly attribute emergent causation to explain the behavior of complex systems whereas experts in fields such as biology and economics do so readily. Therefore we know that it is possible to use this as a generic framework as a generic to explain novel emergent phenomena. Additionally, Jacobson's results provide evidence to support the claim that expertise in certain fields may be built on a deep understanding of this emergent ontological category.

[1] Conceptual change difficulties reported in learning some important science concepts such as electricity in physics (Chi, Feltovich, & Glaser, 1981; White, 1993), gas laws and equilibrium in chemistry (Wilson, 1998), and in the biological sciences such concepts as diffusion, osmosis (Odom, 1995; Settlage, 1994), and evolution (Anderson & Bishop 1986; Brumby, 1984; Jacobson & Archodidou, 2000).

Lastly, there are powerful computer models to facilitate the acquisition of complex systems, however, the literature tells us that certain beliefs appear to limit how readily learners "see" and correctly explain the model's behaviors (e.g., Resnick & Wilensky, 1997). For instance, Resnick (1994) identifies the tendency to attribute centralized control to self-organizing behaviors of multi-agent computer models in StarLogo™. But we do not know the impact of simulations and modeling of different types of complex systems on understanding of emergent behaviors. Nor do we know if all aspects of emergence as demonstrated by these models of complex systems are equally challenging to novice learners.

Our interest in this paper is to take a modest step toward addressing some of these gaps in understanding how knowledge of emergent causal processes, as demonstrated in multi-agent simulations, may affect learning of certain science concepts. More specifically, we seek to identify and describe which emergent behaviors can be learned through simple simulations and modeling of emergent systems and which are more problematic for learners.

In the following sections we will describe the mixed method longitudinal case study of nine science students who participated in five, one-on-one, one-hour long inquiry-based sessions using simulations designed with StarLogo™. We will also describe the coding taxonomy (Complex System's Taxonomy – CST) which we developed to analyze the transcribed audio data collected.

Material and Methods

Sample

We recruited science students, between the ages of 17 and 18, in their freshmen year at a pre-university English college in Quebec (equivalent to grade 12). From this cohort we selected nine case studies using a purposeful sampling strategy (Creswell 2002). A major criterion for selection was the students' level of motivation and persistence[2].

The students' ages and academic experiences guaranteed that their formal knowledge of complex systems and emergent processes was limited or non-existent. However, we administered a pre-test to establish a baseline of their entry-level knowledge of these concepts (these data are not discussed in this paper).

Instruction

The treatment consisted of five, 60 minute one-on-one inquiry-based sessions. Each session was comprised of two major components: (a) StarLogo computer simulations, and (b) cognitive scaffold in the form of coach/interviewer. The simulations were selected based on the ratings of four subject matter experts. The criteria were that the simulations should demonstrate emergent causal processes, and may in fact exhibit other behaviors of complex dynamic systems. The resulting treatment consisting of three simulations, and one tutorial, (Slime - session 1; FreeGas – session 2; StarLogo programming tutorial – session 3; no simulation – session 4; Wolf-Sheep – session 5) selected from a bank of over 12 other existing StarLogo simulations that also were judged appropriate for grade 12 science students. The simulations finally selected also have a prior history of providing learners with opportunities to learn about concepts of complexity (e.g., Resnick & Wilensky, 1997). This should not suggest that each simulation presented the same level of affordance for learning complexity concepts, however, they all held the potential to demonstrate some level of the more anticipated behaviors (i.e., non-isomorphic multiple levels of organization, decentralized control, randomness, nonlinearity, probabilistic behavior, and dynamic homeostatic behaviors). A question of interest that emerged from the observations was the differential effects of the different types of complexity represented in the simulations (i.e., the tightly coupled organization modeled in Slime simulation, versus the dissipative systems of FreeGas, and the somewhat in-between system modeled in Wolf-Sheep). Lastly, we also did not know the impact of presentation sequence but decided to keep this constant across learners to reduce the variability among cases although it prevented us from learning more about this question.

Procedure

Over the period of five one-hour sessions, spanning a 7-week period each of the nine learners met individually with the coach in a research lab and worked with the simulation assigned for the session (see above). As they explored the assigned simulation, learners were asked to describe their observations related to the behaviors of the agents (i.e., slime mould, gas molecule, turtle, wolf-sheep) and construct and articulate possible explanations for these behaviors. The literature suggests that these causal explanations would reveal the underlying component beliefs/mental models (deterministic "clockwork" component beliefs used by novice learners versus nondeterministic "emergent" component beliefs used by experts) used to interpret these phenomena (e.g., Chi, et al, 1994; Jacobson, 2000). These statements could then be coded and triangulated with data collected relating to shifts in component ontological beliefs that forms part of a larger study (Charles, 2003).

Based on the literature (e.g., Resnick, 1994) we anticipated that learners would be able to identify and describe behaviors common to complex dynamic systems during their sessions. Therefore the ability to

[2] Learning Approach Questionnaire (LAQ) created by Donn (1989) was used to assess motivation. We selected participants with high internal motivation to ensure persistence with the task over course of this longitudinal study.

comply a list of the structural similarities between the simulations was viewed as the high level objective of this experience. At the conclusion of each session learners were asked to attempt to produce a list of behaviors exhibited by the simulation. If necessary they were reminded of the list complied from their previous sessions. Lastly, they were provided with a list of concepts, which may be related to either complex or simple systems and asked to construct a concept map. These data are not described in this paper.

Data collection, coding, and analysis

We collected direct observational data (audio and video tapes of the instructional activities), written documents (students' responses at the pretest and posttest), and interview data. A coding scheme entitled Complex Systems Taxonomy (CST) was developed to determine students' conceptual understanding of the various aspects of complex systems. Adapted from Jacobson (2000), it reflects concepts presented by Holland (1995), Bar-Yam (1997), and others. This "fine grain" overly represented coding scheme was used purposefully to ensure that all articulated observations of systems' behaviors could be coded (see Appendix for complete CST). Post analysis results allowed for narrowing of the taxonomy for future use.

Results

One of the major themes constructed from the categories to emerge from the interviews was that the different simulations facilitated the acquisition of different aspects of complex systems. The results in Table 1 represent the total responses aggregated across students. It displays the percentage of responses within each complex systems component.

Table 1: Distribution of responses (percentages) within Complex Systems Taxonomy (CST) for each simulation.

CST Concept	Simulations		
	Slime	FreeGas	Wolf-Sheep
ML	49.1	35.2	32.5
LI	22.4	25.1	35.3
OS	2. 8	14.0	8.6
PR	11.4	19.3	13.4
RB	5.1	3.0	2.3
TA	4.20	0.26	0.19
FL	1.10	0.43	2.90
DE	0.68	1.20	0.70
SR	0.74	0.00	0.13
DC	1.30	0.68	1.40
DI	0.32	0.00	0.38
NL	0.00	0.15	0.38
PA	1.40	0.26	1.00

ML is Multiple Levels of Organization, **LI** is Local Interactions, **OS** is Open Systems, **PR** is Probabilistic Behavior, **RB** is Random Behavior, **TA** is Tags, **FL** is Flows, **DE** is Dynamic Equilibrium, **SR** is Simple Rules, **DC** is Decentralized Control, **DI** is Diversity, **NL** is Nonlinear, **PA** is Pattern Recognition

To answer the question *what difficulties might students experience with learning the concepts involved with emergent causal processes* we analyzed the data both at the level of students and at the level of emergent causal process concepts. Thereby producing the two levels of analysis reported below.

Student level analysis

Figure 1 illustrates the combined scores on the CST for each student across all sessions. On this basis students could be classified into four groups:

- *Sophisticated Emergent Causal Processes (ECP) Identifier* (CST score > 75). This describes Greg who is considered an outlier at the high end.
- *High Moderate Emergent Causal Processes (ECP) Identifier* (CST score between 60 and 70). This describes Mitch, Sidney and Sam.
- *Moderate Emergent Causal Processes (ECP) Identifier* (CST score between 40 and 50). This describes Walter and Norman.
- *Novice Emergent Causal Processes (ECP) Identifier* (CST score between 30 and 40). This describes Emilie, Penny, and Monique (an outlier at the low end).

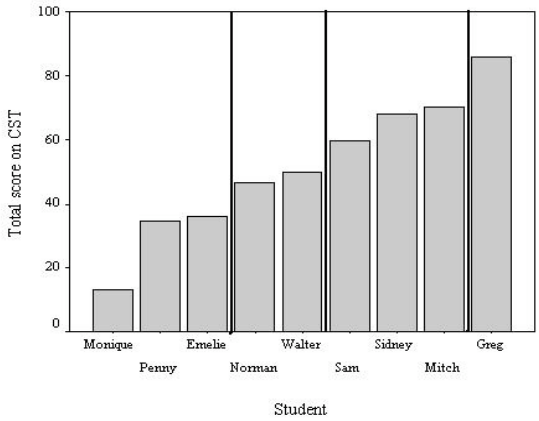


Figure 1: Student's understanding of Complex Systems concepts over three simulations.

Concepts level analysis

The results of Table 2 show the number of statements (relative to each student's total number of statements) that were coded (using the CST) into each Complex Systems concept. Thus, it allows us to make a provisional decision on whether each student observed and therefore discussed the Complex Systems concepts. If one arbitrarily, takes a value of 1 as the cutoff point, we can provisionally conclude that all students including the three Novice ECP Identifiers (Monique, Emilie, and Penny) observed and discussed the

concepts of "multiple levels of organization", "local interactions", and "probabilistic causes". All the other students also observed and discussed the concept of "random behavior". The major difference between the Moderate ECP Identifiers (Norman and Walter) and the High ECP Identifiers (Sam, Sidney, and Mitch) was in the general strength of their responses. On the other hand, the Sophisticated ECP Identifier (Greg) not only had a greater response to the latter concepts, he also observed and discussed more concepts, namely "flows" and "dynamic equilibrium"

Table 2: Relative number of statements made by each student coded into Complex Systems concepts over three simulations.

CST Concepts	Novice ECP			Moderate ECP		High-Moderate ECP			Sophisticated ECP
	Monique	Emilie	Penny	Norman	Walter	Sam	Sidney	Mitch	Greg
ML	6.7	21.0	15.7	16.1	20.9	22.3	30.8	26.0	25.8
LI	3.0	8.5	8.8	11.3	13.0	17.5	17.1	20.6	25.7
OS	0.0	2.2	3.6	2.7	4.1	4.1	2.4	5.4	12.1
PR	2.6	2.3	3.6	5.8	6.7	7.3	10.8	11.6	13.0
RB	0.0	0.5	0.9	3.1	1.0	2.5	3.7	2.7	2.4
TA	0.4	0.4	1.1	1.6	1.1	0.8	0.4	1.5	1.3
FL	0.2	0.0	0.0	0.4	0.0	0.1	0.4	0.2	1.2
DE	0.5	0.1	0.2	0.1	0.6	0.3	0.4	0.6	1.0
SR	0.1	0.0	0.2	0.1	0.0	0.4	0.4	0.0	0.4
DC	0.1	0.6	1.9	0.2	0.3	1.3	0.5	0.3	0.3
DI	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	0.4
NL	0.0	0.0	0.1	0.1	0.2	0.0	0.1	0.2	0.5
PA	0.0	0.0	0.4	0.0	0.8	1.8	0.1	0.1	0.9

ML is Multiple Levels of Organization, LI is Local Interactions, OS is Open Systems, PR is Probabilistic Behavior, RB is Random Behavior, TA is Tags, FL is Flows, DE is Dynamic Equilibrium, SR is Simple Rules, DC is Decentralized Control, DI is Diversity, NL is Nonlinear, PA is Pattern Recognition

The interpretation that concepts, which had low counts on the CST scheme (e.g., random behavior, nonlinear effect, decentralized control, dynamic equilibrium), suggests that students did not observe them is not the only conclusion to be drawn from these data. It may indicate that learners readily recognized the behavior described by the concept and chose to focus instead on other concepts that were more challenging or interesting. It may also indicate that the simulation did not offer sufficient affordances for learning that concept.

Discussion

Chi and colleagues (e.g., Chi et al., 1994; Chi 2000) have long proposed that ontological training will remove ontological barriers, which they speculate create the misunderstandings observed when learning certain scientific concepts. Our study shows that not all of these identified barriers are equally daunting. In fact, our study confirmed that using the selected intervention, it was possible to hurdle two of the barriers (Multiple Levels and Local Interactions) identified as problematic by Chi (2000). Our results also suggest that two (Nonlinearity and Random Behaviors) of a possible six complex systems concepts are either not affected by this intervention with its relative affordances for learning complex systems concepts (i.e., non-isomorphic multiple levels of organization, decentralized control, probabilistic behavior, and dynamic homeostatic); or, that these concepts represent a deeper level of entrenched beliefs and require some other type of intervention or condition before substantial change will be observed. The more important of these two is randomness because it is an addition to the list of barriers identified by Chi (2000).

Adding to the list of Ontological Barriers - "Causal Determinacy"

One of the ontological barriers not identified by Chi (2000) is the attribution of causal determinacy (i.e., difficulty in acquiring the concept of random actions). This current study shows that, possibly because of weak affordances of the simulations, students experienced difficulty with the notion of randomness. Klopfer and Um (2000) in a study of fifth and seventh grade students using StarLogo in a scaffolded learning environment called "Adventures in Modeling" also demonstrated that students experienced difficulties with learning the concept of random events; although in the latter portion of their 14 sessions intervention, students were able to grasp this concept.

The evidence from the study reported here and from the larger study (Charles, 2003) is that all the learners at some level were challenged by randomness. In fact, it was the main stumbling block for Greg who otherwise acquired an understanding of all the emergent causal processes without exceptional cognitive struggle. For example, Greg when provided with an ontological prompt during session 1, answered with an explicit statement describing the Slime mould model as being deterministic. His view was that the computer program limited the options and therefore the outcome was determined a priori, therefore predictable.

> Greg: Yeah, I think it's more of a deterministic system. Because like even looking at the way that this is set up there was a minimum number of turtles that you could have and I think it starts off as a system that has a plan and that all the other variables just act on whether like it's your plan ... so you have a deterministic system.

What this suggests perhaps is that even though learners accept the randomness of some happenings, as indicated in their answers to the question about ants foraging, at a deeper level they struggle to accept the lack of some means of predicting future outcomes (even by infinitesimally small or remote means). This deep level understanding is further confounded by the limitations of the programmed environment of the simulations, which indeed may confirm beliefs that there is some level of predictability because random number generations machines are behind these calculations. This is the level of discussion that Greg, Mitch and Sidney all at some point conducted with the coach.

How then did any of the learners show signs of acquiring a deeper level understanding of this concept? The evidence suggests that Greg was the only case to describe random actions at the deeper level of understanding as an element of true causal indeterminacy and "noise". He appeared to accomplish this as a consequence of both cognitive scaffolding and his domain knowledge. During the final interview session, one year after the intervention, Greg was asked to explain his concept map. In this discussion, he elaborated on the role played by random actions in the behavior of systems. This required him to reflect and in doing so he referenced his course work from biology and how the "noise" of random events creates the "possibilities" of the future states.

> Greg: ...so that creates um, randomness, and that creates possibilities, also. That if there were no random events, then you wouldn't have those possibilities. Um, but all these chance events, they, when they get absorbed into the complex system, they have very little effect. It's like throwing a pebble into a river. Sure, you might course the river in a one in billion chance or something, but chances are it does nothing. It's not going to affect the flow of the river in any way. Uh, so, what that means is that complex systems, they follow more rules of probability, and they, they... so nothing is for sure I guess, there is always the element of chance involved. But they're [complex systems] by and large more predictable than simple systems.

The attribution of causal determinacy is a key obstacle to understanding emergent causal process for most learners. This arises either because of the learners' component beliefs, as in the instantiation of the case study Norman, or because of the confounding of concept and programming limitations as demonstrated by Sidney, Mitch and overcome by Greg. The contention may come as no surprise to those investigating the cognitive processes involved in reasoning about uncertainty (e.g., Shauhnessy, 1992). Metz (1998) points to the spurious causal attributions that result from misunderstanding of randomness and probability. What is surprising is that this same barrier also may account for a major difficulty in learning emergent causal processes such as evolution. This contention is supported by research from Zaïm-Idrissi, Désautels, and Larochelle (1993). In their study working with 15 biology students (master's level) they concluded that the majority of the sample held deterministic forms of reasoning about the topic of evolution. Furthermore, they uncovered several inconsistencies in the belief systems of the study's participants, primarily, the conflict between deterministic and probabilistic reasoning.

Therefore, it is possible that this causal determinacy attribution may be one of the most widely interconnected beliefs that affect other related beliefs such as probabilistic causes, and even decentralized control. It may well fit Chinn and Brewer's (1993) description of the evidentiary supporting schema. They state: "It appears, then, that well-developed schemas are not necessarily entrenched. The key is whether the schema is also embedded in evidentiary support and is used to support a wide range of other theories and observations that the person believes" (p. 17). Future research is required to try and untangle the possible confounding of the simulations' weak affordances and the students' ontological belief about randomness.

Conclusions

Using the complex systems' taxonomy, the results of this inquiry show that all nine case study students had little difficulty developing an understanding of three emergent causal processes: Multiple Levels of Organization, Local Interactions and Probabilistic Behavior. However, the emergent component concepts of Nonlinearity and Randomness were challenging for all. In fact, only one student, Greg, was capable of demonstrating a deep conceptual understanding of these concepts. Furthermore, his understanding grew with maturation over time, with experience from complementary content areas, and cognitive scaffolding from the coach/interviewer. Greg's persistent attempts to reason with these concepts and explain phenomena using these notions (e.g., explaining evolution of a species as dependent upon random events) may also account for his ability to acquire this knowledge.

The results of our study also show that the affordances for learning aspects of emergent causal processes, and concepts of complexity, offered by multi-agent simulations and modeling are highly related to the type of complex system represented and also to the students' background understanding of science. In particular more learners had difficulty learning with representations (simulations) of dissipative system complexity (FreeGas) compared to those using

representations of tightly coupled organization models of complexity (Slime).

In summary, this investigation provides evidence that it is possible, using simple simulations and scaffolding, to facilitate the learning of some aspects of an emergent causal explanatory framework. However, other components of emergence, which are linked to non-deterministic (i.e., randomness) and nonlinear conceptions are not easily acquired and may represent the more deeply entrenched ontological beliefs. Further research is needed to examine these specific aspects of emergent causal frameworks and the effectiveness of other types of instructional simulations and tools.

References

Bar-Yam, Y. (1997). *Dynamics of complex systems*. Reading, MA: Addison-Wesley.

Charles, E. (2003). An Ontological Approach to Conceptual Change: The role that complex systems thinking may play in providing the explanatory framework needed for studying contemporary sciences. Unpublished doctoral dissertation.

Chi, M.T.H. (2000). *Misunderstanding emergent processes as causal*. Paper presented at the Annual Conference of the American Educational Research Association, April 2000.

Chi, M. T. H., Slotta, J. D., & deLeeuw, N. (1994). From things to processes: a theory of conceptual change for learning science concepts. Learning and Instruction, 4, 27-43.

Chinn, C. & Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63(1), 1-49.

Duit, R., Roth, W.M., Komorek, M. & Wilbers, J. (1998). Conceptual change cum discourse analysis to understand cognition in a unit on chaotic systems: towards an integrative perspective on learning in science. *International Journal of Science Education*, 20(9), 1059-1073.

Holland, J. H. (1998). *Emergence: From chaos to order*. Reading, MA: Perseus Books.

Jacobson, M.J. (2000). *Butterflies, Traffic Jams, & Cheetahs: Problem solving and complex systems*. Paper presented at American Educational Research Association annual meeting in Atlanta.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, 6(3.), 41-49.

Klopfer, E., & Um, T. (2000). Young adventurers: Modeling of complex dynamic systems with elementary and middle-school students. *Proceedings of the Third International Conference of the Learning Sciences* (pp. 214-220).

Metz, K.E., (1998) Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction*, 16(3), 285-365.

Penner, D.E. (2000). Explaining systems: Investigating middle school students' understanding of emergent phenomena. *Journal of Research in Science Teaching*, 37(8), 784-806.

Resnick, M. (1994). *Turtles, termites and traffic jams: Explorations in massively parallel microworlds*. Cambridge, MA: MIT Press.

Resnick, M., and Wilensky, U. (1997). Diving into complexity: Developing probabilistic decentralized thinking through role playing activities. *Journal of the Learning Sciences*, 7, 153-172.

Zaim-Idrissi, K., Desautels, J., & Larochelle, M. (1993). "The map is the territory!" The viewpoints of biology students on the theory of evolution. *The Alberta Journal of Educational Research*, 39(1), 59-72.

Acknowledgements

This research was funded in part by the Quebec Provincial Government program *Programme d'aide à la recherche sur l'enseignement et l'apprentissage* (PAREA). We thank Janet Kolodner for her advice during the writing of this paper and Gary M. Boyd for his help to the first author while working on her PhD. thesis at Concordia University, Montreal, Canada.

Appendix

COMPLEX SYSTEM CODING TAXONOMY
1. Local interactions.
2. Simple rules
3. Decentralized control
4. Random behavior
5. Tags
6. Flows
7. Internal models
8. Diversity/ variability
9. Modularity
10. Pattern formation
11. Open/closed systems
12. Multiple Levels
13. Probabilistic
14. Nonlinearity
15. Criticality
16. Dynamic equilibrium
17. Adaptation
18. Selection
19. Time scale.
20. Multiple causality

A Cross-Linguistic Study of Phonological Units: Syllables Emerge from the Statistics of Mandarin Chinese, but not from the Statistics of English

Train-Min, Chen (trainmin@alumni.ccu.edu.tw)
Department of Psychology, National Chung Cheng University
160 San-Hsing, Min-Hsiung, Chia-Yi 621 Taiwan, R.O.C.

Gary S. Dell (gdell@s.psych.uiuc.edu)
Beckman Institute, University of Illinois at Urbana-Champaign
405 N. Mathews Ave, Urbana, Illinois 61801 U.S.A.

Jenn-Yeu, Chen (psyjyc@ccu.edu.tw)
Department of Psychology, National Chung Cheng University
160 San-Hsing, Min-Hsiung, Chia-Yi 621 Taiwan, R.O.C.

Abstract

This study explored the statistical patterns of English and Mandarin Chinese sound sequences, by comparing their learning in a simple recurrent network. Experiment 1 showed that vivid syllable structure emerged from the sound sequence of Mandarin Chinese. Experiment 2 further demonstrated that the emerged syllable structure of Mandarin Chinese is considerably more salient than that of English. We claim that the more salient syllable structure in Mandarin Chinese inputs is one reason why syllable units are particularly emphasized in its processing in comparison to English.

Introduction

According to linguistic theory, the sound patterns of all languages are hierarchical. Segmental speech sounds (or phonemes) are concatenated into syllabic constituents (onset, rhyme), which join to form syllables, which, in turn, are the constituents of larger units such as feet and words. In psycholinguistic theories of production, each kind of unit plays a part, but some units are more salient than others. For example, standard theory (e.g. Levelt, Roelofs, & Meyer, 1999) holds that lexical items are stored as sequences of segments. The organization of these sounds into syllables, however, is not stored, but rather is computed during production. Evidence against stored syllables comes from two principal sources:

(1) *The absence of syllabic speech errors.* For example, exchanges of non-morphemic syllables are very rare. You would never hear "napkin" spoken as "kinnap."

(2) *The absence of syllabic priming effects that cannot be attributed to segmental units*. For example, naming a word is speeded when a masked orthographic prime syllable that matches the initial sounds of the word precedes it. In many such studies, this priming is unaffected by whether or not the prime syllable corresponds to a whole syllable in the target word (e.g. Schiller, 2000).

The conclusion that syllables are not units of storage has been based on studies of Germanic languages such as English and Dutch. Our own production research has demonstrated that this conclusion is not warranted for all languages. Here, we briefly review our studies of production in Mandarin Chinese, which show that the syllable is far more unitary than has been found in English and Dutch. Then, we present two computational studies involving the learning of Mandarin and English sound sequences. These studies suggest that cross-linguistic differences in the salience of the syllable in production emerge from the statistics of the sequences.

Speech Error Data Psycholinguists believe that the commonness of slips of units such as segments or words provides evidence that these units are psychologically real. Research has shown that whole syllables, in contrast, rarely move, at least in English and in related languages, and the few apparent cases can be otherwise explained as morphemic or segmental slips (Chen, 2000). However, this is not the case for Mandarin Chinese. Chen (2000) demonstrated that syllable movement errors indeed happen in the natural speech of Mandarin Chinese at a rate of 10,000 times greater than would be expected if these errors were the result of independent segmental slips. Importantly, these syllable errors "strand" tone; only the segmental part of the syllable moves. One such example is that the word [ching1zhuo2du4] (清濁度 'clarity') slipped to [ching1du2du4], an anticipation of the third syllable [du]. This stranding of tone rules out the explanation that these are slips of morphemes or characters.

Masked Priming Data Studies of masked priming of word naming in Mandarin Chinese also point to the syllable as a unit. Chen, Lin, and Ferrand (2003) found that when the segmental overlap between the target word to be produced and a preceding character prime constituted a complete syllable, the response time was faster than when it did not. This result was obtained for both CVC and CV-glide syllables (Lin & Chen, 2003).

Implicit Priming Data The implicit priming paradigm is a production task that is useful for discovering relevant phonological units. Participants learn several cue-target word pairs, and later must say the target member of the pair as quickly and correctly as possible upon seeing its paired cue word. Using this task in Dutch, Meyer (1991) showed that when target words in a set shared their initial portions,

responses were faster than when they did not. Moreover, when the shared initial portions did not constitute a syllable, priming was still observed and correlated positively with the number of shared segments. Hence, the observed priming effects seem to derive from shared segments. Implicit priming in Mandarin Chinese was quite different from that in Dutch (Chen, Chen, & Dell, 2002). Crucially, priming was only found when target words shared the segments of entire first syllable, or the segments and the tone together. Accordingly, the priming effects in Mandarin Chinese derive from syllable rather than segment sized units.

Why are Mandarin Chinese speakers more sensitive to syllable units than other speakers?

We believe that the production system reflects the language. Relative to English, Mandarin Chinese has few syllable types. English has more than 10,000 syllables, and Mandarin only around 400 (without counting the tone) or 1,200 (when tones are considered). In addition, re-syllabification commonly occurs in English speech, e.g. 'demand it' becomes 'de-man-dit', but not in Mandarin (Kuo, 1994). Moreover, English, but not Mandarin, involves ambisyllabicity, the apparent membership of a consonant in more than one syllable e.g. 'hap-py'. These properties make the syllable a more efficient planning unit for Mandarin Chinese. In this paper, we explore the hypothesis that these properties lead to sound sequences whose statistical structure favors syllabic, as opposed to word and segmental, units. More generally, we claim that the relative importance of various unit types is a product of experience and test this claim by adopting the computational approach of Elman (1990) to phonological sequences.

Analysis of Sequences by Simple Recurrent Network

The Simple Recurrent Network (Elman, 1990)

A simple recurrent network (see Figure 1) is a three-layer feedforward network, in which input, hidden and output layers are linked by forward trainable connections in a distributed fashion, i.e. fully connected. Crucially, the network includes another input layer, the context layer, which serves as the dynamic memory of the network. The connections from context layer to hidden layer are trainable and distributed, but the recurrent connections from hidden layer to context have fixed weights of 1.0 and are one-to-one. Functionally, the recurrent connections behave much like a copier, which duplicate the activation pattern of hidden layer at a particular time step on the context layer. Hence, output at any given time step is shaped by the network's previous internal state together with its current input. These properties make simple recurrent networks useful models of how people implicitly learn the structure of sequences.

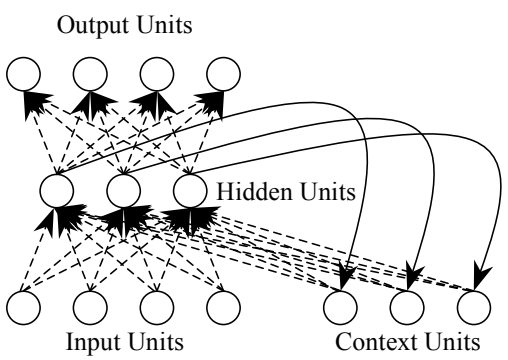


Figure 1. Elman's (1990) simple recurrent network, in which activations are copied from hidden layer to context layer on a one-for-one basis, with fixed weight of 1.0. Dotted lines represent trainable connections.

Word Structure in a Letter Sequence

Elman (1990) examined the statistical structure of English letter sequences by having the simple recurrent network predict the letter that follows the current input letter. Trainable weights were changed to the extent that the prediction was incorrect. The degree of prediction error was highly correlated with word boundaries. Error tended to spike up for word-initial letters, and declined as a function of the serial position for letters within words (see Figure 2). Hence, the relatively higher error for word-initial letters successfully demonstrated that the simple recurrent network discerns the word structure in the letter sequence without providing it with any word boundary cue during training. It appears that the word is the dominant unit, at least in English letter sequences. Next, we perform the same analysis on spoken Mandarin Chinese input.

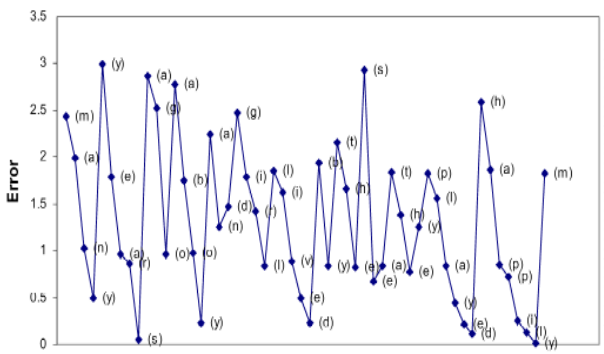


Figure 2. Graph of root mean squared error plotted over time in the letter prediction task (Elman, 1990, Figure 6). The letter to be predicted each time is shown in parenthesis.

Experiment 1: Exploring the Syllable Structure in Mandarin Chinese Inputs

In this experiment, Mandarin Chinese input was assessed by inspecting the relative performance of the network on predicting (1) word-initial sounds, (2) syllable-initial sounds that are not also word initial, and (3) the sounds within the syllable (hereafter, within-syllable sounds). If predicting syllable-initial sounds is harder than predicting within-syllable sounds, syllable boundaries will be protruded, that

is, syllable units will show up. Besides, if predicting the syllable-initial sound is as difficult as predicting the word-initial sound, it suggests that the syllable is the sole emergent unit. This is because word-initial sounds themselves are also syllable-initial, and the syllable unit alone could explain the pattern without postulating a word level.

Method

Simulation Materials and Sound Representation The simulation material came from a 30-minute stretch of a children's radio broadcast program (for ages 6 and upward) downloaded from the "National Education Radio" website at **http://www.ner.gov.tw.** It contained 5,394 sounds (sounds differing only in nasal features were regarded as different sounds), comprising 2,072 syllables and 1,300 words. For the simulation, each sound was represented as a 52-bit binary vector, 47-bit for the segment (because 47 different segments were involved) and 5-bit for the tone (because there are five lexical tones in Mandarin Chinese).

Simulation Design and Network Training The performance of the network on predicting the word-initial, syllable-initial, and within-syllable sounds was examined under 18 conditions created by crossing three factors: (1) tonal information (Syllable condition: tone distributed to each sound of the syllable; Rhyme condition: tone only distributed to sounds of the rhyme; Without condition: no tonal information), (2) the number of training epochs (20, 40), (3) the number of hidden units (25, 100, 200). Performance of the network was evaluated by two kinds of scores: (1) the Error Rate, calculated by regarding the output vector incorrect if the proposed target vector was not its closest vector, and (2) the Euclidian Distance between the actual output vector and its target vector. For both ways of scoring, the larger scores index greater unpredictability and, hence, a more salient boundary.

Before training, the connection weights were initialized randomly in the range of ±0.5. Training began with presenting the network a sequence of input vectors one at a time, and having the network learn to predict the next by adjusting the connection weights with the backpropagation algorithm. Leaning rate and momentum were set to 0.3 and 0.9 respectively.

Results and Discussion

Throughout the study, the Error Rate and Euclidian Distance displayed the same pattern. We present only the latter measure in Figure 3, which illustrates the results of single, but typical, condition. Sounds at word and syllable boundaries were much more difficult to predict than within-syllable sounds. The difficulty of predicting the syllable-initial sounds, however, was quite similar to that of the word-initial sounds. This pattern held no matter how many hidden units the network was equipped with, how many epochs of training had passed, and how (or even whether) the tone was represented. These findings strongly suggest that the syllable is functioning as a unit. Predictability is very high within the syllable, and the word boundary does little to increase uncertainty beyond that associated with the syllable boundary. Whether *any* word structure exists will be statistically examined in Experiment 2 when we directly compare Mandarin Chinese with English.

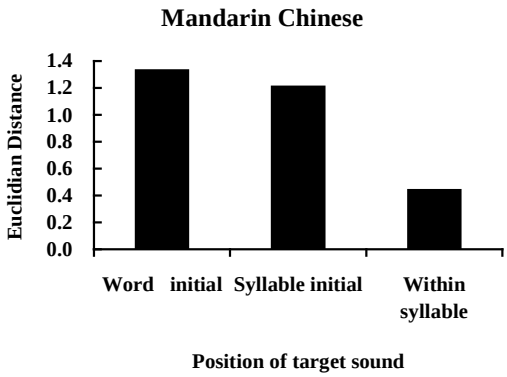


Figure 3. Average Euclidian Distance for predicting sounds in different positions. (Simulation condition: 100 hidden units, 20 epochs, tone was distributed to each sound of the syllable)

An important side result of this experiment is that the pattern shown in Figure 3 was, for the most part, independent of how we represented tone, or even whether we represented it at all. For example, if tone was not represented, the condition corresponding to the one illustrated in Figure 3 led to distances of 1.00, 0.94, and 0.56 for word-initial, syllable-initial, and within-syllable sounds, respectively. Thus, the segmental pattern alone is more than enough to motivate dominant syllable-sized units. In fact, the speech error study of Chen (2000), the implicit priming study of Chen, Chen, and Dell (2002), and the masked priming study of Chen, Lin and Ferrand (2003) all suggested that tone-less or segmental syllables as well as syllables with tone function as important production units in Mandarin. The findings of Experiment 1 are quite consistent with these data.

Experiment 2: Comparing Sound Patterns in English and Mandarin Chinese

The second experiment compared the sound distributions in English and Mandarin Chinese inputs directly. This was achieved by replicating the prior experiment using English and Mandarin Chinese versions of comparable simulation materials. The experiment also manipulated the nature of the representation of diphthongs, that is, whether they are considered to be one or two sounds. Because the prior experiment had shown that the supra-segmental (tone) information, the number of the hidden units, and the number of training epochs did not matter, the present experiment was conducted without supra-segmental information (tone or stress), and with a constant 100 hidden-unit network trained for 20 epochs.

Method

Simulation Materials The simulation materials consisted of 10 short English-Mandarin Chinese bilingual children's stories, downloaded from the "Mandarin Daily News" website at **http://www.mdnkids.org.tw/**. The English versions contained a total of 6,511 sounds when diphthongs were counted as two sounds and 6,243 sounds when diphthongs were counted as one sound, 2,482 syllables and 1,949 words. The Mandarin Chinese versions contained 7,116 sounds when diphthongs were counted as two sounds and 6,472 sounds when diphthongs were counted as one sound, 2,743 syllables and 1,860 words. The principle of representing the sounds for simulation was identical to the prior experiment.

Simulation Design Four conditions were created by crossing two factors: (1) the number of sounds that diphthongs denote (one, two), and (2) languages (English, Mandarin Chinese). Aside from the Error Rate and Euclidian Distance, another score, Syllabic Saliency, was created for representing the degree of saliency of syllable structure. It was defined as where, in percentage terms, the performance of predicting the syllable-initial sound locates on a scale that is maximal at the performance of predicting the word-initial sound and minimal at the performance of predicting the within-syllable sound.

Results

The results with Euclidian Distance and Error Rate were statistically indistinguishable and so we continue to report only the distance measure. The main finding from the experiment was that, as expected, Mandarin Chinese differed considerable from English. Treating the 10 stories as "subjects" in an analysis of variance with language, type of boundary, and diphthong representation as independent variables yielded a strong interaction between language and boundary type, $F(1,18) = 197.2$, $p < .0001$. Figure 4 shows the findings from the diphthong-as-two-sounds condition. In Mandarin the predictability at syllable and word boundaries was nearly identical. The Syllable Saliency here was 96%, that is, predicting sounds at syllable boundaries was almost as inaccurate as predicting them at word boundaries. For English, the Syllable Salience in this condition was 35%, much lower than in Mandarin, $F(1,9) = 25.9$, $p < .001$. In fact, for English, syllable-initial prediction accuracy was actually closer to within-syllable than word-initial accuracy. The results were similar, but less dramatic, when the diphthong was treated as a single sound. The Syllable Saliencies for Mandarin and English were 88% and 43%, respectively, $F(1,9) = 19.4$, $p < .002$. Pooling across the diphthong treatment yielded a strong effect of language on this measure, $F(1,9) = 24.6$, $p < .001$.

Clearly, in Mandarin, the predictability was close for word-initial and syllable-initial sounds. Were word-initial sounds any less predictable than syllable-initial ones? For English, they definitely were, $F(1,9) = 28.7$, $p < .001$. For Mandarin, the very small effect was not significant in the diphthong-as-two-sounds condition, but it was in the diphthong -as-one-sound condition, $F(1,9) = 4.4$, $p < .03$.

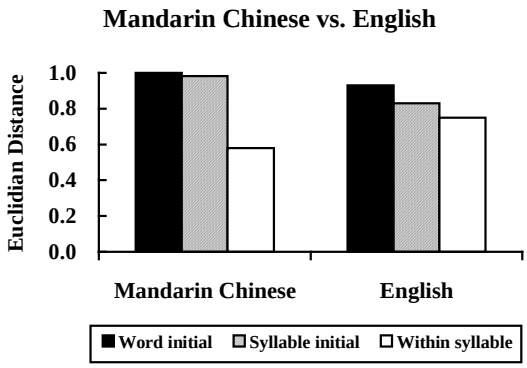


Figure 4. Average Euclidian Distance for predicting sounds in different positions in Mandarin Chinese and English (Simulation condition: diphthong as 2 sounds, 100 hidden units, 20 training epochs, without tonal information)

Discussion

To summarize, three major findings were obtained. First, vivid syllable structure emerged from the sound sequence of Mandarin Chinese. Second, the emerged syllable structure is more salient in Mandarin Chinese than in English. Third, equivalent syllable structure was found even when supra-segmental information was removed from the sound sequence. Implications of these results are discussed below.

As described in the introduction, psycholinguistic studies demonstrated that the role of the syllable is not equally emphasized in the production of English and Mandarin Chinese, a finding that hints that the sound patterns the language presents should reflect such difference. This is exactly what we demonstrated in this experiment. In Mandarin, the predictability of a sound was almost entirely determined by whether or not it is at a syllable boundary. In English, word structure was more salient, and the predictability within a syllable was not that much greater than that at syllable boundaries that are not word boundaries.

A stronger, but more speculative, interpretation of our findings makes reference to the particular kind of model that we used to assess predictability, the simple recurrent network. This network architecture has been offered as an account of phonological retrieval in production (e.g. Dell, Juliano, & Govindjee, 1993). One of the advantages of such an account is that one does not need to explicitly include or exclude particular kinds of units. Rather, the weights acquired through learning lead to activation states with greater or lesser correspondence to discrete units at several levels. Hence, the learner is not faced with the all-or-none decision as to whether to have a syllable level in the system. To the extent that different languages possess gradations in the salience of units such as the syllable, this connectionist approach may help explain the cross-linguistic variation

Another finding of note was that a strong syllable structure emerged in Mandarin Chinese even when supra-segmental (tone) information was not considered. This suggests that the segmental syllable, i.e. the syllable without the tone, has statistical support in the input, and may function as a processing unit. Psycholinguistic studies of Mandarin support this hypothesis. For instance, analysis of natural speech errors indicated that sometimes a syllable moves to a new location, leaving its tone behind (Chen, 2000). That is, the slipping unit was a segmental syllable. Pan, Chen, and Chen (1999) demonstrated this effect with experimentally generated slips. Furthermore, implicit priming and masked priming findings are robust both for syllables with tones and for segmental syllables (Chen, Chen & Dell, 2002; Chen, Lin & Ferrand, 2003).

References

Chen, J.-Y. (2000). Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia, 43(1),* 15-26.

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by implicit priming task. *Journal of Memory and Languages, 46,* 751-781.

Chen, J.-Y., Lin, W.-C., & Ferrand, L. (2003). Masked priming of the syllable in Mandarin Chinese speech production. *Chinese Journal of Psychology, 45(1),* 107-120.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and Content in Language Production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17,* 149-195.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14,* 179-212.

Kuo, F.-L. (1994). *Aspects of segmental phonology and Chinese syllable structure.* Ph.D. thesis, University of Illinois at Urbana-Champaign, Illinois.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-75.

Lin, W.-C., & Chen, J.-Y. (2003). Masked priming of the segmental syllable in Mandarin Chinese speech production: More evidence. Presented orally at the 42nd Annual Meeting of the Chinese Psychological Association, October 4-5, Taipei, Taiwan.

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language, 30,* 69-89.

Pan, H.-C., Chen, T.-M., & Chen, J.-Y. (2000). GuoYu YinJie de DengLu LiCheng (The role of syllable in phonological encoding in Mandarin Chinese). Presented orally at the First Taiwan Cognitive Science Conference: A Review and a Prospect of Cognitive Psychology in Taiwan, June 23-25, Chiayi, Taiwan.

Schiller, N. O. (2000). Single-word production in English: The role of subsyllabic units during phonological encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 512-528.

Toward a Model of Comparison-Induced Density Effects

Jessica M. Choplin (jchoplin@depaul.edu)
DePaul University Department of Psychology
2219 North Kenmore Avenue
Chicago, IL 60614-3504

Abstract

A model of the effects of distribution density on evaluations of attribute values is proposed in which biases created by language-expressible magnitude comparisons (e.g., "I waited longer for the bus today than I did yesterday") serve as the mediating mechanism. The biases created by comparisons as well as the mechanisms by which comparison-induced biases could produce density effects are described. Simulation data demonstrate signature characteristics of comparison-induced density effects. An experiment found preliminary evidence in support of the view that some density effects might be comparison-induced.

Density Effects

Evaluations of attribute values such as grades (Wedell, Parducci, & Roman, 1989), taste (Riskey, Parducci, & Beauchamp, 1979), visual velocity (Sokolov, Pavlova, & Ehrenstein, 2000), prices (Niedrich, Sharma, & Wedell, 2001), income (Hagerty, 2000) and so forth often depend upon the density—or frequency—of the distribution from which judged values are drawn (Krumhansl, 1978; Parducci, 1965, 1995). In particular, evaluation functions are typically concave (downward) for positively skewed distributions and convex (concave upward) for negatively skewed distributions. Values drawn from positively skewed distributions are also often judged larger than are values drawn from negatively skewed distributions.

Several explanations for these effects have been proposed. Parducci's (1965) Range-Frequency Theory assumes that people are aware of and use percentile rank information to evaluate attribute values. Range-Frequency Theory explains the finding that evaluation functions are often concave for positively skewed distributions, because the density at the lower end of positively skewed distributions gives low values larger percentile rank scores than they would have had otherwise. The slope of the function becomes shallow at the sparse upper end of the distribution where percentile rank scores increase at a slower rate. The reverse pattern of changes in percentile rank scores in negatively skewed distributions explains the finding that evaluation functions are often convex for negatively skewed distributions.

Haubensak (1992) suggested an alternative explanation for density effects on evaluations of sequentially presented values. He argued that since people do not know the distribution density and range in advance, they tend to assume that early values are typical or average and assign them intermediate verbal labels or category ratings. After these initial labels or category ratings have been assigned, people are obliged to use them consistently. Since early values are most likely to come from the dense portion of skewed distributions, the portion of the range at the dense end of these distributions will be smaller than the portion of the range at the sparse end. To cover the entire range of values the remaining verbal labels or category ratings would have to be assigned asymmetrically.

In this paper, I propose yet another possible explanation for density effects. Namely, that some density effects might be comparison-induced (Choplin & Hummel, 2002). Verbal comparisons will tend to bias values apart in dense regions making the slope of the evaluation function steep and bias values together in sparse regions making the slope of the evaluation function shallow. These biases would make evaluation functions concave for positively skewed distributions and convex for negatively skewed distributions. The assignment of verbal labels or category ratings to these biased values might explain why values drawn from positively skewed distributions are often judged larger than are values drawn from negatively skewed distributions.

I start by reviewing the basic tenets of Comparison-Induced Distortion Theory (Choplin & Hummel, 2002) and describing how comparisons could produce density effects. I present simulation data to demonstrate signature characteristics of comparison-induced density effects and how they differ from density effects produced by other mechanisms. I then describe an experiment in which I found preliminary support for the view that some density effects might be comparison-induced.

Comparison-Induced Distortion Theory

The basic idea behind Comparison-Induced Distortion Theory (Choplin & Hummel, 2002) is that language-expressible magnitude comparisons suggest quantitative values. To investigate the meanings of English age comparisons Rusiecki (1985) gave his participants sentences such as "Mary is older than Jane" and "Martin's wife is older than Ken's wife" and asked them to report the ages they imagined. Rusiecki found considerable agreement in the values imagined by his participants. In response to the comparison "Mary is older than Jane" participants imagined Mary to be 20.2 years on average and Jane to be 17.9 years on average. In response to the comparison "Martin's wife is older than Ken's wife" participants imagined Martin's wife to be 37.2 years on average and Ken's wife to be 33.0 years on average.

Of particular interest to the current discussion, the age differences imagined by Rusiecki's (1985) participants were remarkably similar. Regardless of the particular ages they imagined, participants imagined a difference between the ages of approximately 2 to 5 years (slightly larger for larger values)—not 1 month or 30 years. Inspired by these results, Rusiecki argued that comparisons suggest quantitative differences between compared values. I will henceforth call

these quantitative differences "comparison-suggested differences," because they are the differences suggested by comparisons. In the case of age comparisons, for example, Rusiecki's results demonstrate that comparison-suggested differences are approximately 2 to 5 years (for ease of discussion I operationally define the comparison-suggested difference implied by age comparisons to be 3.5 years).

Choplin and Hummel (2002) proposed a model of attribute evaluation in which magnitude comparisons (like those investigated by Rusiecki, 1985) bias evaluations of magnitude values. In particular, they suggested that evaluations of magnitude values might be vulnerable to bias whenever values differ from the values suggested by comparisons. For example, if the actual age difference between two people were 1.5 years (i.e., less than the comparison-suggested difference of 3.5 years), then a comparison would tend to bias evaluations of their ages apart—toward a difference of 3.5 years. The younger person would be evaluated younger than she or he would have been evaluated otherwise and the older person would be evaluated older than she or he would have been evaluated otherwise. If the actual age difference between two people were 5.5 years (i.e., more than the comparison-suggested difference of 3.5 years), then a comparison would tend to bias evaluations of their ages together—again toward a difference of 3.5 years. The younger person would be evaluated older than she or he would have been evaluated otherwise and the older person would be evaluated younger than she or he would have been evaluated otherwise.

Formally, the comparison-suggested value of the smaller of two compared values (E_S; E for Expected) and the comparison-suggested value of the larger of two compared values (E_L) can be calculated from the comparison-suggested difference, D:

$$E_S = S_L - D \quad (1a)$$

$$E_L = S_S + D \quad (1b)$$

where S_L and S_S (S for Stimulus values) are the values of the larger and smaller values unbiased by comparisons respectively. Represented values are assumed to be a weighted mean of the values unbiased by comparisons and the comparison-suggested values:

$$R_S = wE_S + (1\text{-}w)S_S \quad (2a)$$

$$R_L = wE_L + (1\text{-}w)S_L \quad (2b)$$

where w is the relative weights of the two values, is bound between 0 and 1, and is constrained so as to prevent impossible values (e.g., negative years or sizes of geometric figures) from being represented. For example, assuming a comparison-suggested difference, D, of 3.5 years, a comparison between a 22-year old and a 28-year old would bias evaluations of their ages toward each other. If the weight given to comparison-suggested values were .2, then the represented age of the 22-year old would be 22.5 years and the represented age of the 28-year old would be 27.5 years. That is, the age of the 22 year old would be evaluated, i.e., treated, as if it were half a year older and the age of the 28 year old would be evaluated as if it were half a year younger.

Comparisons Might Create Density Effects

Comparison-induced biases like those just described might produce density effects. Consider, for example, the positively skewed distribution of ages presented in Figure 1 which might be approximately representative of the ages of students in a typical undergraduate classroom. Filled-in arrows represent biases created by comparisons between values that are closer together than the comparison-suggested difference and that are, therefore, biased apart by comparisons. Outlined arrows represent biases created by comparisons between values that are farther apart than the comparison-suggested difference and that are, therefore, biased together by comparisons. Values in dense regions (i.e., 18 – 22 years in Figure 1) are more likely to be closer together than the comparison-suggested difference and as a result comparisons will more likely bias evaluations apart. Values in sparse regions (i.e., 22 – 28 years) are more likely to be farther apart than the comparison-suggested difference and as a result comparisons will more likely bias evaluations together. I propose that this difference in the effects of comparisons within dense regions versus the effects of comparisons within sparse regions might produce comparison-induced density effects.

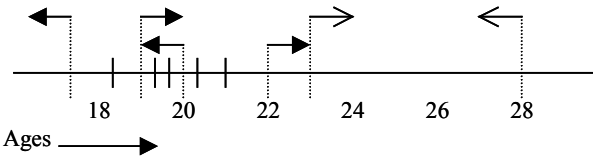


Figure 1: Comparison-induced biases that might occur in a positively skewed distribution of ages.

Modeling Density Effects

To model comparison-induced density effects, Comparison-Induced Distortion Theory requires several assumptions. First, an operational definition of the comparison-suggested difference (D) is required. In the preceding sections, for example, an adequate operational definition of the difference suggested by age comparisons was obtained from Rusiecki's (1985) study in which he queried his participants as to the differences they imagined. Alternatively, an adequate operational definition might be obtained from common real-world differences.

Second, because the number of comparisons people could hypothetically articulate as well as the sequences in which they could hypothetically articulate them is—in most cases—indefinite, assumptions about which comparisons get articulated are required. Almost any comparison scheme would produce density effects. Comparing each value to the value presented one item back, for example, would produce density effects. In the modeling presented below, I assumed that the to-be-judged item is only compared to one other item. Additionally, I assumed two constraints on the selection of this comparison item: the similarity between the to-be-judged item and candidate comparison items and the sequence in which values were presented. Although these assumptions were optional, I utilized them because they are

psychologically realistic and they have a long history in models of categorization and the psychology of judgment (see, for example, Haubensak, 1992; Nosofsky & Palmeri, 1997; Smith & Zarate, 1992).

To model the constraint of similarity on the selection of comparison items, I (along with Shepard, 1987) assumed that similarity is an exponentially decreasing function of the distance between item values. For every recently presented item j, it's similarity to i, the to-be-judged item, (η_{ij}) is calculated as:

$$\eta_{ij} = e^{-cd_{ij}} \quad (3)$$

where c is a sensitivity parameter and d_{ij} is the weighted distance between i and j in similarity space across all relevant dimensions weighted by the importance of each dimension (see, for example, Nosofsky & Palmeri, 1997).

To model the constraint provided by the sequence in which items are presented, I calculated the activation (a_{ij}) of each candidate comparison item j as:

$$a_{ij} = M_j \eta_{ij} \quad (4)$$

where M is the memory strength of exemplar j and is given by: $M_j = \alpha^{t(i) - t(j)}$ where α represents the memory decay on each trial and is bound between 0 and 1 and where t(i) and t(j) are the trials on which i and j were presented respectively (see Nosofsky & Palmeri, 1997). Selection of the item to which the to-be-judged item is compared could be accomplished a number of different ways. The choice axiom might be used to make selection stochastic. In the simulations below, maximum activation (a_{ij}) was used to make selection deterministic.

Simulation Using Artificial Values

The purpose of this simulation was to demonstrate how the model proposed above might create density effects and to point out signature characteristics of comparison-induced density effects that differentiate them from density effects created by other mechanisms. To demonstrate how the model presented above would create density effects, a computer-generated sequence of 500 values drawn from a log-normal distribution was created from equations 5 and 6.

$$V_{normal} = \sqrt{-2 \log R_1} \sin(2\pi R_2) \quad (5)$$

$$V_{\log - normal} = e^{\sigma V_{normal}} \quad (6)$$

where R_1 and R_2 are random, computer-generated values between 0 and 1. Equation 5 produced a normally distributed sequence and Equation 6 changed that sequence into a log-normally distributed sequence. To skew the log-normal distribution, σ was set at .9.

To model recall of the item to which the to-be-judged item was compared, the parameter α, representing memory decay, was arbitrarily set at .985 thereby minimizing memory losses. The parameter c, representing sensitivity to differences, was set at 0.3. Within the sequence, the second value was compared to the first value; the third value was compared to whichever of the first or the second value had the highest activation (a; see Equation 4); the fourth value was compared to whichever of the first, second, or third value had the highest activation, and so forth. In these simulations, only the most recent 7 values were candidates for comparison. Recalled values were biased by the comparison on the trial on which they were judged, but were not biased further by subsequent comparisons.

To model comparison-induced distortions, the comparison-suggested difference, D, was set at 0.38 and the weight given the comparison-suggested values, w, was .5. As suggested in Figure 1, values from the dense region were more likely to be compared to values that were less than a comparison-suggested difference away than were values from the sparse region. The values that were smaller than 1.5 (the dense lower region) were most similar to a value that was less than a comparison-suggested difference away 86.1% of the time (329/382). By contrast, the values that were larger than 1.5 (the sparse upper region) were most similar to a value that was less than a comparison-suggested difference away 40.2% of the time (47/117).

The results are presented in Figure 2. Generated values are plotted along the horizontal axis. The value of each item is plotted on the vertical axis. The filled-in squares represent comparison-biased values and the outlined circles represent unbiased values.

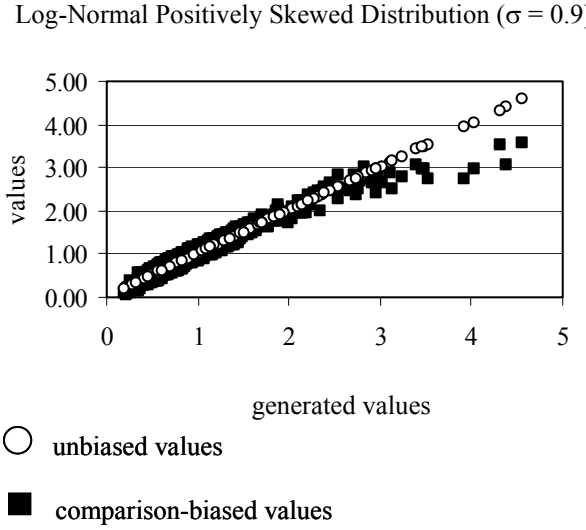


Figure 2: Simulation Results.

As shown in Figure 2, comparisons between each value and the most recent similar value created biases. Consistent with previous research using category ratings as the dependent measure, these biases produced a concave evaluation function. Seemingly contrary to previous research, however, comparisons had a tendency to bias values downward instead of upward.

This seeming contradiction can be reconciled by noting that category ratings depend not only upon representations of values but also upon the function mapping representations to category ratings. A number of functions could produce high category ratings. For example, if people were to use the range of values to make category ratings as proposed by Volkmann (1951), then even if comparisons biased representations in one direction (perhaps, as

measured by reproduction), category ratings could be biased in the other direction (see Biernat, Manis, & Kobrynowicz, 1997).

To demonstrate this possibility, range scores (i.e., [value on trial t minus smallest value up to trial t] divided by [largest value up to trial t minus smallest value up to trial t]) were calculated from the comparison-biased values. The results—after the initial 35 trials in which the range was established—are plotted in Figure 3. Filled-in squares represent comparison-biased range scores. The range transformation makes comparison-biased values comparable to the predictions of Range-Frequency Theory and so range-frequency compromise values (with the weight given to frequency set at .35) are also plotted in Figure 3 and represented as outlined triangles (see Parducci, 1965). The comparison-biased range scores mirrored the unbiased range-frequency compromise scores, suggesting that in some cases the effects of density on people's category ratings might be comparison-induced.

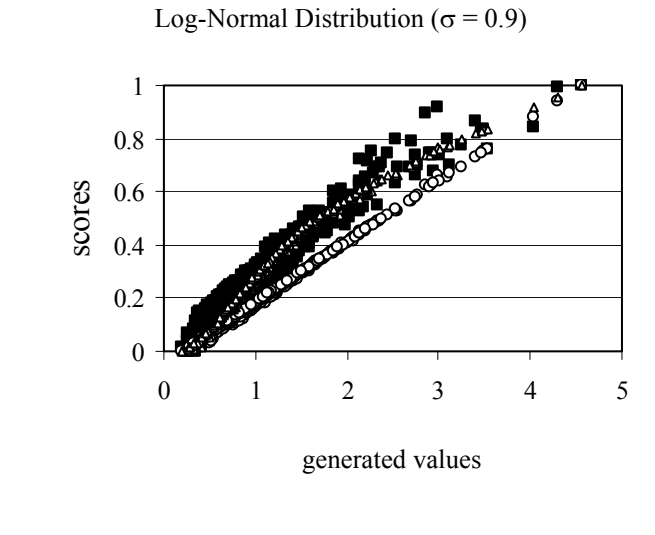


Figure 3: Simulation Data. Range scores on comparison-biased values mirror Parducci's (1965) range-frequency compromise scores.

Mirroring range-frequency compromise scores, the generated values in this simulation had comparison-biased range scores that were larger than their unbiased range scores 97.4% (487/500) of the time. I have successfully fit the model of comparison-induced density effects presented here to the results of several published density effect studies (e.g., Riskey et al., 1979) under the assumption that the comparison-biased represented values are mapped to category ratings using Volkmann's (1951) range function.

This modeling points out several signature characteristics of comparison-induced density effects that differentiate them from density effects produced by other mechanisms (e.g., range-frequency compromise). The comparison-induced biases in this simulation depended solely upon the model's knowledge of the comparison-suggested difference (D), the importance of the comparison (w), and the value retrieved for comparison. The model has no knowledge of the density of the distribution or of the percentile ranks of values and so the percentile ranks of values do not affect the model's judgments on individual trials. Rather, density affects aggregate data, because values in dense regions are more likely to be compared to values that differ from them by less than a comparison-suggested difference (and less likely to be compared to values that differ from them by more than a comparison-suggested difference) than are values in sparse regions. By contrast, Range-Frequency Theory assumes that people have implicit knowledge of percentile rank information and use that knowledge in making judgments on individual trials. Due to this difference, Range-Frequency Theory predicts that density effects ought to be observable on individual trials and Comparison-Induced Distortion Theory predicts that they ought not to be.

Experiment

The purpose of this experiment was to investigate whether the signature characteristics of comparison-induced density effects demonstrated in the modeling presented above could be observed empirically. Participants imagined that they were spending 25 days in rural Minnesota during the middle of winter and had to rely upon public transportation. The length of time they had to wait for the bus varied each simulated day and they indicated how aversive the wait would be. Half of the participants judged wait times drawn from a negatively skewed distribution and the other half judged wait times drawn from a positively skewed distribution.

Method

Participants. Seventy-three people volunteered to participate after being approached by the experimenter on the University of California, Los Angeles campus or in the surrounding community (36 in the positively skewed condition and 37 in the negatively skewed condition).

Materials and Procedure. A random sequence of 10 wait times drawn from a negatively skewed (7, 10, 13, 16, 16, 16, 19, 19, 19, and 19 minutes) or a positively skewed (7, 7, 7, 7, 10, 10, 10, 13, 16, and 19 minutes) distribution was created for each participant. Each participant's sequence was presented twice. An initial sequence of 5 days was inserted at the start of the sequence to control for primacy effects, introduce participants to the range of values they would see in the experiment, and to measure participants' baseline evaluations prior to being exposed to the density manipulation. The wait times on these 5 days were 7, 19,

13, 19, and 7 respectively and were the same for all participants. On each simulated day (simulated within a single session), the experimenter verbally told participants how long the fictitious wait for the bus was that day. Participants indicated how aversive they imagined that wait would be using a line-analogue measure in which they placed a tick at the spot along the line that was analogous to how aversive the wait was (see Schifferstein & Frijters, 1992). A stop mark on the left-hand side of the line was labeled 0=not bad and a stop mark on the right-hand side of the line was labeled 10=extremely bad. The 25 lines were presented on a single, one-page experimental handout and labeled Day 1 through Day 25.

Results and Discussion

To reduce variance caused by idiosyncratic reactions to wait times, participants' judgments during the initial sequence were used as a baseline. Each participant's judgments on trials 6 through 25 were divided by the average of her or his judgments on trials 4, 5, and 6.

Distribution density effects were revealed by differences between judgments in sparse regions versus differences between judgments in dense regions. Among participants whose wait times were drawn from the negatively skewed distribution, the difference between judgments of 7-minute wait times and judgments of 13-minute wait times (i.e., the sparse region) was reliably smaller than the difference between judgments of 13-minute wait times and judgments of 19-minute wait times (i.e., the dense region), $t(36) = 3.99$, $p < .01$. Among participants whose wait times were drawn from the positively skewed distribution, the difference between judgments of 7-minute wait times and judgments of 13-minute wait times (i.e., the sparse region) was approximately the same size as the difference between judgments of 13-minute wait times and judgments of 19-minute wait times (i.e., the dense region), $t < 1$. A 2 (distribution) x 2 (region) Mixed-Factors ANOVA found that this interaction was significant [$F(1,71) = 6.11$, $MSE = 0.22$, $p = .01$].

Because the initial sequence of 5 days inserted at the start of the experiment introduced participants to the entire range of wait times and was the same for all participants, Haubensak's (1992) model is not a viable model of the observed density effects (but note that the density effects observed in this experiment were smaller than the density effects often observed). Range-Frequency Theory and Comparison-Induced Distortion Theory remain as viable models of the observed density effects.

Range-Frequency Theory assumes that people have implicit knowledge about the percentile ranks of stimulus values and use that knowledge to judge stimulus values on particular trials. It, therefore, predicts density effects on individual trials. By contrast, Comparison-Induced Distortion Theory assumes that comparisons produce the same biases regardless of the type of distribution from which values are drawn (as long as D, w, and the values to which they are compared remain constant). It predicts density effects not on individual trials, but rather only in the aggregate and it does so only because the values to which judged values are compared differ across distributions.

The predictions of Range-Frequency Theory and Comparison-Induced Distortion Theory were tested by concentrating on differences between successive wait times of 3 minutes. At 3-minute differences the stimulus one back will likely be the most similar recent value, although occasionally the identical value 2 trials back may be the most similar recent value. To-be judged values drawn from the negatively skewed distribution were preceded by a value that was 3 minutes away 41.5% of the time (292/703). Of these to-be-judged values, 69.9% (204/292) were larger than 13, i.e., were in the dense region, and 14.7% (43/292) were smaller than 13, i.e., were in the sparse region. The to-be-judged values drawn from the positively skewed distribution were preceded by a value that was 3 minutes away 40.6% of the time (278/684). Of these to-be-judged values, 72.7% (202/278) were smaller than 13, i.e., were in the dense region, and 12.2% (34/278) were larger than 13, i.e., were in the sparse region.

Contrary to the predictions of Range-Frequency Theory and consistent with the predictions of Comparison-Induced Distortion Theory, differences between successive judgments (when actual differences were 3 minutes) were not correlated with differences in percentile rank. These correlations were not significant for descending ($r = .065$, $F < 1$) or ascending ($r = -.030$, $F < 1$) pairs from the positively skewed distribution or for descending ($r = -.047$, $F < 1$) or ascending ($r = -.050$, $F < 1$) pairs from the negatively skewed distribution.

Consistent with the predictions of Comparison-Induced Distortion Theory and not predicted by Range-Frequency Theory, the differences between judgments of successive wait times that were different by 3 minutes were biased apart, i.e., larger than their rightful proportion of 25% of the range (using participants' responses on trials 4, the large end of the range, and 5, the small end of the range, as the baseline). The differences between judgments of values that differed by 3 minutes were 33.2% of the range on average (SD = 25.1%) in the positively skewed distribution [which was significantly larger than their rightful proportion of 25%, $t(277) = 5.46$, $p < .01$] and were 44.6% of the range on average (SD = 49.6%) in the positively skewed distribution [which was also significantly larger than their rightful proportion of 25%, $t(291) = 6.77$, $p < .01$]. Further analyses did not find differences between the two distributions or between regions within the two distributions, or interactions between them. The differences between judgments of values that differed by 6, 9, and 12 minutes did not differ from their rightful proportions of the range (all t's < 1), perhaps because more similar recent items were recalled instead.

General Discussion

A model of distribution density effects in which verbal comparisons such as "I waited longer for the bus today than I did yesterday" create the observed biases was proposed.

Modeling demonstrated that comparisons might produce density effects. An experiment found preliminary empirical support for this proposal.

Future research will investigate density effects using reproduction dependent measures and investigate the predicted disassociation between representations of values (perhaps as measured by reproduction, see Figure 2) and category ratings (see Figure 3, Biernat et al., 1997). Future work will also investigate effects of distribution density on recall of values from memory. Comparison-Induced Distortion Theory predicts density effects on recall of values from memory and Range-Frequency Theory does not (see Choplin & Hummel, 2002, for a discussion).

Although in my view many density effects are likely to be comparison-induced, I do not assume that all density effects are comparison-induced. Density effects observed when all values are presented simultaneously in ascending or descending order (e.g., Wedell et al., 1989) strike me as cases where density effects are particularly likely to be categorization-induced as Parducci (1965) suggested. Additionally, even if density effects are found to be comparison-induced, the equations used to formalize Range-Frequency Theory will likely still provide a useful heuristic for predicting effects of density on judgment.

Conclusions

Some density effects might be comparison-induced. Comparisons of values in dense regions will tend to bias values away from each other, while comparisons of values in sparse regions will tend to bias values toward each other. These biases could produce density effects.

Acknowledgments

I thank Allen Parducci, Doug Wedell, and Gordon Logan for helpful conversations and Deborah Moradi for running the experiment.

References

Biernat, M., Manis, M., & Kobrynowicz, D. (1997). Simultaneous assimilation and contrast effects in judgments of self and others. *Journal of Personality and Social Psychology, 73*(2), 254-269.

Choplin, J. M., & Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General, 131*(2), 270-286.

Hagerty, M. R. (2000). Social comparisons of income in one's community: Evidence from national surveys of income and happiness. *Journal of Personality and Social Psychology, 78*, 764-771.

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 303-309.

Krumhansl, C. L. (1978). Concerning Applicability of Geometric Models to Similarity Data: Interrelationship between Similarity and Spatial Density. *Psychological Review, 85*, 445-463.

Niedrich, R. W., Sharma, S., & Wedell, D. H. (2001). Reference price and price perceptions: A comparison of alternative models. *Journal of Consumer Research, 28*(3), 339-354.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Parducci, A. (1965). Category judgments: A range-frequency model. *Psychological Review, 72*, 407-418.

Parducci, A. (1995). *Happiness, pleasure and judgment: The contextual theory and its applications*. Mahwah, NJ: Lawrence Erlbaum.

Riskey, D. R., Parducci, A., & Beauchamp, G. K. (1979). Effects of context in judgments of sweetness and pleasantness. *Perception & Psychophysics, 26*, 171-176.

Rusiecki, J. (1985). *Adjectives and Comparison in English*. New York: Longman.

Schifferstein, H. N. J., & Frijters, J. E. R. (1992). Contextual and sequential effects on judgments of sweetness intensity. *Perception & Psychophysics, 52*, 243-255.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3-21.

Sokolov, A., Pavlova, M., & Ehrenstein, W. H. (2000). Primacy and frequency effects in absolute judgments of visual velocity. *Perception & Psychophysics, 62*, 998-1007.

Volkmann, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Roherer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 279-294). New York: Harper & Row.

Wedell, D. H., Parducci, A., & Roman, D. (1989). Student perceptions of fair grading: A range-frequency analysis. *American Journal of Psychology, 102*(233-248).

Visual Cues to Reduce Errors in a Routine Procedural Task

Phillip H. Chung (Pchung@Rice.Edu)
Department of Psychology, 6100 Main Street
Houston, TX 77005 USA

Michael D. Byrne (Byrne@Rice.Edu)
Department of Psychology, 6100 Main Street
Houston, TX 77005 USA

Abstract

This paper reports an experiment designed to evaluate the efficacy of visual cues as error "interventions" in computer-based routine procedural tasks. Using two separate tasks with a well-documented error prone step, the effects of several visual cues were compared. The findings provide support for goal selection driven by environmental cues in routine procedural tasks. The importance of cue timing and movement and meaningfulness characteristics, particularly in dynamic tasks with external pressures, is demonstrated.

Introduction

History and Motivation

With the introduction of automation and computers, an outstanding arena for human error has been established. Subsequently, much effort has been given to categorize errors occurring in such situations, yet for the most part understanding of this very human phenomenon remains fairly nebulous. As John and Kieras (1996) stated in the 90s, "No methodology for predicting when and what errors users will make as a function of interface design has yet been developed and recognized as satisfactory…even the theoretical analysis of human error is still in its infancy."

Over the years, several elaborate models and taxonomies of human error have been developed for the purpose of qualitative diagnosis (e.g., Reason, 1990). Although useful for post hoc explanations, the predictive power of these is quite limited. Results from controlled studies with various error "interventions" may extend our understanding of why such cognitive errors arise, helping us not only evaluate (post hoc) but also design (predict) safer machines. Application areas to benefit abound, from aerospace to medicine.

Postcompletion Errors

Noting the general lack of specificity in the existing theories of human error, Byrne and Bovair (1997) moved to develop a computational theory for one widely cited (e.g., Rasmussen, 1982; Young, 1994) omission error, postcompletion error. Postcompletion errors can be broadly defined as errors that occur when the task structure demands "that some action…is required after the main goal of the task…has been satisfied or completed," (Byrne & Bovair, 1997, p. 32). With this particular class of error, the actor possesses the correct knowledge necessary to execute the task, usually performed frequently and correctly. Yet, for even operators highly familiar with the task, the isolation of a postcompletion step within the task structure makes omissions there not unlikely. This is particularly true when the actor is further affected by external factors such as a working memory load and/or fatigue, as well as internal tendencies such as hillclimbing (Gray, 2000; Polson & Lewis, 1990).

Some commonplace examples include forgetting to remove the original after making a photocopy, leaving a card in the ATM after withdrawing cash, and failing to replace the gas cap after filling up a car. Byrne and Bovair (1997) hypothesized that these errors were due to excessive working memory load leading to goal loss, or an omission of a step from the task at hand. Since with postcompletion errors the actor omits a specific subgoal rather than forgetting what to do altogether (the overlying main task goal), the source of the error was thought to more likely be working memory than long-term memory. A more recent study by Reason (2002) examined the photocopy example in detail, finding postcompletion errors to be the most common type of omission in that task (Figure 1).

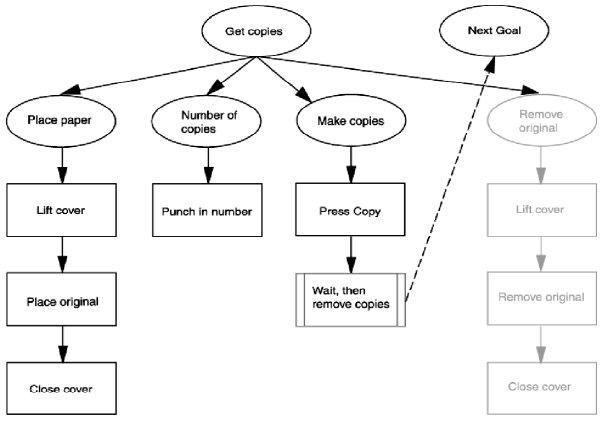


Figure 1: Photocopy task structure.

Three high-level explanatory observations were provided:

1. The emergence of the last copy generates a strong but *false completion signal* since the main goal of copying is achieved before all necessary steps (subgoals) are complete.

2. The proximity of this false signal to the end of the main task allows for the *attention to be increasingly diverted* to the subsequent task.
3. The emergence of the last copy indicates that it is no longer necessary to put in another original leaving it *functionally isolated*.

Hierarchical Control Structures and Goal Management

Many of the assumptions behind the current theory of postcompletion error reside on the foundational concept of hierarchical control structures and their retention by skilled operators. In previous studies by Byrne and Bovair (1997) and Serig (2001), participants reliably generated errors at the postcompletion steps both within subtasks as well as within the larger task, in keeping with the idea of a hierarchical task structure. Cognitive modeling work by Kieras, Wood, and Meyer (1997) has also provided strong evidence to suggest that even well practiced experts, such as telephone assistance operators, do not abandon such task hierarchies.

As Altmann and Trafton (1999) propose, the ability to break down complex tasks and problems into hierarchies and subgoals, "may be to complex cognition what the opposable thumb is to complex action." Traditionally, these types of goal-based processing strategies have relied solely on a "task-goal" stack that essentially predicts perfect memory for old goals. However, their activation-based model of memory for goals (MAGS) offers an alternative account to this approach that provides a more straightforward account for the types of errors found in human behavior.

In essence memory and the environment (i.e., dual-space, Rieman & Young, 1996; internal and external representations, Zhang & Norman, 1994) are substituted for a goal stack, and task goals are considered as ordinary memory elements with encoding and retrieval processes that must overcome noise and decay. Retrieval cues from the environment dictate the reactivation of suspended goals (e.g., Figure 1, in grey) with perceptual heuristics acting as a substitute for the stack-native last in, first out rule. This model makes several predictions about postcompletion errors and the characteristics of a successful cue:

1. Any *salient* cue (e.g., a loud beep) should be sufficient to prime a postcompletion action (suspended goal).
2. It should *not* be necessary to put the postcompletion action on the critical path.
3. Reminders at the start will *not* help a PCE at the end (masked by other goals).
4. Just-in-time priming from environmental cues are the *only* reliable reminder.

Previous Study

A previous experiment (Chung, 2004) examined the effects of a simple visual cue (red singleton onset) and a downstream error cost (in the form of a resultant mode error) on postcompletion error commission. Although neither was found to cause significant change in reaction times or error commission at the postcompletion step, the results did generate some valuable implications. First, the fact that the visual cue did not significantly reduce the number of postcompletion errors committed by the participants suggested that the cue lacked sufficient salience to prime the suspended postcompletion goal. While the sudden onset of a large red dot (against a black and white console) next to the button that needed to be pressed seemed informative enough, omissions were made regardless. Undoubtedly, participants had sufficient understanding of the task and scenario, since they could not proceed to testing without completing extensive training.

Neither did the downstream error cost (resulting in a mode error) bring about a significant change in behavior at the postcompletion step for participants. While it was expected that the visual cue would be more effective between the two treatments, it was also hypothesized that the downstream error cost would cause a change in behavior. However, the number of postcompletion errors in this condition was not significantly different from the control group, suggesting that it did not provide any significant advantage (or disadvantage) for participants. This perhaps follows findings by Serig (2001) that demonstrated error commission to be relatively independent of negative or positive feedback.

Experiment

Two tasks

Along with the original Tactical task introduced in the study by Byrne and Bovair (1997), a separate Medical task was added to distinguish the effects of the interventions. The task also included a potential postcompletion step (determined via task analysis) where the interventions could be implemented, as with the Tactical task.

Intervention Implementation

Two different interventions were introduced in this experiment: an enhanced visual cue (see Figure 2) and a mode indicator (Figure 3). These were developed specifically to address issues brought up by the findings in previous work (Chung, 2004), help pinpoint the characteristics of a successful intervention, and evaluate the predictions of MAGS (Altmann & Trafton, 1999), should one or both have a significant effect.

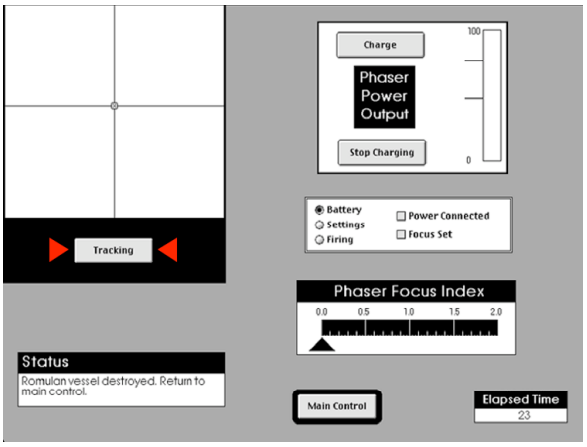


Figure 2: Tactical interface with cue (two arrows).

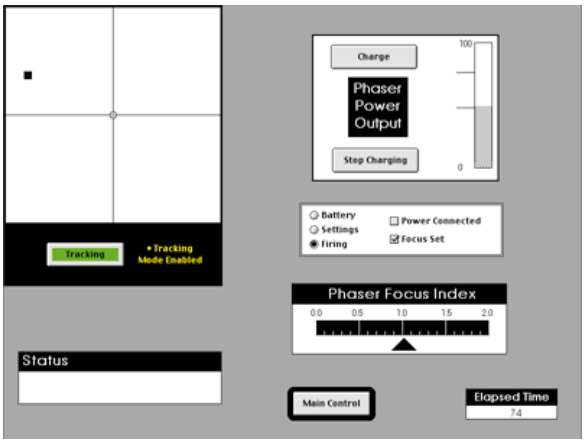


Figure 3: Mode indicator (highlighting on "Tracking" and "Tracking Mode Enabled" message).

Intervention Attributes

In a study by Monk (1986), auditory cues were employed to drastically reduce the occurrence of mode errors. Monk (1986) notes that display changes are similarly effective when the person is required to look at the relevant parts of the display at the appropriate moment in the dialogue. The related automatic processing of novel peripheral cues regardless of whether or not they are informative has been well documented (e.g., Remington, Johnston, & Yantis, 1992). Colorful visual cues are known to be effective and necessary in guiding individuals to select points of activity, such as the push plate on a door (Norman, 1988).

Research (Sutcliffe, 1995) and real-world practice indicates that the visual attributes most effective for attracting attention (warnings and indicators) on a computer interface, in order, are as follows: movement (blinking or change of position), shape and size (character font, shape of symbols, text size, size of symbols), color, brightness, shading and texture (different texture or pattern), and surroundings (borders, background color). Sutcliffe (1995) advises that these should be applied sparingly, however, as the presence of many conflicting stimuli can essentially dull their individual effectiveness. Red, green, and yellow are recommended as the optimal colors for status indicators, each corresponding to its meaning on a traffic light. To draw attention, white, yellow, and red are most effective, although yellow offers the best visibility.

Based on these recommendations and the characteristics of the failed cue from previous work (Chung, 2004), alternating red and yellow blinking arrows (Figure 2) were used for the visual cue. As per the Altmann and Trafton's (1999) predictions, the cue appeared "just-in-time" at the postcompletion step. In contrast, the mode indicator consisted of green and yellow highlighting on the "Tracking" button along with other contextual indication appearing *before* the postcompletion step. When combined with the given if-then rule at training (i.e., "If you see a mode indicator light and message, the system is on"), the mode indicator was expected to prime the corresponding goal (the postcompletion step) of turning off the Tracking system. Once the participant finishes the intermediate steps and hits the "Tracking" button a second time, the indicator disappears to indicate that the Tracking mode has ended. Exact placement of the interventions was determined through pilot studies.

Method

Participants

Ninety-one undergraduate and graduate students from Rice University aged 18 to 35 participated for course credit in a psychology course and additional cash prizes ranging from $10 to $40.

Materials

Materials consisted of a short paper-based quiz, paper-based instruction manuals for each of the four tasks (Tactical, Medical, and two filler tasks), Apple iMac computers running the Tactical and Medical applications written in Macintosh Common Lisp, stereo headphones, and a web-based general questionnaire for demographic information.

Design

This study used a two-factor between participants design with two independent variables, task and intervention. Task consisted of two conditions (Tactical and Medical) to compare the effectiveness of the interventions across task and interface. Intervention consisted of three conditions: control (no intervention), visual cue (alternating red and yellow blinking arrows), and mode indicator (mode indication for the system state change). Participants were randomly assigned to one of the six groups.

The primary dependent measure was the number of postcompletion errors made during the Tactical and Medical tasks. Other dependent measures of interest included the overall number of errors per task and performance on a

concurrent working memory task (described in Byrne & Bovair, 1997).

Procedure

Participants were run in two sessions spaced two days apart. The first session served as a training session using written documentation for each of the tasks. The major steps of the two target tasks (Tactical or Medical) are outlined in Table 1 and essentially consist of a series of key presses and mouse clicks and movements. Order of training and group assignment were randomized for every participant. Once participants successfully completed the training trial with the manual and logged three subsequent error-free trials, they were allowed to move on to the next task.

Table 1: Steps in each task.

Tactical	Medical
Charge Phaser (5 substeps)	Insert Cassette (1 substep)
Set Focus (3 substeps)	Program Rate (2 substeps)
Track Target (3 substeps)	Program Vol (2 substeps)
Fire Phaser (4 substeps)	Start Flow (2 substeps)
Return to Main Control (1 substep)	

Errors resulted in warning beeps and messages and participants were returned to the main control to restart the task. This was to prevent participants from completing training without having gone through each of the tasks at least four times with all steps done correctly and completely. When training was complete, they were reminded that they would compete for prizes in two days and given a short quiz to ensure that they had accurate working knowledge of the tasks.

The second session consisted of the test trials for both the Tactical and Medical tasks. In random order, participants completed seventeen trials of their assigned postcompletion task (Tactical or Medical) and eleven trials for each of the two filler tasks, for a total of thirty-nine trials on the test day. At testing, the experiment program emitted beeps on error commission to warn individuals but did not immediately return them to the main control or provide warning messages, as in training. Participants were encouraged to work both accurately and quickly by means of a scoring system, prizes, and an onscreen timer. A three-letter span auditory working memory task was introduced in all task conditions at testing.

Results

Data from 82 of the original 91 participants was used in the final analysis. The primary reason for the loss of data was participant failure to show up at their assigned testing date. Only one participant was removed as an outlier (Medical, cue condition). Groups broke down as shown in Table 2 below.

Table 2: Participants per group.

Condition	Control	Cue	Mode
Tactical	14	16	13
Medical	14	12	13

Postcompletion Error Frequency

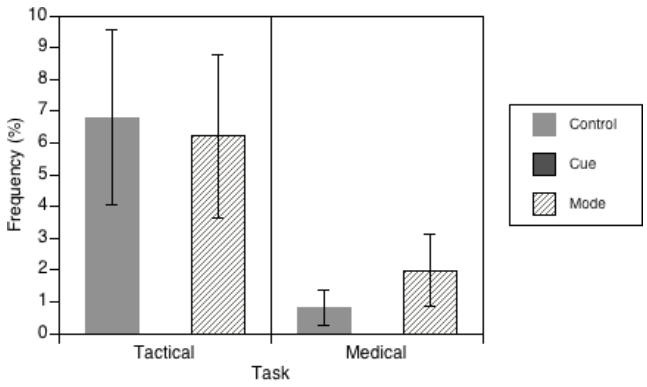


Figure 4: Postcompletion error frequency (std. error bars). Cue condition is 0% for both Tactical and Medical tasks.

Our primary measure of interest was the frequency of errors at the postcompletion step (out of seventeen trials) in both tasks. This is the step immediately following completion of the main task goal. For the Tactical task, mean postcompletion error frequencies were 6.81%, 0%, and 6.21% for the control, cued, and mode indicator conditions, respectively (Figure 4).

Analysis of variance showed the effect of intervention to be reliable, $F(2, 76) = 4.061$, $p = .021$, but not the interaction of intervention by task, $F(2, 76) = 1.86$, $p = .162$. Planned comparisons confirmed our hypothesis, as participants made significantly less errors in the cued condition versus the control, $t(76) = 3.14$, $p = .002$, and even versus the mode indicator group, $t(76) = 2.81$, $p = .006$. In comparison, the mode indicator failed to produce reliable differences with the control group, $t(76) = .263$, $p = .793$.

In the simpler Medical task, mean errors at the postcompletion step were very low: 0.82%, 0%, and 1.99% for the control, cue and mode indicator conditions, respectively. Again, none of the twelve participants in the Medical cued condition made a single postcompletion error in all seventeen of their trials. The same planned comparisons done on the Tactical task revealed no reliable differences across intervention and task.

Total Errors

The average number of total errors (out of all possible steps) was found to be higher for the Tactical task than the

Medical: 0.67 in the Tactical versus 0.28 in the simpler Medical task, $F(1, 76) = 14.60$, $p < .001$. Differences across intervention were not reliable, $F(2, 76) = 2.24$, $p = .113$, although it should be noted that the total number of errors was slightly higher for both the cue and mode indicator conditions in both tasks.

Working Memory Task

Participants showed no reliable differences in working memory task performance regardless of task $F(1, 76) = 3.47$, $p = .07$ or intervention, $F(2, 76) = 1.09$, $p = .342$.

Discussion

Our findings generally corroborate our hypothesis for the visual cue. As reported, all sixteen participants in the just-in-time cue condition of the Tactical task exhibited error-free performance at the postcompletion step on all seventeen of their trials. In contrast, the control and mode indicator groups showed mean postcompletion error frequencies between six and seven percent. Given the lack of reliable differences across intervention for overall error rates and performance on the working memory task, there seems to be no reason not to attribute the difference in postcompletion error frequency to the success of the intervention.

Nevertheless, our expectations for the mode indicator were not met. While the just-in-time cue reduced the postcompletion error mean to nil, the mode indicator had hardly any effect relative to the control. This was despite the fact that all participants were given equal training and the mode indicator was made as large, if not larger, than the flashing arrows in the cued condition. With the additional novel appearance of the crosshairs (Tactical) and display information (Medical), the state change should have been noticeable. Thus, its failure does not seem attributable to a lack of knowledge or relative visibility.

The Medical task was ineffective as a parallel of the Tactical task, perhaps primarily due to its substantially shorter length (see Table 1). It took participants nearly one quarter of the time taken to finish the Tactical task and simply failed to generate sufficient error rates to prove useful for comparing the effects of the interventions. However, it is notable that the visual cue also completely eliminated postcompletion errors in the Medical task as in the Tactical task.

Validation of MAGS

Our findings generally fell in line with the predictions of Altmann and Trafton (1999), given that the cue in the previous experiment (Chung, 2004) failed from a lack of salience. As claimed, the new ("just-in-time") cue was sufficient to prime the postcompletion step, making it unnecessary to place the postcompletion action on the critical path. Moreover, the mode indicator (a state change "at the start") did not sufficiently prime the postcompletion step that followed. It was likely "masked" by the intermediate steps or goals, as they explained.

These results support the idea of goal selection in tasks as a product of environmental cues. Hence, it follows that postcompletion errors are often generated by "false completion signals" (Reason, 2002), such as the emergence of the copy in the photocopy task. Likewise, the visual cue implemented here, in the form of two blinking arrows, was able to prime the postcompletion goal sufficiently to be correctly retrieved.

Conclusion

Several guidelines for the design of safe interfaces used in routine procedural tasks can be gleaned from this work. Interventions should be made to appear "just-in-time," as to reduce demands on memory. Asynchronous cues like the mode indicator place their own demands on memory, since there are steps intermediate to the step they are meant to prime. Even negative feedback (*after* the postcompletion step), as a downstream error cost (Chung, 2004) or as a reprimand from an "overseer" (Serig, 2001), has demonstrated no reliable reduction in the frequency of errors. Postcompletion errors cannot simply be willed away.

Also, it seems that movement and/or shape (meaning) are strong determinates of whether or not a cue is attended to (Sutcliffe, 1995). A cue (a simple red singleton) used in previous work (Chung, 2004) appeared at the same exact location as the just-in-time cue in this experiment, yet generated no significant reduction in error frequency. The mode indicator, which relied on static contextual cues, also had no reliable effect. It was made static (as in most real-world applications) since the nature of such cues is that there are intermediate steps between their onset and the step they are meant to prime. Blinking would unnecessarily attract visual attention to an inactive control.

Implementing a successful error intervention may not, however, be so simple as merely adding visual cues with these properties to the interface. Differences in task (e.g., length) and interface (e.g., background color) characteristics also attenuate the effectiveness of these cues, as demonstrated by the Medical task. Additionally, the fact that our participants had explicit training on the meaning of the cue must be considered. Simply placing blinking arrows or other novel cues on the interface would affect naïve users differently from those who had been trained.

The failure of a singleton onset (Chung, 2004) and subsequent success of two blinking arrows in this experiment may at least partially be explained by the speed at which our visual attention shifts in procedural tasks with medium to high level of skill and external pressures. Hence, while the cue used in the previous experiment also appeared in temporal conjunction with the completion of the previous step and in spatial proximity to the targeting window, it was overlooked. In contrast, the successful blinking cue continued to generate attention-capturing movement until the postcompletion step was satisfied. Moreover, it offered immediate information (arrows pointing to the correct button) about its meaning.

Hence, these findings come as further empirical evidence for cue-driven goal selection in procedural tasks. More specifically, they highlight the importance of cue timing and the visual properties of movement and meaningfulness. Follow up inquiry is underway to determine the individual strengths of these properties. Such data will be vital for any truly predictive theory of human error.

Acknowledgments

We would like to acknowledge the support of the Office of Naval Research under grant number N00014-03-1-0094. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ONR, the U.S. Government, or any other organization.

References

Altmann, E. M. & Trafton, J. G. (1999). Memory for goals: An architectural perspective. *Proceedings of the twenty-first annual conference of* the Cognitive Science Society (pp. 19-24). Hillsdale, NJ: Erlbaum.

Byrne, M. D., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science. 21(1),* 31-61.

Chung, P. H. (2004). Visual cues to reduce error in computer-based procedural tasks. Masters thesis, Rice University, Houston, TX.

Gray, W. D. (2000). The nature and processing of errors in interactive behavior. *Cognitive Science. 24(2),* 205-248.

Monk, A. (1986). Mode errors: a user-centered analysis and some preventative measures using keying-contingent sound. *International Journal of Man-Machine Studies, 24,* 313-327.

Norman, D. A. (1988). The Psychology of Everyday Things. New York: Basic Books.

Polson, P. G., & Lewis, C. H. (1990). Theory-based design for easily learned interfaces. *Human-Computer Interaction, 5,* 191-220.

Rasmussen, J. (1982). Human Errors: A taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents, 4,* 311-335.

Reason, J. (1990). Human Error. Cambridge, UK: Cambridge University Press.

Reason, J. (2002). Combating omission errors through task analysis and good reminders. *Quality and Safety in Healthcare, 11,* 40-44.

Remington, R.W., Johnston, J.C., & Yantis, S. (1992). Involuntary attentional capture by abrupt onsets. *Perception & Psychophysics, 51*, 279-290.

Rieman, J., Young, R. M., & Howes, A. (1996). A dual-space model of iteratively deepening exploratory learning. *International Journal of Human-Computer Studies, 44,* 743-775.

Serig, E. M. (2001). Evaluating Organizational Response to a Cognitive Problem: A Human Factors Approach. Doctoral dissertation, Rice University, Houston, TX.

Sutcliffe, A.G. (1995). Human-Computer Interface Design. London: Macmillan Press Ltd.

Young, R. M. (1994). The unselected window scenario: analysis based on the SOAR cognitive architecture. *Proceedings of the Computer Human Interaction (CHI) Conference*, 1994.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18,* 87-122.

Imagistic Processes in Analogical Reasoning: Conserving Transformations and Dual Simulations

John J. Clement (clement@srri.umass.edu)
Scientific Reasoning Research Institute
College of Natural Sciences and Mathematics
and School of Education
Lederle GRT 434
University of Massachusetts
Amherst, MA 01003 USA

Abstract

The classical theory of analogical reasoning focuses on mappings between discrete symbols as the mechanism of analogy evaluation and transfer. This paper introduces several other analogy evaluation strategies discovered in expert reasoning protocols: bridging analogies, conserving transformations, dual simulations used to detect perceptual-motor similarity, and overlay simulations. These findings provide evidence for the hypothesis that certain analogical reasoning processes can be imagery based.

Earlier work on higher order reasoning has indicated that expert subjects use various methods to generate analogies spontaneously when solving difficult problems (Clement, 1988), and that evaluating the validity of such analogies is essential to using them (Clement, 1989). That is, even if one has generated a confidently understood analogous case, one must evaluate one's confidence in the validity of the analogy relation to have confidence in transferring results to the target. The classical theory of analogical reasoning (Gentner, 1983; Holyoke and Thagard, 1989; Forbus, et al, 1997) focuses on mappings between discrete symbols as the mechanism of analogy evaluation and transfer. This paper examines several other analogy evaluation strategies observed in expert think aloud protocols. The data base for the study comes from professors and advanced graduate students in scientific fields who were asked to think aloud about a variety of problems. This paper focuses on two mathematicians solving physics problems they found difficult. By focusing on problems with which they were unfamiliar (i.e., a problem on the frontier of their own personal knowledge). it is plausible that the thought processes analyzed will share some characteristics with hypothesis formation and model construction processes used on the frontiers of science.

An example of a problem where analogy evaluation is important is the "Sisyphus problem" in Figure 1A: "You are given the task of rolling a heavy wheel up a hill. Does it take more, less, or the same amount of force to roll the wheel when you push at x, rather than at y? Assume that you apply a force parallel to the slope at one of the two points shown, and that there are no problems with positioning or gripping the wheel. Assume that the wheel can be rolled without slipping by pushing it at either point."

One expert subject proposed the analogy that the wheel acts like a heavy lever perpendicular to the slope, with its fulcrum at the point of contact. Intuitively, the lever would be easier to move by pushing at X, suggesting that the same would be true for the wheel. But in the wheel the point of contact is moving, and ordinarily lever fulcrums do not move. In addition some subjects assume that the fulcrum should instead be at the wheel's center. Therefore the evaluation of the validity of the analogy relation (shown as the dotted line between A and B in Figure 1) was in question. This is distinguished from the subject's confidence in his understanding of the analogous case B itself, which was quite high in this case.

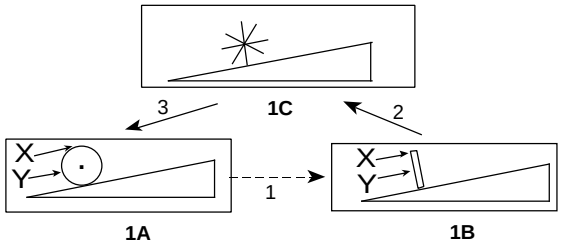


Figure 1: S2's lever analogy for the wheel

Bridging Analogies

One method for evaluating this analogy used by this subject was the bridging analogy shown in Figure 1C of a spoked wheel without a rim. By breaking the problem of confirming a "farther" analogy into the problem of confirming two "closer" analogies. such a bridge can make it easier to develop confidence that the wheel does work like the lever in Figure 1B (a correct analysis). Bridging analogies are defined as occurring when the subject finds or generates an intermediate case which shares features with both the target and source analog. Their value has been documented previously in a number of expert problem contexts and in instructional applications (Clement, 1986). While it can be very helpful to subjects, bridging in itself is an incomplete strategy for analogy evaluation, since each half of the bridge must itself be evaluated. Therefore bridging is most useful in conjunction with other evaluation methods and it adds to, rather than reduces, the number of tasks to be performed.

This raises the problem of why experts bother to consider bridging cases at all, since they seem to create more work.

Conserving Transformations

In this section I present examples of a second evaluation strategy called conserving transformations and argue that it is distinctly different from the commonly cited method of matching discrete features for evaluating an analogical relationship. A paradigmatic case of a conserving transformation (although he did not identify it as such) is Wertheimer's method for determining the area of a parallelogram by cutting one end off and moving it to the other end to form a rectangle. A transformation is an action that changes a system 1 to system 2. If 2 is the same as 1 with respect to a feature or relationship R, then the transformation conserves R. An example of a conserving transformation in the Sisyphus problem occurred when subject S7 changed the problem to an analogous one involving an almost-vertical cliff with gear teeth:

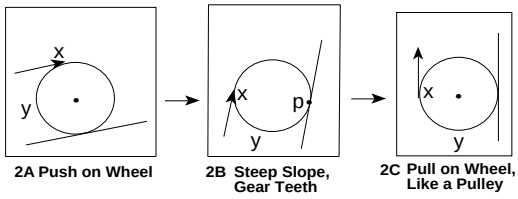


Figure 2: Wheel analogy series of S7

(Brackets in transcripts denote interpretations from viewing tape, while parentheses denote observed actions.)

01 S: "Suppose it were tilted steeply and you did that; so steep as to be almost vertical. (Draws Figure 2B). It seems like it [the wheel] would skid out from under you the other way [down along the cliff]. This (moves hands as if turning an object clockwise) would get away from you here [at point p]. Let's assume it's gear toothed [gear teeth on the wheel and the cliff] and that it won't slip. "

The change from situation A to B in Figure 2 appears to be a double transformation consisting of: the change of slope, and the addition of gear teeth. One can define the "targeted relationship" as the one for which an explanation or prediction is sought in the target situation (e.g. the relation between the force required and its location on the wheel). In his further work on the problem S7 never questions the validity of these transformations, and assumes that the targeted relationship in the problem situation is not affected by them. One can surmise that this occurs because the gear teeth transformation is a standardized one in physics and both are intuited to be irrelevant to the relationship of interest in the problem, i.e. they are conserving transformations. The origins of this kind of intuition have been studied since Piaget's early conservation experiments but are still poorly understood. (Case 2C will be discussed later.)

The hand motions over the drawing here provide one source of evidence on the use of dynamic imagery. Although the drawing can be an external support for a static visual representation, it does not depict movements, so it is reasonable to hypothesize that the subject is performing a mental imagistic simulation of the wheel slipping down on the cliff. The change in slope simplifies the problem by changing it to one in which forces act mostly along only one dimension.: upward and downward. Since the problem already specified no slipping, the gear teeth do not add new information but may help in imagistically simulating what will happen in the analogous case. Thus they may be an example of what I have called an "imagery enhancement" strategy (Clement, 1994, 2003).

The transformations appear in this case to be a means of both generating and evaluating the new analogy. Clement (1988) found that of a collection of 31 spontaneous analogies generated by ten experts, a greater number of analogies were generated via such transformations than those generated via an association to another case already in memory. However, the present paper focuses on the possible analogy evaluation function of transformations rather than on their analogy generation function.

Spring problem. A more substantial transformation is illustrated by the passage below from S2's solution to the following spring problem: "A weight is hung on a spring. The original spring is replaced with a spring made of the same kind of wire, with the same number of coils, but with coils that are twice as wide in diameter. Will the spring stretch from its natural length more, less, or the same amount under the same weight? (Assume the mass of the spring is negligible.) Why do you think so?" Earlier this subject has considered long and short horizontal bending rods made of the same wire as the spring and bent by hanging the same amount of weight on one end as an analogy for the spring problem (Figure 3B). Knowing that the longer rod will bend more suggests to him that the wider spring stretches more. In the following passage he evaluates that analogy by speaking of rolling up the bending rod into a spring (Figure 3C is discussed later):

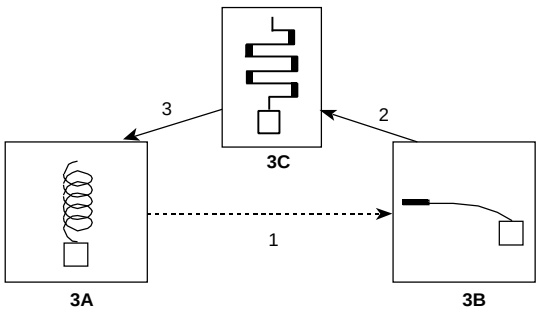


Figure 3: Rod analogy and zigzag bridge of S2

102 S: "You can imagine a spring...and you know...there's no difference between the top and the bottom. It's a symmetric situation…

105 S: You take your [straight, horizontal] wire, you say 'OK, you think it's the bending that does it. Well, then

let's bend it [by pulling down on one end of the straight wire]. And then let's roll it up [around a vertical axis] to make the spring. And you get a spring which stretches more and more at the bottom. The loops are wider apart!

107 S: Stretch it [a normal spring]: you don't get this increase of the distance between the loops toward the bottom. You just get a uniform stretching. And therefore the stretched spring cannot be understood as a rolled up bent spring."

This argues against bending as the source of stretching. Here the subject describes a very explicit spatial transformation between the spring and the rod. The sequence is: he generates the rod analogy; he simulates bending in the rod; he evaluates the analogy by transforming it back into a spring; there is a conflict with a known property of the spring, and he discounts the rod analogy. This evaluation is extremely valuable in that it gives him information arguing that the conjectured mechanism of bending is invalid. (In fact springs stretch primarily via twisting in the wire, not bending.) Note the imagery report in line 102. These passages suggest the attempt to use a visual transformation to evaluate the validity of a tentative analogy. The evaluation is influential in that it leads to discounting the validity of the analogy. Griffith, Nersessian, and A. Goel (2000) have also designed and investigated a computer program which successfully accounts for a number of features of this protocol and others collected for the spring problem. Transformations played an important role in modifying and improving faulty analogies or models in their program. However, they did not examine the role of conserving transformations as a means for analogy evaluation.

Prior to these sequences the subject had generated not only the analogous rod case, but what appeared to be a complete mapping of symbolic features between the rod case and the spring case. Bending, length, and slope, in the rod were mapped onto stretching, width, and slope in the spring. The relation of <greater length causes greater bending in the rod> had been mapped to the sought-after relation of <greater width causes greater stretch in the spring>. Therefore the transformations above do not appear to be adding any new elements to the mappings. Rather, they seem to be increasing the subject's confidence that he has found an important visual mismatch in the slope feature. They are new ways to *arrive at* the same mappings. That is, the transformations are a *means* to determining a match or a mismatch as the outcome, not just the notation for a mismatch as read off from two different lists. The notion that the transformation should be conserving is quite plausible. If the main mechanism is bending, this "winding up" transformation is locally perpendicular to the bending, therefore it could very well be a conserving transformation. Instead of transferring the "result" from the base to the target by using an explicit set of correspondences, in the present model this can be simply "read off" from an image derived from the imagistic results simulated in the base being transformed back to the target. Thus the conserving transformation strategy is a process that can work independently from an explicit feature matching process.

A traditional approach to analogy evaluation focuses on determining that multiple similarities between the base and target are sufficiently important. In contrast, a conserving transformation strategy need only focus on determining that a single transformation from base to target is sufficiently unimportant (irrelevant to the targeted relationship). This may mean that confirmation of an analogy via a conserving transformation can require considerably less work than confirmation via mapping.

Dual Simulation

Case 1. There is evidence in the protocols for a very direct strategy for analogy relation evaluation termed "dual simulation". A brief example that hints at this possibility follows where S2 says:

(Line 23) "Surely you could coil a spring in squares, let's say, and it would behave more or less the same".

There is not very much data in this statement, but it is plausible that the subject created an image of a square spring, simulated the effect of hanging a weight on it, and found this to be similar to the image of hanging a weight on a normal spring. However, the resolution of the perceived similarity appears to be at a low level of detail.

It is doubtful that his conclusion here is from "looking up a fact in memory", because of the novelty of the square case. (Later simulations by S2 with the square coil lead to imagining one side acting like a wrench to twist the next side. This produces an Aha episode with the insight that torsion is a major mechanism of stretching in the spring, and predicts correctly that the wider spring will stretch more, but that is the topic of another study (Clement, 1989)).

Dual simulation depends upon the process of **imagistic simulation** discussed in Clement (2003). That article found evidence for such an internal process from several observation categories for external behavior: **personal action projections** (spontaneously redescribing a system action in terms of a human action, consistent with the use of kinesthetic imagery), **depictive hand motions**, and **imagery reports**. The latter occurs when a subject spontaneously uses terms like "imagining," "picturing," a situation, or "feeling what it's like to manipulate" a situation. In several of the present cases one sees **dynamic imagery reports** (involving movement or forces). None of these observations are infallible indicators on their own, but as multiple instances accumulate, they can be taken as evidence for imagery. Taken together with the subject's new predictions, the observations above can be explained via *imagistic simulations* wherein a somewhat general perceptual motor schema assimilates the image of a particular object and produces expectations about its behavior in a subsequent dynamic image, or simulation.

The process of **dual simulation** can be summarized as follows. Imagistic simulations of the target and the analogous case are each run in as much detail as possible. The dynamic images of the behavior of each system are then

compared; and they may be inspected for certain aspects. If their behavior "appears" to be the same, the analogy relation receives some support, depending on the level of certainty in the comparison.

Case 2. More data is present in the following episode of an analogy to a two-dimensional spring made of zig-zagging wire that lies in a single vertical plane, shown in Figure 3C.

> 23 S2: "I wonder if I can make the spring..which is a 2 dimensional spring..but where the action ..isn't at the angles...it's distributed along the length… I have a visualization... Here's a .. a bendable bar, and then we have a rigid connector...(draws more bars connected in a zig zag, two dimensional shape). And when we do this what bends...is the bendable bars...and that would behave like a spring. I can imagine that it would.... it would stretch, and you let it go and it bounces up and down. It does all the things."

Here the conjunction of the dynamic imagery reports and the comparison of the two systems gives more support to the hypothesis that a dual simulation is occurring to compare the target and the zig zag cases. The dual simulation appears to establish the analogous case as being relevant and plausibly analogous in that its behavior is similar, at least at a gross level of qualitative behavior, to the target. But this does not tell the subject whether the two systems exhibit the same relationship between width and stretch. Thus in the above cases dual simulation appeared to serve only as a check on the initial plausibility of the analogy.

One then needs to be clear that dual simulation as an analogy evaluation strategy does not necessarily mean confidently simulating the targeted relationship in both base and target. In that case there would be no need for an analogy because the target could have been directly simulated on its own. However, the examples presented indicate that dual simulation can still help one determine whether the target and base are similar with respect to other important behaviors, thereby increasing one's confidence that the analogy is sound (or eliminating the analogy from consideration).

Overlay Simulation

Lever case: There is evidence for the existence of a more precise type of dual simulation that I term "overlay simulation" where the image of one simulation takes place "on top of" a second image. Although I have separated them in Figure 1 for clarity, S2 actually drew his lever analogy (Figure 1B) directly on top of the wheel (Figure 1A) and compared the movement of the wheel and the lever. This meant that the arrow symbolizing the application of a force by pointing to the top of the wheel was also pointing to the top of the lever. When two separate systems are represented as overlapping in the same external diagram with salient features aligned I term this an *overlay diagram*. This supports the interpretation that internal dynamic images of the two systems and their actions were overlapping in the same way. I call this hypothesized internal aspect here an *overlay simulation* as a special type of dual simulation. Presumably the alignment of key features made it easier for him to compare the expected movements and resistances of the wheel and the lever as he simulated each of them.

Spokes case: Overlay simulation may also be responsible for the power of S2's "spoked wheel without a rim" bridging analogy shown in Figure 1C. For the spoke that is touching the ground, the spoke can be seen as a lever with its fulcrum at the ground. This means that the entire wheel of spokes can be seen at any one time as equivalent to a single lever, supporting the analogy on the right hand side of the bridge BC in Figure 1. This subject spoke of a tireless, rimless wheel. Again this is shown separately in figure 1C for clarity, but in fact the spokes were inscribed within the rim of a circular wheel in the subject's drawing. So on the other side of the bridge AC, the spokes are envisioned at the same size as the original wheel, and this may make it easy to sense via dual simulation that they behave in the same way as the wheel when a force is applied. In particular, the way the rimless spoked wheel "rocks" on each spoke over a short distance can be seen as similar to the way the original wheel rolls. That is, it appears, especially with many spokes, to have the same kind of motion in a mental simulation and therefore be amenable to the same type of analysis with respect to the causes of motion. Although such arguments must be bolstered mathematically to make them rigorous, as a form of heuristic reasoning, this type of qualitative argument can be quite compelling.

Pulley case: As a third example the case of the pulley analogy in Figure 2C was also used by S7 in the Sisyphus problem. He believes that perhaps the push needed at X on the wheel is smaller than at Y, similar to a pulley where the force applied to the end of the rope need only be half of the weight of the wheel. As part of an attempt to evaluate that analogy, S7 speaks and gestures as if alternating between seeing the same drawing (Figure 2C) as a wheel and a pulley, referring to it differently as one or the other in alternate fashion. Continuing from segment 01 above:

> 05 S7: What it feels like is the weight of it [wheel in Figure 2B]-; is pretty close to parallel with what you've got if you go roll it with a complete vertical. It now begins to feel like a pulley...(Draws Figure 2C) What the vertical is over here no longer matters perhaps but we'll say it's er, gear toothed again.
>
> 06 S: ...And you're over here pulling like this [at x]. That feels like you're on the outside of a pulley pulling up.
>
> 07 S: And since you say it doesn't slip, then this thing over here (points to line in upper right of Figure 2C and adds upward pointing arrowhead to it) must be providing the other half of it, something it feels, in which case it's a classic pulley; no, it can't be classic pulley. But it's, like a classic pulley in which now you only need half of the force. If the weight of the thing is 10 lbs. here, it feels like 5 would work here (writes 5 on upper left of C) and 5

over here (writes 5 on upper right) as though it were a pulley… So let's imagine it is a pulley.

08 S: [In] this new point of view, it feels like working at X [on the edge of the wheel] is better [than at the center].

The personal action projections and alternating references to both the wheel and the pulley systems while staring at the same diagram 2C provide initial evidence for an overlay simulation here that compares the system of rolling the wheel straight up a vertical cliff to the pulley system. Presumably it is easy in an overlay simulation to switch rapidly between simulations of the two cases, (which happens at least five times above). Again, although the imagery is probably assisted in this case by the drawings, the drawing cannot be providing perceptions of forces or motions involved, and so I hypothesize that these are imagined via imagistic simulations. Some evidence for kinesthetic imagery is indicated by personal action projection phrases like "feels like you're on the outside of a pulley pulling up" and "you're over here pulling" in the transcript, and such imagery is clearly not already enacted in the static drawings.

Later he expresses some reservations about the pulley analogy however: "This rope wrapping around here..doesn't feel to me necessarily like...pushing (moves hand . to r.) on the outside of a wheel." But in the passage below he appears to reevaluate the analogy positively by (1) generating a bridging analogy; and (2) using overlay simulations by simulating different systems in alternating fashion using the original wheel drawing. Therefore this final example is more complicated because it combines these two strategies.

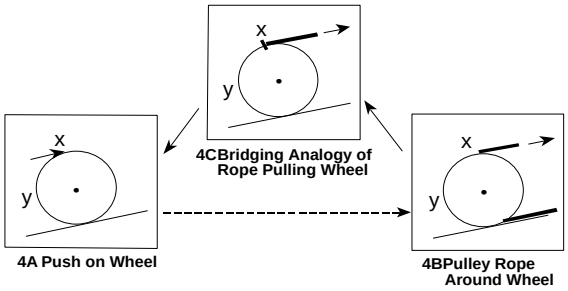


Figure 4: Second analogy series of S7

162 S7: (Looking at Figure 4A) I've got my full (holds both hands out as if pulling a rope and shakes them slightly) power available- and where would I apply that? My instinct tells me [it is easier to apply force at] X again but that er, but again it's in terms of a pull and not a push. I'd have to get a grip.(closes eyes) Assuming that's not a problem, then pulling should be the same as pushing.. Seems clear that- (silently holds both hands out as if pulling a rope for 4 sec.)...So we attach a rope to one of the teeth [as in 4C but staring at the same Figure 4A], now it becomes more like the pulley problem (holds r. hand out as if pulling a rope for 3 sec)…the teeth at the bottom are playing the role of-; the pulley doesn't look so bad after all. And you hang on for all you're worth up there, to keep it from rolling; to keep it balanced.

Figure 4C shows how a rope attached to the edge of the wheel at X can be seen as an intermediate bridging case between the original problem and the pulley case in 4B. Although I have drawn three cases in Figure 4 for clarity, in fact S7 used only Figure 4A while talking about the three cases: the pushed wheel, the pulley, and the rope attached to the tooth at X on the wheel. One can hypothesize that the internal overlay simulations create a context whereby the alignment between trajectories and forces in imagistic simulations of different cases, as well as the evaluation of the validity of the analogies between the cases, can be more easily made.

163 S7: Seems a lot easier than getting down here behind it [at "Y" in Figure 4A] and pushing. Why? because of that coupling pulley effect. It seems like it would be a lot easier to hold it here [at "X"] for a few minutes (Holds hands in "pulling" position) than it would be to get behind it… yeah, my confidence here is much higher now, that it's right… [easier to push at X] And so the pull--it just felt right with the pulley feeling. Now pushing (lays extended finger on paper pointing up slope at X in Figure 4A and moves it toward X) uh,.. it's got to be the same problem…

178 I: Do you have a sense of where your increased confidence is coming from?

179 S: It's the pulley analogy starting to feel right.

The subject's thinking here appears to determine whether the forces on the edge of the wheel and on the rope from the pulley "feel" the same as he performs an imagistic simulation of each case. The bridging case in Figure 4C of a rope tied to the wheel at point X appears to serve the purpose of setting up two pairs of cases (base:bridge B:C and bridge:target C:A) that are "closer" to each other than AB. In other words the bridging case creates two analogy pairs that are more perceptually similar. This may be an important advantage if the evaluation of each pair is being done via a dual simulation of the cases. This provides one answer to the earlier question of why bridging can be useful to a subject even though it seems to add more work in creating additional analogy relations.

In fact the underlined references to feeling forces, personal action projections, and hand motions in the above passages provide evidence for the involvement of kinesthetic imagery and for dual imagistic simulations when he is comparing two cases. One can hypothesize that Figure 4A is acting as an overlay diagram for an overlay simulation. The use of an overlay diagram and references that the wheel problem solution "felt right with the pulley feeling." supports the hypothesis that dual simulations are being used to evaluate these analogies. Thus this last example illustrates the combined use of overlay simulation (as a special kind of dual simulation) and bridging as analogy evaluation strategies.

Conclusion

In summary, rather than a single process for mapping elements in a discrete symbolic representation, a number of

additional processes for evaluating an analogy relation have been identified, namely: bridging analogies, conserving transformations, dual simulations to detect dynamic similarity, and overlay simulations. Roughly, conserving transformations work by allowing the subject to detect the causal, perceptual motor irrelevance to a targeted relationship, of making a transformation on a case. Dual simulations work by allowing the subject to detect a causal, perceptual motor similarity between base and target. Overlay simulations are a special type of dual simulation in which the image of one case is overlaid and aligned on top of the other case to make comparisons more precise. An intermediate bridging case is a higher order strategy that can facilitate making one of the above processes easier to perform. The relationship of these strategies to discrete feature mappings is still unclear, but when subjects can articulate such mappings, that may add another important kind of precision to the process of analogy evaluation.

Implications. These findings add to previous evidence (Casakin and Goldschmidt 1999; Clement 1994, 2003: Craig, Nersessian and Catrambone, 2002; Croft and Thagard, 2002; Trickett and Trafton, 2002) for formulating the general hypothesis that many analogical reasoning processes can be imagery based. Also, the wheel problem transcript provides evidence that imagery and runnability are transferred from base to target. Clement (2003) extended this theme by examining evidence for the transfer of imagery and runnability from source analogues to explanatory models and hypothesized that this may be an important source of model flexibility, providing an argument for the importance of such processes. The importance of bridging analogies as an instructional technique has been documented previously (Clement, 1989), and the same may very well be true for conserving transformations (Wertheimer, 1959), and overlay simulations/animations. Thus much work remains to be done in this area.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants RED-9453084 and REC-0231808. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Casakin, H, & Goldschmidt, G. (1999). Expertise and the use of visual analogy: Implications for design education. Design Studies, 20:153--175.

Clement, J. (1986). Methods used to evaluate the validity of hypothesized analogies. *Proceedings of the Ninth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Clement, J. (1988). Observed methods for generating analogies in scientific problem solving. *Cognitive Science, 12*: 563-586.

Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. Glover, J., Ronning, R., and Reynolds, C. (Eds.), *Handbook of creativity: Assessment, theory and research*. NY: Plenum, 341-381.

Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving. In Tirosh, D. (Eds.), *Implicit and explicit knowledge.* Norwood, NJ: Ablex Publishing Corp.

Clement, J. (2002). Protocol evidence on thought experiments used by experts. In Wayne Gray and Christian Schunn, Eds., *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society 22*, 32. Mahwah, NJ: Erlbaum.

Clement J. (2003). Imagistic simulation in scientific model construction. In Proceedings of the *Twenty-Fifth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Craig, D. L., Nersessian, N. J., & Catrambone, R. (2002). Perceptual simulation in analogical problem solving. In: *Model-Based Reasoning: Science, Technology, & Values*. (pp. 167--191). Kluwer Academic / Plenum Publishers, New York.

Croft, D., & Thagard, P. (2002). Dynamic imagery: A computational model of motion and visual analogy. In L. Magnani and N. Nersessian (Eds.), *Model-based reasoning: Science, technology, values.* New York: Kluwer/Plenum, 259-274.

Forbus, K., Gentner, D., Everett, J., and Wu, M. 1997. Towards a computational model of evaluating and using analogical inferences. *Proceedings of the 19 th Annual Conference of the Cognitive Science Society*, pp 229-234, LEA, Inc.

Gentner, D. (1983). Structure-mapping a theoretical framework for analogy. *Cognitive science*, 7, 155-170.

Griffith, T. W., N. J. Nersessian, and A. Goel (2000). Function-follows-form transformations in scientific problem solving. In *Proceedings of the Cognitive Science Society* 22, 196-201. Mahwah, N.J.: Lawrence Erlbaum

Hegarty, M. (2002). Mental visualizations and external visualizations. In Wayne Gray and Christian Schunn, Eds., *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* 22, 40. Mahwah, NJ: Erlbaum.

Holyoak, K. J., & Thagard, P. R. (1989). A computational model of analogical problem solving. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 242-266). New-York: Cambridge University Press.

Trickett, S. and Trafton, J. G. (2002) The instantiation and use of conceptual simulations in evaluating hypotheses: Movies-in-the-mind in scientific reasoning. In Wayne Gray and Christian Schunn, Eds., *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* 22, 878-883. Mahwah, NJ: Erlbaum.

Wertheimer, M. (1959). *Productive thinking*. New York: Harper & Row.

Dumb mechanisms make smart concepts

Eliana Colunga (colunga@psych.colorado.edu)
Department of Psychology, 345UCB
Boulder, CO 80309 USA

Linda B. Smith (smith4@indiana.edu)
Department of Psychology, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

There is an ongoing debate on the nature of the processes and knowledge involved in learning language. On one side of the debate, people argue that children learn words through deliberative processes that use propositional conceptual knowledge; on the opposing side, people argue that children learn words through automatic processes and knowledge based on learned associations among perceptual features. In this paper we concentrate on the Animate/Inanimate distinction as evidenced in children's novel noun generalizations. The results of two experiments with 3-year-olds and adults suggest that 1) automatic processing guides children's generalizations of novel nouns and 2) "conceptual" knowledge may be formed as a web of learned correlations.

Background

The nature of the processes and knowledge involved in children's learning of new nouns is a highly contentious issue. Children generalize names for things in appropriate ways depending on the kind of thing being named. For example they generalize names for artifacts by their shape; names for non-solid substances by their material; and names for animates by their shape and texture. The debate centers on the nature of the processes and knowledge involved in this behavior. On one side of the debate, people argue that children reason about category membership using slow, deliberative, conscious processes and propositional beliefs about categories and category structure (Gelman & Markman, 1987; Keil, 1994; Kemler-Nelson, Russell, Duke & Jones, 2000). On the opposite side, people support fast, unconscious, automatic processes and knowledge in the form of correlations among perceptual features (Smith, 1995; Smith, 2000; Smith, Colunga & Yoshida, 2003). Sometimes this debate has been framed in terms of whether children's early word learning is "smart" (reflective, conceptual) or "dumb" (built from more basic general and automatic processes). In this paper we concentrate on the Animate/Inanimate distinction as shown in children's novel noun generalizations. We show evidence that automatic processing guides children's generalizations of novel nouns, and suggest that the "conceptual" knowledge that enters word learning, even in adults, may be me made out of correlations of perceptual features.

One widely used task to study children's word learning biases is the novel noun generalization task. In this generalization task, children are shown an exemplar like those in Figure 1. They are told its name, and then asked what other things have the same name. When shown an exemplar like that in Figure 1a, with cues indicating it is a depiction of an animate thing, children systematically generalize its name only to new instances that match in both shape and texture, but not to things that match in shape only or texture only. When shown an entity without such features, children systematically generalize the name to new instances that match in shape, whether they match in other properties or not, as shown in Figure 1b. Jones & Smith, (1991) and others have suggested that children learn correlations between features and category structure -- between having eyes and being in a category organized by shape and texture, and between being angular (and without animacy features) and being in a category organized by shape. Consistent with theories and evidence on attentional learning, they suggest that in the novel noun generalization task, these features automatically increase attention to relevant properties, enabling children to attend to the right similarities according to the kind of the object at hand. This correlational learning account has been supported by showing that the requisite correlations between features and category organization exist across the first 300 nouns that children learn (Samuelson & Smith, 1999).

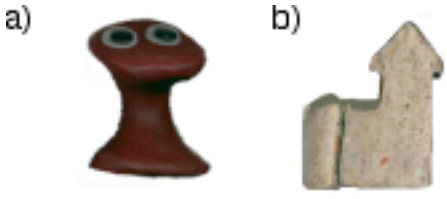


Figure 1. Examples of an exemplar with cues indicating it is a depiction of an animate thing, and an exemplar with artifact cues.

Recently, Booth and Waxman (2002) provided support for the alternative "smart" interpretation of children's performance in this task. Booth and Waxman presented exemplars in a context that construed them as animate or inanimate. The disambiguating context consisted of brief stories in which the experimenter gave the child information about the exemplar. For example, in the animate condition,

the experimenter introduced an exemplar as a Teema and explained that the Teema has a mommy and a daddy that love it very much and gave it hugs and kisses. Similarly, in the inanimate condition, the experimenters showed the same exemplar but this time introduced it as a Teema which is used by astronauts in their trips and will be replaced if it breaks. Their results showed that 3-year-olds can use the information in the stories to guide their generalizations of the novel noun – in the Animate condition, children generalized the novel name for the exemplar to other objects matching in shape and texture, but in the Artifact condition they generalized the novel name to any object that matched the exemplar on shape. Booth and Waxman take this result to mean that children word learning is "smart", a result of deliberative processes operating on conceptual knowledge in the form of an unitary concept of animacy. Their account is illustrated in Figure 2a.

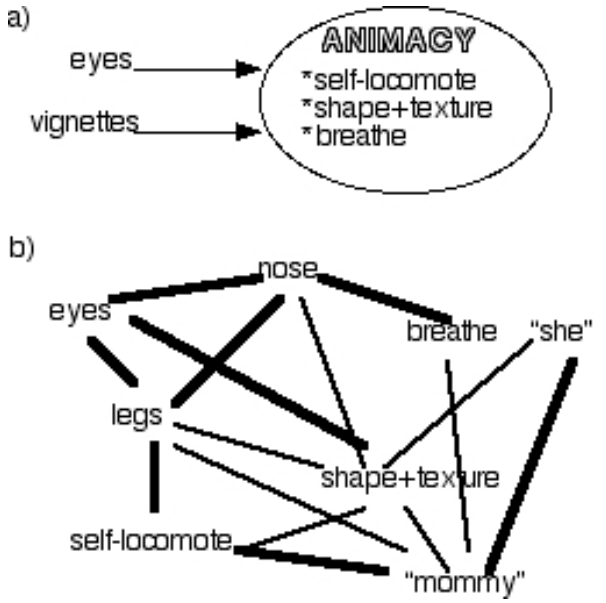


Figure 2. In Booth & Waxman's account, eyes, like vignettes, serve as a gateways to children's concept of animacy. In the correlational learning account, eyes and words like "she" or "mommy" are part of a web of correlations that is learned from the regularities that exist in animate categories in the world.

According to Booth and Waxman, the stories affect children's extension of novel nouns because they provide conceptual information relevant to the exemplar's ontological kind. That is, the Animate story identifies the exemplar as an animate, and the Artifact story identifies the exemplar as an artifact. Once children know the ontological kind of the exemplar, they have access to all the knowledge regarding that ontological kind, including the fact that shape and texture are central features for Animates and shape is the central feature for Artifacts.

The correlational learning account can also explain Booth and Waxman's findings. Novel nouns are extended on the basis of learned correlations among perceptible properties. Having eyes correlates with having a mouth, correlates with being called "he" or "she", correlates with animate-like motion patterns, correlates with attention to shape and texture. This web of correlations is the "knowledge" used in the novel noun generalization task through automatic processes, rather than through deliberative reasoning. Importantly, by this account any one cue can activate any other part of the web, *depending on the degree to which they have been correlated in the learner's experience*. Under this view, Booth and Waxman's results can be explained as a consequence of learned correlations among perceptible properties, including words. The vignettes shift children's attention because they are made out of words – words that correlate with categories of animates that are organized by shape and texture (like zebra, or snake) and words that correlate with categories of artifacts that are organized by shape (like hammer, or cup).

If the correlational story is correct, these words should cue attention to shape and texture in the Animate condition and attention to shape in the Artifact condition automatically, and without being strung together to form a coherent story *about* the exemplar. A strong test of this prediction would be to prime the children with animate-correlating or artifact-correlating words presented merely as a list and measure how this affects their performance in the novel noun generalization task. Under Booth and Waxman's explanation, reading children these words in a list should not have an effect since they are not put together into a coherent story that presents conceptual information regarding ontological status, nor are they presented as referring in any way to the exemplar. Thus, if children's novel noun generalizations are shifted by this priming, this attentional shift will not be attributable to any kind of deliberative process that reasons about the ontological status of the exemplar using the information given by the experimenter. Experiment 1 tests this prediction.

Experiment 1

Methods

Participants. 24 3-year-old children with a mean age of 42.6 months (range: 38.4-45-5 months) participated.

Materials. The stimuli consisted of two sets of eight abstract objects that could be construed as either animate or inanimate. Each set consisted of an exemplar object and 7 test objects that matched the exemplar in none, one, or more features of shape, texture or color. Figure 3 shows the diagnostic items, the ID test object matched the exemplar in all shape, color and texture features (but was different in size), there were also test objects that matched the exemplar in shape only, color only, texture only, shape + texture, shape + color, color + texture, and none of these features.

Two lists of words, one of animate-correlating words, the other of artifact-correlating words, were selected from the Animate and Artifact vignettes in Booth and Waxman (2002). Table 1 shows the words used.

Table 1: List of animate-correlating and artifact-correlating words selected from Booth & Waxman (2002).

Animate	Artifact
mommy	take
daddy	worn
love	break
sleep	make
hugs	bought
kisses	use
hungry	
walking	
gobbled	
happy	

Procedure. Children were randomly assigned to either the Animate or Artifact condition. During the Priming Phase, the child and the parent were asked to repeat the words said by the experimenter. The experimenter then went through the corresponding list saying each word once and letting parent and child repeat it. None of the test objects or exemplars were in view during the Priming phase. The Priming Phase was followed by the Testing Phase. The list of words was put away and children were informed that now they were going to play a different game and they were introduced to one of the exemplars "This is a Teema." and then asked for each of the test objects in that set "Is this a Teema?". Each of the test objects were presented twice in one of two previously generated random orders. Then the second set was presented, preceded by a second Priming Phase, for a total of 28 trials. The order of the sets was counterbalanced across conditions.

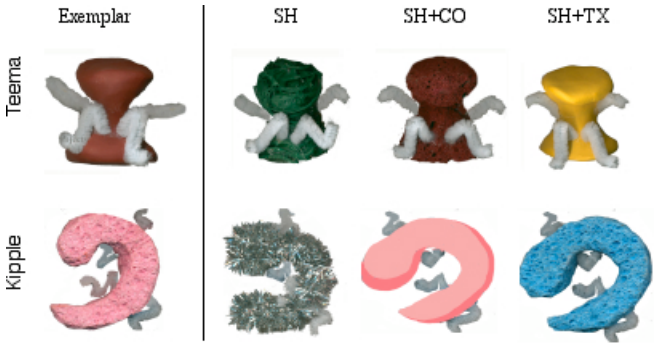


Figure 3. The two exemplars and some of their corresponding test items.

Results

The number of "yes" responses (the name applies) was submitted to a 2(Priming List)x7(Test Item) repeated measures ANOVA. The analysis revealed a main effect of Priming List ($F_{(1,22)}$ = 5.38, P < .05). That is, children overall said "yes" more when primed with the Artifact list of words than when primed with the Animate list. There was also a main effect of Test Item ($F_{(1,22)}$ = 37.21, P < .001). No interactions were significant.

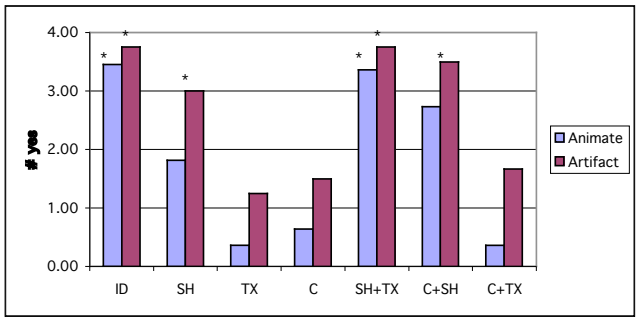


Figure 4. Results for Experiment 1.

Furthermore, post-hoc analysis revealed different patterns of noun extension in the two conditions. As shown in Figure 4, children in the Animate condition extend the name to the ID test object and to the shape+texture match, whereas children in the Artifact condition extend the name to the ID test object, the shape-only, the shape+texture and the shape+color matches. We counted the number of children who said "yes" more than expected by chance on each condition. On the critical test item, the shape-only match, there was a significant difference on the number of children who said "yes" more than expected by chance in the Animate versus the Artifact condition ($\chi^2(1,N=24)$ = 4.19, P <0.05). That is, children in the Artifact condition extended the name of the exemplar to all the test objects that matched it in shape, regardless of their size, color or texture, but children in the Animate condition extended the name of the exemplar more conservatively, only to those test objects that matched it in both shape and texture.

Discussion

At the very least, the results of Experiment 1 showed where the Animacy/Artifact information in Booth and Waxman's vignettes comes from. Just hearing these words, without weaving them into a coherent story, was enough to shift children's attention to features typical of animate vs. artifact categories – shape only for artifacts, shape+texture for animates. But more importantly, children's novel noun generalizations were influenced by these words in a priming paradigm, without these words being heard at the same time that the exemplar or test objects were present. Apparently, these words automatically activate the attentional biases with which they are associated, a fact consistent with well-supported ideas about memory processes (). The result thus supports the idea that children's novel noun generalizations depend on automatic processes that operate on learned correlations.

But is it possible that the concept of animacy is nothing but learned correlations of perceptible features? At least intuitively, it would seem that adults have a concept of animacy that corresponds to the kind of conceptual knowledge Booth and Waxman are talking about. Certainly, adults know the word "animate" and understand the words "living thing". Do these concepts enter into their generalizations of novel nouns? Will adults differentially

generalize names for ambiguous stimuli if told the exemplar represents an animate entity vs. an artifact? One reason to believe they may not is that "animate" – the word – as such, is probably not very correlated with perceptual category formation, that is, with deciding the range of instances that go in a category. If the knowledge used in this task comes from learned correlations, the word "animate" in and of itself may not be potent enough to activate attention to shape and texture. A better cue would be one that is strongly associated with perceptual category decision and category name extensions.

In contrast, by Booth and Waxman's account, the word "animate" should directly activate adults' concept of animacy, indeed, and should therefore direct adults to the relevant knowledge within that concept, enabling them to reason that the relevant properties for categorization are shape and texture.

Experiment 2 tests these ideas using two experimental conditions. In one, adults were told the exemplar is an animate or an artifact; in the other, they were given additional perceptual cues (motion) correlated with animates or artifacts.

Experiment 2

Methods

Participants. 40 undergraduate students participated in this experiment.

Stimuli. The two sets of objects used in Experiment 1 were used in this experiment.

Procedure. Participants were randomly assigned to one of 4 conditions: WordOnly-Animate, WordOnly-Artifact, Word+motion-Animate, and Word+motion-Artifact. Participants were introduced to the exemplar of a set and told "This is a Teema. It is an animate. It is a living thing" in the WordOnly-Animate condition; "This is a Teema. It is an artifact. It was made in a factory" in the WordOnly-Artifact condition. In the Word+motion condition, the exemplar was presented with the corresponding phrases but they were also moved in a walking or slithering motion in the Animate condition or in a rolling or hammering motion in the Artifact condition. Each participant saw both sets and the order of the sets was counterbalanced across conditions. As in Experiment 1, each participant was queried on each of the 7 test-objects twice, for a total of 28 trials.

Results.

We first consider performance in the WordOnly conditions (Figure 5). The number of "yes" responses was submitted to a 2(Kind) x 7(TestItem) repeated measures ANOVA. The analysis revealed a main effect of TestItem ($F_{(1,18)}$ = 41.763, $P < 0.001$) and no other main effects or interactions. There was no significant effect of Kind, that is of hearing Animate versus Artifact instructions ($F_{(1,18)}$ = 1.153, $P > 0.2$). Post-hoc analysis on the critical test item, the shape-only match, also yielded no significant difference between conditions ($\chi^2(1,N=20)$ = .2197, $P > 0.05$). In short, when the ontological kind information consisted solely of words like "animate" and "artifact", attention was not shifted consistently with ontological kind.

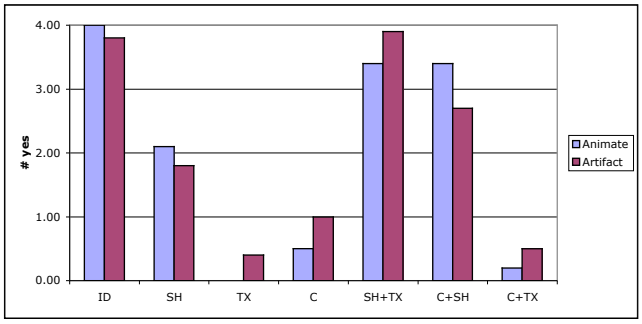


Figure 5. Results for the Word Only condition in Experiment 2.

Figure 6 shows adults' performance in the Word+motion conditions. Again, the number of "yes" responses was submitted to a 2(Kind) x 7(TestItem) repeated measures ANOVA. The analysis revealed a main effect of TestItem ($F_{(1,18)}$ = 26.543, $P < 0.001$) and a significant interaction between Kind and TestItem ($F_{(2,18)}$ = 3.024, $P < 0.01$). Post-hoc analysis confirmed a different pattern of responses in the Animate and Artifact conditions for the shape match. There was a significant difference on the number of adults who said "yes" more than expected by chance on the shape-only match in the Animate versus the Artifact condition ($\chi^2(1,N=20)$ = 5.4945, $P <0.025$). As predicted, participants were more likely to accept the shape only match in the Artifact than in the Animate condition. Thus, adult's novel noun generalizations, when given additional perceptual cues highly correlated with ontological status, did generalize names according to kind – by shape and texture for Animates and by shape for Artifacts.

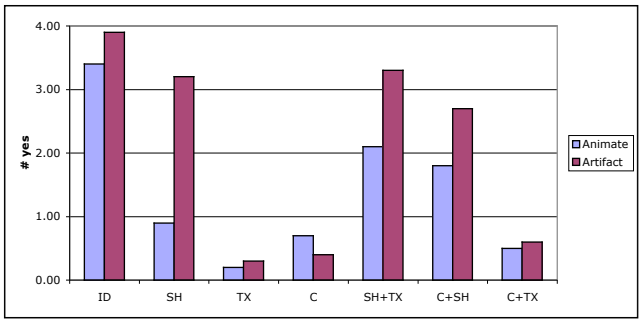


Figure 6. Results for the Word + motion condition in Experiment 2.

Discussion

As expected from the correlational learning account, adults do not show evidence of the animate/artifact distinction in their novel noun generalizations when the only

information they get about the kind of the exemplar is in the form of phrases like "is an animate" and "is a living thing". They do, however, show differing patterns of generalizations when perceptual correlations, in the form of motion cues, are added. This result suggests that, although adults most likely understand what the word "animate" and what the phrase "living thing" mean, this may not be the kind of information that enters in generalizing a novel noun. It also suggests that reasoning about the implication of the ontological status of the exemplar is not the kind of process that primarily guides the extension of novel nouns. Instead, this result suggests that the knowledge that constrains novel noun generalizations is formed by learned correlations of perceptual features and category structure.

General Discussion

Put together, the results of the two experiments support the idea that automatic processes operating on learned correlations of perceptual information can guide word learning. The results of Experiment 1 showed that just listening to a list of words correlated with animates (or a list of words correlated with artifacts) is enough to shift children's attention to the features that typically organize categories of animates (or artifacts). The fact that the lists of words -- without forming complete sentences, without referring to the exemplar, without even co-occurring with the exemplar – can shift attention in kind-specific ways suggests an explanation to Booth and Waxman's results that does not need to appeal to any form of deliberative reasoning on the party of the child. It could be, however, that when given complete sentences forming coherent stories, children do take advantage of that information and more deliberative processes that work on this more propositional information are invoked, but the results in Booth and Waxman (2002) and the results in Experiment 1 can be explained as simple priming, operating on previously learned associations. Further experiments are necessary to determine when this more "conceptual" knowledge is used in extending a novel noun.

Furthermore, the results of Experiment 2 suggest that what we think of as more "conceptual" knowledge may be, at the end, nothing more than a web of correlations, including perceptual features, words, category structure, contexts, and so on. The adults in this experiment failed to shift their attention when explicitly told that the exemplar was animate or artifact, but had no problem doing so when given further correlational support, like watching the to-be-construed-as-animate exemplar "walk" or used as a hammer. Clearly, adults do have highly abstract knowledge about animates and artifacts, and in a different task they might show this. For example, if the task were to make inferences about the exemplar, being told that it is to be construed as an animate might lead to more sound reasoning than watching it "slither". However, it seems that when it comes to extending a novel noun, adults, like children, rely on automatic processes guided by learned correlations.

Conclusion

Extending a novel noun in ways consistent with the ontological kind of its referent is certainly a "smart" thing to do – it allows word learning to proceed quickly and carves the world into useful partitions. However, this "smartness" may be the product of rather "dumb" processes such as associative learning and generalization by similarity. Through the application of these dumb mechanisms in a world that presents regularities we may create smart concepts.

References

Booth, A. E., & Waxman, S. (2002). Word learning is "smart": Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition, 84,* B11-B22.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: the role of categories and appearances. *Child Development*, *58*(6), 1532-1541.

Jones, S., & Smith, L. B. (2002). How children know the relevant properties for generalizing object names. *Developmental Science, 5*, 219 - 232.

Jones, S. S., & Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development, 8(*2), 113-139.

Keil, F. C. (1994). Explanation, association, and the acquisition of word meaning. *Lingua, 92*, 169-196.

Kemler-Nelson, D. G., Russell, R., Duke, N., & Jones, K. (2000). Two-year-olds will name artifacts by their functions. *Child Development, 71*(5), 1271-1288.

Samuelson, L., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*, 1–33.

Smith, L., Colunga, E., & Yoshida, H. (2003). Making an ontology: Cross-linguistic evidence. In D. O. Rakison, L. (Ed.), *Early category learning and concept development: Making Sense of the blooming, buzzing, confusion.* Oxford: Oxford University Press.

Smith, L. B. (1995). Self-organizing processes in learning to learn words: Development is not induction. In C. A. Nelson (Ed.), *Basic and applied perspectives on learning, cognition, and development. The Minnesota Symposia on Child Psychology, Vol. 28* (pp. 1-32). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Yoshida, H., & Smith, L. B. (2002). Shifting ontological boundaries: How Japanese- and English-speaking children generalize names for animals and artifacts. *Developmental Science.*

Making the Implausible Plausible

Louise Connell (louise.connell@ucd.ie)
Department of Computer Science, University College Dublin
Belfield, Dublin 4, Ireland

Abstract

Whenever we must evaluate a theory, consider an excuse, or appraise a situation, we must judge how plausible things appear to us. In short, plausibility judgement occupies a central position in human cognitive life. Recently, it has been shown that the plausibility of a scenario depends on how its events can be connected (Connell & Keane, in press). In this paper, two experiments examine how a normally implausible scenario can be made to seem plausible by forcing a connection between its events. Results show that people's perceptions of a scenario's plausibility can be manipulated by encouraging them to represent events in a causal chain or temporal sequence.

Introduction

Every day, in many different situations, we judge plausibility. Whether evaluating a theory, considering the plot quality of a movie, or listening to child explain how a dish came to be broken, we are assessing how plausible a scenario seems to us.

Across the cognitive science and cognitive psychology literature, plausibility judgement has been shown to be useful in a diverse range of cognitive tasks. People often use plausibility judgements in place of costly retrieval from long-term memory, especially when verbatim memory has faded (Lemaire & Fayol, 1995; Reder, 1982; Reder, Wible & Martin, 1986). Plausibility is also used as a kind of cognitive shortcut in reading, to speed parsing and resolve ambiguities (Pickering & Traxler, 1998; Speer & Clifton, 1998). In everyday thinking, plausible reasoning that uses prior knowledge appears to be commonplace (Collins & Michalski, 1989), and can even aid people in making inductive inferences about familiar topics (Smith, Shafir & Osherson, 1993). It has also been argued that plausibility plays a fundamental role in understanding novel word combinations by helping to constrain the interpretations produced (Costello & Keane, 2000; Lynott, Tagalakis & Keane, in press). Yet, despite its apparent usefulness in cognitive life, the study of plausibility judgement in its own right has been neglected in cognitive science until recently.

The Knowledge-Fitting Theory of Plausibility

Recently, Connell and Keane have proposed the Knowledge-Fitting Theory of Plausibility (2003a, 2003b, in prep; Connell, 2004) to rectify this oversight. According to the Knowledge-Fitting Theory, a plausible scenario is one that fits well with our knowledge of the world. In other words, a plausible scenario has good *concept-coherence*. A concept-coherence view of plausibility suggests that when people make a plausibility judgement, they relate the current scenario to their prior experience, and in some way assess whether it fits in with what they have experienced in the past. For example, take the scenario "*The bottle rolled off the shelf and smashed on the floor.*" People might understand this scenario by drawing a causal inference between the events – that is, the bottle falling *caused* it to smash on the floor. This might lead them to judge this scenario as being highly plausible because prior experience tells them that fragile things often break when they fall on hard surfaces. Put simply, the scenario has a certain concept-coherence. In contrast, if the scenario was "*The bottle rolled off the shelf and melted on the floor*", people might consider it less plausible because they cannot connect the events, as their past experience has few examples of falling fragile objects melting on contact with floors. In other words, this scenario lacks a certain concept-coherence.

The Knowledge-Fitting Theory is supported by a number of empirical findings. For example, Black, Freeman and Johnson-Laird (1986) disrupted the sequence of events in short stories and found that people's plausibility ratings were sensitive to the degree to which the overall concept-coherence of the story had been altered; when people could no longer infer connections between events, they no longer considered the stories plausible. Connell and Keane (in press; Connell, 2004) investigated this issue further, and found that the plausibility of a scenario is not only affected by *whether* the events can be connected, but also by *how* the events are connected. For example, events linked by causal inferences (i.e., *X* was caused by *Y*) were judged to be more plausible than events linked by temporal inferences (i.e., *X* happens after *Y*). These types of scenario were both considered more plausible than scenarios where the events could not be connected at all (i.e., unrelated events). Connell and Keane (2003a, 2003b, in prep.) suggest that causal connections have better concept-coherence than temporal connections because they fit more closely with prior experience, and this makes causal scenarios seem more plausible.

The Current Study

When we judge the plausibility of some scenario in everyday life, it is often with the objective of accepting or rejecting the presented scenario. For example, we may choose to accept or reject an excuse based on whether we find it plausible or not. However, plausibility is not always a binary variable (i.e., a choice between plausible and implausible) (e.g., Black et al., 1986; Connell & Keane, in press). Rather, it may sometimes be considered as a sliding scale between plausibility and implausibility, and we may judge a particular scenario as lying somewhere along this scale. In addition, the plausibility of a particular scenario may not have a constant value: for example, temporal scenarios can be made to seem less plausible when attention is drawn to their non-causal nature (Keane, Connell & O'Donoghue, in prep.).

This paper investigates whether an implausible scenario can be made seem plausible by forcing a particular connection between its events. The Knowledge-Fitting Theory holds that if we cannot connect the events in scenario, we will find it implausible. However, it is possible that if we manage to connect unrelated events in some way, then the plausibility of the scenario might increase. For example, the scenario "*The bottle rolled off the shelf and melted on the floor*" seems generally implausible, but it is possible to construct some set of circumstances that makes it appear more plausible (e.g., the bottle melted because the floor was very hot, because the house was on fire). This suggests that the concept-coherence and plausibility of a normally implausible scenario could be manipulated by encouraging people to make particular connections between events. Therefore, the first experiment asks people to make specific causal or temporal connections between unrelated events, and examines how this influences their decision to accept or reject the scenario. In the second experiment, the same manipulation is used to show how different connections also influence people's plausibility ratings. These results are then related back to the Knowledge-Fitting Theory, and are used to examine what kind of relationship exists between binary and scale plausibility judgements.

Experiment 1

In Experiment 1, participants are presented with implausible scenarios and are asked to judge whether each scenario is plausible or not. Connell and Keane (in press) show that scenarios with no connection between events are considered implausible, while those with causal and temporal connections are considered plausible. This experiment leads people to think about how the events in a scenario may be causally or temporally connected, and examines how these different connections can make people accept as plausible a scenario that would normally be rejected as implausible. For example, take this scenario: "*The teacher misspelled a word. The vase smashed.*" If people are encouraged to represent this normally implausible scenario with a causal or temporal connection between events, then this may lead them to perceive the scenario as plausible. For example, if this scenario was represented within a specific temporal frame (e.g., the vase smashed a second or two after the teacher misspelled a word), then it may have sufficient concept-coherence to appear plausible to some people. Alternatively, if this scenario was represented with a causal chain between events (e.g., the vase smashed because the teacher bumped against it when taking a step back to examine the misspelled word), then it may have sufficient concept-coherence to appear plausible to many people.

This experiment uses different types of question to encourage people to make particular connections between events. In the no-connection control condition, participants judged the plausibility of the scenario directly after reading it, and were not asked to connect the events in any particular way. In the other conditions, participants had to answer a question that encouraged them to represent the scenario in a particular way before judging its plausibility; the causal condition required participants to connect the events as cause and effect, while the temporal condition required participants to connect the events as a temporal sequence. In short, people are expected to find few scenarios plausible in the no-connection condition (poor concept-coherence), more scenarios plausible in the temporal condition (better concept-coherence), and most scenarios plausible in the causal condition (best concept-coherence).

Method

Materials & Design. Materials consisted of twenty "implausible" scenarios, each consisting of two sentences describing unrelated events. The materials were constructed by creating twenty causal scenarios (e.g., "*The surgeon performed the operation. The patient recovered.*") and then randomizing the combinations of first and second sentences (e.g., "*The surgeon performed the operation. The candle flickered.*"). Thus, each scenario contained two events where the cause (Event A) was followed by a different, unrelated effect (Event B). The experimental design was a single between-participants factor (connection type), with three conditions (causal, temporal, no-connection). A between-participants design was chosen to avoid possible confounds (e.g., participants forming a causal connections in all presented scenarios).

Participants. Thirty-six student volunteers from University College Dublin participated in this experiment.

Procedure. Participants were randomly assigned, in equal numbers, to one of the conditions in the experiment.

Instructions stated that each scenario was taken from a story and consisted of two events: Event A and Event B. An example not used in the materials was given:

Event A: *The boy kicked the football.*

Event B: *The branch snapped.*

In the causal condition, participants were asked to write down their answer to the question "Why do you think Event B happened?" and were presented with a sample answer "The branch snapped because the football hit it hard, because the boy was aiming at the tree." In the temporal condition, participants were asked to write down their answer to the question "How long after Event A do you think Event B happened?" and were presented with a sample answer "The branch snapped 2 or 3 seconds later." In all conditions, participants were then given a forced-choice plausibility judgement "Do you find this scenario plausible?" and asked to circle "Yes" if they would accept the scenario as plausible, and to circle "No' if they would reject the scenario as implausible. The scenarios were presented on separate pages in random order, resampled for each participant.

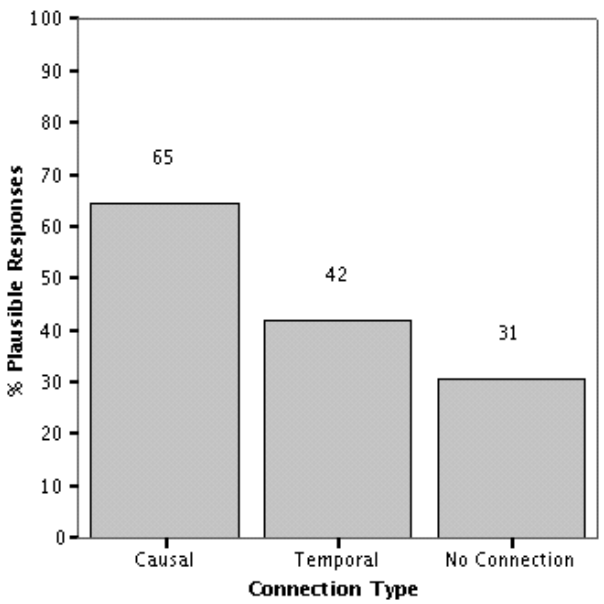


Figure 1: Percentage of scenarios accepted as plausible for each connection type in Experiment 1.

Results & Discussion

The results were in line with predictions, and are shown in Figure 1. Three scenarios were considered plausible by more than 80% of participants in the no-connection control condition, and these were excluded from the analysis in all conditions. While this paper lacks sufficient space to discuss participants' answers in detail, people reported little difficulty in making the connections in the temporal condition, and only occasional difficulty in the causal condition.

People's willingness to accept the scenarios as plausible was influenced by how they had been encouraged to represent the connections between events, as shown by chi-squared analysis, $\chi^2 = 47.74$, $df = 2$, $p < 0.0001$. When asked to represent the events as a specific temporal sequence, people accepted significantly more scenarios as plausible (42%, compared to 31% control), $\chi^2 = 5.31$, $df = 1$, $p < 0.05$. However, the greatest change occurred when people were asked to represent a causal chain between the events, with over twice the number of scenarios being perceived as plausible (65%, compared to 31% control), $\chi^2 = 45.77$, $df = 1$, $p < 0.0001$. In this way, causal connections were reliably better than temporal connections at making scenarios appear plausible (65% compared to 42%), $\chi^2 = 20.71$, $df = 1$, $p < 0.0001$.

So what makes implausible scenarios suddenly appear plausible? Why do people perceive the plausibility of a scenario differently when they are asked why or when events happened? After all, the actual connection made between unrelated events is arbitrary: people are free to come up with any explanation or time frame they choose to connect the events. There is nothing stopping people from causally connecting the events in the no-connection control condition, nor is anything stopping them from causally representing the scenario when answering the temporal question. Indeed, it could be argued that the 31% plausible responses in the no-connection control condition result from the times that people managed to make a causal connection between events without any guidance. In effect, the pattern of results suggests that a kind of "cognitive laziness" is at play, and that people do not put any more effort into representing the scenarios than is absolutely necessary. In the no-connection control condition, most scenarios are judged as implausible because no obvious connection between the events comes to mind. However, in order to be able to answer the temporal question, a certain amount of extra effort must go into connecting the events. If the resulting representation has sufficient concept-coherence, the scenario then seems plausible. Lastly, answering the causal question requires quite a lot of effort, as people must explicitly lay out the circumstances that brought about the second event. The resulting causal representation is likely to have good concept-coherence, and so the scenario is likely to appear plausible. This "cognitive laziness" view is consistent with other studies that have demonstrated people's reluctance to infer causal relations unless prompted to do so (e.g., Keane, 1997; McKoon & Ratcliff, 1992).

This experiment shows that people will accept an implausible scenario as plausible if they are encouraged to connect events in a certain way. However, it is also possible to elicit a more fine-grained judgement of plausibility. When people accept a scenario, is it because they judge it to be very plausible or just moderately plausible? This question is addressed in the next experiment.

Experiment 2

Experiment 2 is identical to Experiment 1, except participants are asked to rate the plausibility of scenarios on a scale from 0-10 instead of simply choosing whether the scenario seems plausible or implausible. In other words, participants are presented with implausible scenarios and are asked to judge how plausible they find each scenario. As with Experiment 1, scenarios in the causal and temporal conditions are expected to be rated as more plausible than those in the no-connection control condition because of their greater concept-coherence. However, there are two possibilities for how the causal and temporal conditions may be distinguished.

The first possibility is that when people in the causal condition accept a scenario as plausible, they actually consider it to be *highly* plausible. This means that causal connections between unrelated events would be considered to have very good concept-coherence; indeed, just as good as for more straightforward causal scenarios. If this were the case, the results would mirror those of Connell and Keane (in press, experiment 1), where causal scenarios were rated as highly plausible (7.8 out of 10) and temporal scenarios were rated as only moderately plausible (4.2 out of 10).

The second possibility is that, although implausible scenarios may become acceptably plausible in the causal condition, they will never seem *highly* plausible. This means that people will perceive causal connections between unrelated events to be of a lower quality (i.e., have poorer concept-coherence) than more straightforward causal scenarios (as in Connell and Keane's study). If this were the case, then ratings in the causal and temporal conditions would be expected to be capped at a level of moderate plausibility.

Method

Materials & Design. The materials and design were the same as in Experiment 1.

Participants. Thirty-six student volunteers from University College Dublin, who had not taken part in Experiment 1, participated in this experiment.

Procedure. The procedure was the same as in Experiment 1, except that participants were asked for a plausibility rating rather than a forced choice judgement. In all conditions, participants were asked "How plausible do you find this scenario?" and asked to circle a rating on a scale from 0 – 10. A rating of 0 was described as meaning the scenario was "not at all plausible", while 5 meant "moderately plausible" and 10 meant "completely plausible".

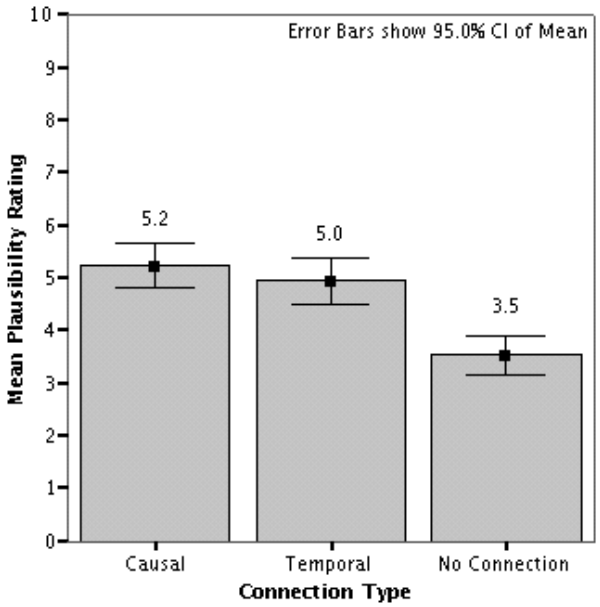


Figure 2: Mean scenario plausibility ratings for each connection type in Experiment 2.

Results & Discussion

The results were in line with predictions, and are shown in Figure 2. As before, three scenarios that were considered plausible by more than 80% of participants in Experiment 1's no-connection control condition were excluded from the analysis in all conditions. Analyses of variance are performed by-participants (F_1) and by-items (F_2), treating participants and items as random factors, respectively.

People's plausibility ratings were influenced by how they had been encouraged to represent the connections between events, as shown by the main effect of connection type, $F_1(2, 33) = 6.30$, $p < 0.005$; $F_2(2, 32) = 31.36$, $p < 0.0001$. Planned comparisons showed that, when asked to represent the events as a specific temporal sequence, people judged the scenarios to be significantly more plausible (5.0, compared to 3.5 control), $F_1(1, 22) = 9.53$, $p < 0.005$; $F_2(1, 16) = 50.58$, $p < 0.0001$. Similarly, when people were asked to represent a causal chain between the events, they perceived the scenarios as being significantly more plausible (5.2, compared to 3.5 control), $F_1(1, 22) = 11.74$, $p < 0.005$; $F_2(1, 16) = 53.23$, $p < 0.0001$. However, there was no difference between the temporal and causal conditions; people did not consider causal

connections between events to be any more plausible than temporal connections, $F_1 < 1$; $F_2(1, 16) = 1.30$, $p > 0.25$.

This experiment shows that, although implausible scenarios may become acceptably plausible in the causal condition, they can never seem *highly* plausible. In other words, while unrelated events can be causally connected, they do not fit with prior knowledge quite as well as more obvious causal connections. For example, we may causally connect the events in the scenario "*The teacher misspelled a word. The vase smashed.*" by assuming that the vase smashed because the teacher bumped against it when taking a step back to examine the misspelled word. While this scenario may just about seem acceptably plausible, it does not seem highly plausible. It certainly does not seem as plausible as a more straightforward causal scenario like "*The cat knocked over a vase. The vase smashed.*" Similarly, for the temporal condition, the results show that connecting events in a specific time frame (e.g., the vase smashed seconds after the teacher misspelled a word) makes a scenario seem somewhat plausible. However, temporal scenarios are considered only moderately plausible at best (Connell & Keane, in press), which suggests that a temporal connection between unrelated events fits with prior knowledge about as well as any other temporal connection. In short, the concept-coherence and plausibility of a normally implausible scenario can be manipulated by encouraging people to make particular connections between events, but the scenario will generally not be judged more than moderately plausible.

So what is the relationship between judging *whether* a scenario is plausible and judging *how* plausible it is? In Experiment 1, we saw that 65% of scenarios in the causal condition were considered acceptably plausible, but yet in Experiment 2, these same scenarios received a plausibility rating of only 5.2 / 10. Similarly, 42% of scenarios in the temporal condition were considered acceptably plausible, and yet were also rated at 5.0 / 10. Analysis of the percentage of plausible responses in Experiments 1 and the mean plausibility ratings in Experiment 2, for each scenario in each condition, shows a direct linear relationship between a scenario's plausibility rating and its acceptability (see Figure 3). This relationship has a significant correlation of $r = 0.88$, $N = 60$, $p < 0.0001$. In short, scenarios with a high plausibility rating will be accepted by most people, while scenarios with a low plausibility rating will be rejected by most people. This suggests that there is no absolute plausibility threshold, above which a scenario will be accepted by everyone as completely plausible. Rather, it depends on what level of acceptability is desired. For example, if we wish 90% of people to accept a scenario, then it should have a mean plausibility rating of approximately 7 out of 10.

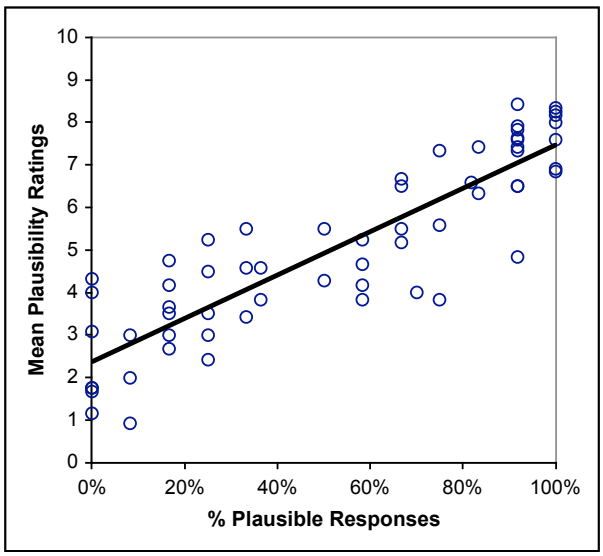


Figure 3: Relationship between scenarios' plausibility ratings and their frequency of acceptance.

General Discussion

This paper shows that an implausible scenario can be made seem plausible by forcing a particular connection between its events. Encouraging people to represent events in a causal chain or temporal sequence, without specifying what the connection should be, alters their perceptions of a scenario's plausibility. In Experiment 1, people are shown to accept 65% of scenarios as plausible once they had explicitly noted a possible causal chain, and 42% of scenarios once they had explicitly noted a possible temporal frame for the events. This compares to a low rate of acceptance for the same scenarios when people are free to make any connections they choose. In Experiment 2, people are shown to consider scenarios as moderately plausible when they are guided into connecting events causally or temporally. In contrast, the same scenarios receive low plausibility ratings when people are free to make any connections they choose. Thus, the novel empirical work reported here demonstrates how people's perceptions of plausibility can be influenced by the circumstances surrounding the task. These findings have implications for any research making use of plausibility judgements, in fields including memory, discourse comprehension, reasoning and conceptual combination.

According to the Knowledge-Fitting Theory of Plausibility (Connell & Keane, 2003a, 2003b, in prep; Connell, 2004), plausibility judgement is about assessing concept-coherence. This view holds that when people make a plausibility judgement, they relate the current scenario to their prior experience, and in some way assess whether it fits in with what they have experienced in the past. Depending on how we

represent a scenario in the first place (i.e., how we connect its events) will therefore determine how well it fits with our prior knowledge. What this paper demonstrates is that the representation of a scenario varies according to how we are encouraged to connect the events. In other words, the concept-coherence and plausibility of a scenario can be manipulated by guiding people towards certain kinds of representation.

The results reported here suggest that people judge plausibility with a certain "cognitive laziness". This means that they do not put any more effort into representing the scenarios than is absolutely necessary. When presented with a scenario, if a possible connection between events does not immediately leap out, then people do not take the trouble to connect the events and instead dismiss the scenario as implausible. However, if circumstances require, people are perfectly capable of connecting even the most disparate events in a coherent manner. For example, the scenario "*The teacher misspelled a word. The vase smashed.*" contains events that are quite difficult to connect. However, people were well able to connect these unrelated events in the causal condition of both experiments, as evinced by the wide and creative variety of causal chains given – e.g., the vase smashed because the teacher bumped against it when stepping back from the blackboard, or because the teacher smashed it in a temper after realising the mistake, or because it was knocked over by a student eager to correct the teacher's error. It is only when circumstances demand it that people overcome their cognitive laziness and take the trouble to reason out a possible connection between events. Indeed, this "cognitive laziness" view is not without its advantages. As well as allowing people to judge plausibility with the least amount of computational expense, it is also tends towards false negatives rather than false positives. This makes it quite a sound approach, as it is safer to reject a scenario that later proves viable than to accept one that later proves unviable.

Acknowledgments

This work was supported by funding from the Irish Research Council for Science, Engineering and Technology, under the Embark Initiative. Thanks also to Dermot Lynott for valuable comments, and to Amy Bohan and Fergus Toolan for experimental assistance.

References

Black, A., Freeman, P., & Johnson-Laird, P. N. (1986). Plausibility and the comprehension of text. *British Journal of Psychology, 77*, 51-60.

Collins, A., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13*, 1-49.

Connell, L. & Keane, M. T. (2003a). PAM: A cognitive model of plausibility. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Connell, L., & Keane, M. T. (2003b). The knowledge-fitting theory of plausibility. *Proceedings of the Fourteenth Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 40-45). Ireland: Trinity College Dublin.

Connell, L. (2004). *A cognitive theory and model of plausibility.* Ph.D. thesis, Department of Computer Science, University College Dublin, Ireland.

Connell, L., & Keane, M. T. (in press). What Plausibly Affects Plausibility? Concept-Coherence & Distributional Word-Coherence As Factors Influencing Plausibility Judgements. To appear in *Memory and Cognition*.

Connell, L., & Keane, M. T. (in prep.). The knowledge-fitting theory of plausibility. *Manuscript in preparation*.

Costello, F., & Keane, M. T. (2000). Efficient Creativity: Constraints on conceptual combination. *Cognitive Science, 24*, 299-349.

Keane, M. T. (1997). What makes an analogy difficult?: The effects of order and causal structure in analogical mapping. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 946-967.

Keane, M. T., Connell, L., & O'Donoghue, N. (in prep.). Questioning plausibility judgements: how questions and inferences affect plausibility. *Manuscript in preparation*.

Lemaire, P. & Fayol, M. (1995). When plausibility judgments supersede fact retrieval: The example of the odd-even rule effect in simple arithmetic. *Memory and Cognition, 23*, 34-48.

Lynott, D., Tagalakis, G., & Keane, M. T. (in press). Conceptual combination with PUNC. To appear in *AI Review*.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440-466.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 940-961.

Reder, L. M. (1982). Plausibility judgments vs. fact retrieval: Alternative strategies for sentence verification. *Psychological Review, 89*(3), 250-280.

Reder, L. M., Wible, C., & Martin, J. (1986). Differential memory changes with age: Exact retrieval versus plausible inference. *Journal of Experimental Psychology: Learning, Memory & Cognition, 12*(1), 72-81.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49*, 67-96.

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory and Cognition, 26*, 965-978.

Chess Masters' Hypothesis Testing

Michelle Cowley (cowleym@tcd.ie)
University of Dublin, Trinity College,
Dublin 2, Ireland

Ruth M. J. Byrne (rmbyrne@tcd.ie)
University of Dublin, Trinity College,
Dublin 2, Ireland

Abstract

Falsification may demarcate science from non-science as the *rational* way to test the truth of hypotheses. But experimental evidence from studies of reasoning shows that people often find falsification difficult. We suggest that domain expertise may facilitate falsification. We consider new experimental data about chess experts' hypothesis testing. The results show that chess masters were readily able to falsify their plans. They generated move sequences that falsified their plans more readily than novice players, who tended to confirm their plans. The finding that experts in a domain are more likely to falsify their hypotheses has important implications for the debate about human rationality.

Hypothesis Testing

People understand everyday and scientific phenomena by generating hypotheses to explain them. They achieve a true understanding only by *testing* hypotheses by searching for proof. There are two main ways people can test the truth of their hypotheses. They can either seek *confirmation*: evidence that is consistent with a hypothesis, or *falsification*: evidence that is inconsistent with a hypothesis. Falsification is generally considered better than confirmation: no matter how much evidence is gathered to confirm a hypothesis, there remains the possibility of refutation later (Popper, 1959). Confirmation could lead to the endorsement of untrue ideas and so if people are rational, they should attempt to falsify their hypotheses. Many cognitive scientists have interpreted experimental findings on hypothesis testing within the framework of falsification (e.g., Wason, 1960). The conclusion has sometimes been reached that when people fail to attempt to falsify, they fail to think rationally.

Early research on hypothesis testing found that people were prone to a *confirmation bias*: they tended to search for confirming evidence and avoid falsifying evidence (e.g., Wason, 1960). Confirmation bias has sometimes been viewed as evidence of human irrationality, for example, it may lead people to form prejudiced beliefs (e.g., Aronson, 1995). But the idea that human hypothesis testing is irrational presents a paradox: How can it be flawed given that it has led to important civil, technological and scientific discoveries? There are two possible answers: one possibility is that testing hypotheses through confirmation is more useful than indicated by a Popperian analysis, and a second possibility is that people are more capable of falsification than experimental evidence has revealed so far. We will first outline the view that confirmation is a useful strategy to test hypotheses and the view that falsification may be conceptually impossible (e.g., Poletiek, 1996). Then we will show that falsification is in fact possible in a domain that has been a trusted test-bed for theories of cognition for almost forty years: chess problem solving. We will consider experimental results that testify to high levels of falsification in the hypothesis testing of chess masters (Cowley & Byrne, 2004).

Confirmation: Vice or Virtue?

Irrational hypothesis testing in the form of confirmation bias was first reported in the 2-4-6 task (Wason, 1960). Participants in this task are required to discover the rule to which the number triple 2-4-6 conforms. The participants are analogous to scientists and the rule is analogous to a law of nature to be discovered. Participants test their hypotheses by generating other number triples and they are told by the experimenter whether each triple conforms to the rule or not. The rule in the 2-4-6 task is the deliberately general rule of 'any ascending numbers'. The salient features of the 2-4-6 triple tend to induce incorrect hypotheses, for example, participants tend to focus on its properties of even numbers and numbers ascending in twos. Participants who generate these hypotheses, 'ascending even numbers' or 'numbers ascending in twos' can discover the real rule 'any ascending numbers' in only one way: by generating triples that falsify their hypothesis. For example, a participant could try to falsify the 'ascending even numbers' hypothesis by generating the triple '3-5-7'. They would discover their hypothesis is false when the experimenter tells them that '3-5-7' is consistent with the real rule. But participants overwhelmingly generated confirming triples such as '10-12-14'. The triple confirms their hypothesis and it is also consistent with the real rule and so the experimenter tells them '10-12-14' is consistent. They announce their incorrect hypothesis as the rule and fail to solve the task correctly.

Confirmation bias has been demonstrated many times in the 2-4-6 task and in other related laboratory tasks, for example, in a task in which participants are required to discover the law governing the motion of particles in an artificial universe displayed on a computer screen (e.g., Mynatt, Doherty, & Tweney, 1978).

But do people confirm their hypotheses only in artificial laboratory tasks? Perhaps they are better able to falsify in real world contexts where they can access their knowledge about the task? In fact, the tendency to confirm has been observed in NASA Apollo mission scientists (Mitroff, 1974), and in the notes of scientists such as Alexander Graham Bell (Gorman, 1995). It is possible that confirmation is useful. For example, the participants who successfully discovered the rule in the complex artificial universe task tended to be those who confirmed their hypotheses in the early stages of their attempted discovery of the rule, and then tried to falsify when they had a well-corroborated hypothesis (Mynatt et al., 1978). Perhaps it is only when a hypothesis worth testing has been established that it is necessary to attempt to falsify it. Confirmation and falsification may be complementary strategies for successful hypothesis testing.

But it is also possible that people do not falsify because they cannot (Poletiek, 1996). According to this view when people generate a hypothesis it is their *best guess* about the truth, and it does not make sense for them to try to show that their best guess is wrong. In a version of the 2-4-6 task, participants were encouraged to generate their best guess about what the rule might be and then they were instructed to perform falsifying tests on it. The instruction to falsify decreased the number of positive triples, such as '10-12-14', which is consistent with the best guess 'even ascending numbers'. The instruction to falsify also increased the generation of negative test triples, such as '3-5-7' which is inconsistent with the best guess 'even ascending numbers'. However, this test is only a falsifying one if the participant expects the experimenter to say that the triple is consistent with the real rule (and then the participant would know that the hypothesis 'even ascending numbers' was wrong because ascending odd numbers are consistent too). But if the participant generates the inconsistent '3-5-7' triple and expects the experimenter to say that it is *not* consistent with the real rule, then they have attempted to confirm their hypothesis (albeit with a negative triple). In fact, participants generated triples that were inconsistent with their hypothesis (negative triples) *but* they expected them to be inconsistent with the experimenter's rule. The participants could not seem to make sense of the instruction to falsify. The instruction to falsify may be impossible to carry out (Poletiek, 1996).

Given the ideas that confirmation is useful and falsification is impossible, does it follow that the normative prescription of falsification is flawed, rather than human rationality? Perhaps, not. Even when a hypothesis is the best guess it is not necessarily an accurate representation of the truth. We turn to the case for falsification next.

Falsification and the Path to Truth

Consider the following example:

> *You are a scientist and your job is to identify the cause of a dangerous new disease. You identify a previously unrecognized virus in tissue samples of symptomatic patients and your hypothesis is that this 'new virus' is the cause of the disease. However, other scientists have identified two viruses, including your new virus in their tissue samples. They hypothesize that it is the 'other virus' and not the new virus that is the cause. Both hypotheses have confirming evidence. A case is reported where the new virus is present and the other virus is absent. What should you conclude?*

A situation similar to this one faced scientists working on the cause of the SARS epidemic. They concluded that the 'new virus' hypothesis was correct. The case where the 'other virus' was absent falsified the 'other virus' hypothesis and corroborated the 'new virus' hypothesis. The example illustrates how important falsification can be.

There are many situations in which it is helpful to anticipate the ways in which a hypothesis or plan could go wrong. For example, it may be helpful to falsify in interactions with a collaborator or opponent, whether in contexts such as political or social engagement, or in games such as tic-tac-toe or poker. The importance of considering what might go wrong is observed in cases of military strategy, for example, in Northern Ireland (Mallie, 2001). Attempts to falsify hypotheses, particularly plans of action, could help reduce costly mistakes.

The merits of falsification are not lost on experts, as the SARS example illustrates. It may even be the case that the ability to falsify is part of what makes an expert (Cowley & Byrne, 2004). The competitive nature of science may ensure that different groups of scientists attempt to falsify their opponent's theories even if they only attempt to confirm their own. The refutation of a theory is often discovered by someone who did not invent the theory (Kuhn, 1996). Hypothesis testing in scientific discovery may benefit from a strategy of attempting to confirm a hypothesis until there is sufficient corroboration for it to be considered seriously, and then attempting to falsify it, just as in the 'artificial universe' task (Mynatt et al., 1978). Perhaps more importantly, experts may generate high quality hypotheses from the outset. An exceptional scientist such as Alexander Graham Bell may have tended to confirm rather than falsify his hypotheses because his expertise ensured that his hypotheses were exceptionally good (and there is a smaller potential set of falsifying evidence for a good quality hypothesis than for a poor one).

As these observations suggest, a more systematic study of expert hypothesis testing is warranted. We chose the game of chess as our expert domain because it meets the essential criteria: it is possible to identify a large sample of experts whose expertise is objectively defined and categorised relative to each other, and it is a task that draws directly on participants' expert knowledge and experience.

Chess and Hypothesis Testing

Studies of chess have contributed substantially to understanding cognition, including problem solving (Newell & Simon, 1972), chunking in working memory (Chase & Simon, 1973), and expertise (De Groot, 1965). Findings from research on chess have successfully explained expertise in non-game domains such as physics (e.g., Larkin, McDermott, Simon, & Simon, 1980). Chess offers great potential for an investigation of expert hypothesis

testing. Of course, choosing a move in chess may depend on a variety of processes including accessing a large repository of chunked domain knowledge about possible opponent moves (e.g., Chase & Simon, 1973; Gobet, 1998a). Our suggestion is that hypothesis testing may be one of several important processes for selecting a move in chess. We expect that expert master players will be better than novices at falsifying their planned moves by thinking about opponent moves that could ruin their plan (for details see Cowley & Byrne, 2004).

Our key research question is, do experts and novices differ in their ability to find refutations to lines of play in chess? We conceptualize hypothesis falsification in chess as finding opponent moves that refute the moves a player examines for play. The opponent moves could ruin the player's plan and worsen the player's position. We address an important aspect of choosing a move that has never been systematically investigated: the evaluation of move sequences.

Hypothesis Testing in Chess

The overall goal of chess is to checkmate the opponent by attacking the opponent king and eliminating all the possible ways the opponent king can escape the attack. Chess thinking may consist of exploring different alternative paths in a 'problem space' (Newell and Simon, 1972). The problem space consists of the initial problem state, that is, the start of the game, intermediate problem states, for example, capturing an opponent piece, and the end state (checkmate). Progress from state to state is achieved through *operators*, that is, in chess the way chess pieces are allowed to move. For example, a bishop operates diagonally backwards and forwards and captures on the square it lands on for any one move.

At the beginning of a game of chess the two players have equal numbers of pieces and theoretically equal chances of securing a win, loss or draw. To secure the best possible result the players must play moves they hypothesize to be so good that they cannot be refuted (Saariluoma, 1995). Refutation (that is, hypothesis falsification) occurs when the opponent plays a move that results in a worsening of the player's position. For example, a player may play a move that he or she plans to be a good move, but the opponent replies with a move that stops the player's plan. The opponent's play worsens the player's position and reduces the player's chance of a win.

There may be three major processes in the choice of a chess move: exploration, elaboration and proof (DeGroot, 1965). Evidence of hypothesis testing is available in the proof process. A chess player tests how good a move is by mentally generating move sequences following on from that move. For example, a move sequence might be: "If I move my knight to that square, you might move your pawn to attack my knight, and then I will have to retreat, and that is really bad for me…". In this example, the move sequence is evaluated as leading to a negative outcome: a falsification has been found for the knight move. Move sequences can be evaluated as leading to either a positive, negative or neutral outcome for a chess position. We conceptualize move sequences that are judged to lead to a positive outcome as akin to evidence confirming that a particular move is a good move. Move sequences evaluated as leading to a negative outcome are akin to evidence falsifying a move that was thought initially to be a good move. Move sequences evaluated as leading to a neutral outcome are neutral evidence.

Accessing Hypothesis Testing in Chess

We carried out an experiment on hypothesis testing in chess players (see Cowley & Byrne, 2004, for details). The 20 participants (19 men and 1 woman) were registered members of the Irish Chess Union. The participants were classified according to the Elo system, which calculates expected playing strength value on the basis of tournament and league results, and the value varies from approximately 1000 for an absolute novice and over 2800 for the world champion. We tested *experienced* novices (mean rating of 1509) and experts (mean rating 2240). The expert group included experts from different Elo categories of expertise, including one grandmaster (Irish Elo >2500) two international masters (Irish Elo > 2300), three Fide masters (Irish Elo > 2200, i.e. International Chess Federation masters), and four initial category experts (Irish Elo > 2000). All international class masters living in Ireland at the time participated in the study (for further information on participant details see Cowley & Byrne, 2004).

We presented the participants with six board positions, three normal and three random (as well as an initial practice position). The board positions were chosen from games in chess periodicals. They were middle game positions with 22-26 pieces to ensure complexity and to rule out the chances that the masters' had seen them before. Importantly, they were 'equality outcome' positions, where there were equal chances with best play for both black and white pieces. This constraint ensured that there would be no obvious confirming or falsifying move sequences. The positions were chosen with the assistance of a chess expert (who was not a participant in the study). See figure 1 for an example of a chess position used.

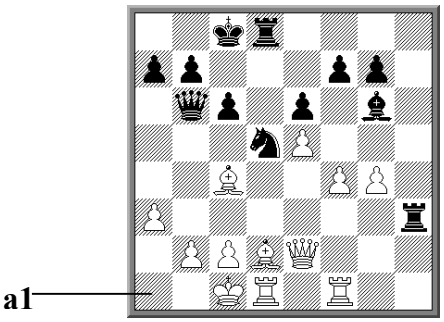


Figure 1: Position 1 with white to play (and the co-ordinate a1 is also illustrated in this diagram).

The participants' task was to, "choose a move you would play in the way you are used to going about choosing a move in a real game". They were given instructions to think aloud, and their verbalizations were recorded by dictaphone. It is instructive to focus on the master level players (for

comparison with masters studied in the chess literature previously) and to this end we selected the think-aloud protocols of five *master level* players (i.e. 1 Grandmaster, 2 International Masters, and 2 Fide International Chess Federation Masters), and compared them to the think-aloud protocols of five novice chess players, chosen at random from the full sample of novices (for other analyses see Cowley & Byrne, 2004).

Moves examined by the player during think-aloud are verbalized using algebraic chess notation, for example a sequence of moves verbalized was: f5 exf5 gxf5 Bh5 Qg2 Rh4. This notation describes each piece and the location of the square it will go to on the chess board. Each square on a board has a location name called an algebraic coordinate. The letters a-h are horizontally along the bottom of the board. The numbers 1-8 are vertically up the board. Each type of piece is given a letter in upper case format. Each coordinate is given a letter in lower case format alongside a number. So for example, the move 'Ra1' refers to a rook piece (R) moving to the a1 square. Or, the rook could go to the b1 square to the right of a1 (Rb1). A sequence of such moves is a move sequence. All of the players were sufficiently fluent with algebraic notation to be able to 'think aloud' using it. Three minutes thinking time was allotted for choosing a move as it is just over the average time per move in tournament play. Exposure for each board position was timed using a standard tournament chess clock, each clock was set at three minutes and when the clock's flag fell participants were told that their time was up.

To accurately access hypothesis testing we also needed participants to provide us with an evaluation of each move sequence that they examined. However, spontaneous evaluation in chess has a low probability of verbalization (Newell & Simon, 1972). Accordingly we used a combined methodology of think-aloud followed by retrospective evaluation. Verbalized move sequences were recorded not only by dictaphone but also by the experimenter (the first author) in algebraic notation concurrent with think-aloud. The experimenter asked the participants for their evaluation of each move sequence, by first saying back the move sequence immediately after each chess problem to reduce retrospective error and interference (Ericsson & Simon, 1993). The participants were then asked to evaluate each move sequence as having lead to a positive, negative or neutral outcome for their positions.

Scoring confirming and falsifying hypothesis tests

The transcribed think-aloud protocols for the responses to the normal board positions were segmented into episodes, move by move. We constructed 'problem behavior graphs' (using Newell and Simon's guidelines) for the responses to the three normal board positions for each of the ten selected participants (thirty problem behavior graphs in total). These graphs plot each move sequence and its corresponding retrospective evaluation. To illustrate we present in Figure 2 a small fragment of a master's problem behaviour graph for one board position.

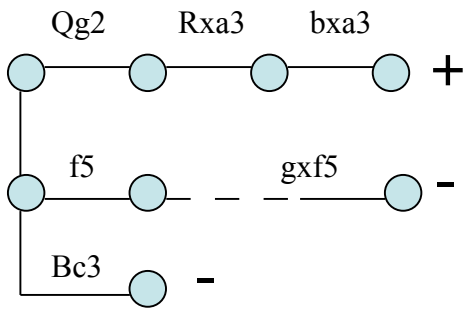


Figure 2: A fragment of a problem behaviour graph constructed from a chess master's protocol.

Each line across represents a move sequence. The order of search is from left to right, then down. Each circle (i.e. node) represents a new position following a move made in the problem space. For example, Qg2 means the player thought aloud about the possibility of moving his queen to the g2 square. Next the player thought aloud about a possible reply from his opponent to his Qg2 move, that is, the move Rxa3 where the opponent moves their rook to the a3 square, and the x indicates that the rook captures a piece, in this case a pawn. Next the player thought aloud about his reply to this opponent's move, that is, bxa3 where he would move his pawn on the b file to a3 (pawn moves do not have a P in front of them), and capture the opponent's rook. The plus sign shows that the player evaluated this move sequence as positive for him. The next line sequence begins with the player thinking about f5, that is, a pawn moves to the f5 square. The next utterance the player makes is gxf5, that is, the pawn on the g file of the board captures a pawn on the square f5. This move is only possible for the player and not his opponent. The player has generated a move sequence that mentions only his own moves and does not mention opponent moves. The dashed line captures these *skipped* moves. The minus shows a negative evaluation. Each problem behavior graph incorporates the think-aloud move sequences with the retrospective evaluation (positive, negative, or neutral).

We used *Fritz 8* (one of the most powerful current chess programs) to obtain an objective evaluation of the chess position that occurs at the final move of each sequence (i.e. terminal node). For readers familiar with *Fritz 8*, we used the infinite analysis module, in which each move sequence is evaluated at least from 11ply from the terminal node (see Chabris & Hearst, 2003). The evaluations provided by *Fritz 8* enable us to identify move sequences that would genuinely be positive or negative for a player. We could distinguish between the move sequences that a player indicated as leading to a positive outcome for their position and that the program established would lead to a positive outcome if played, from the move sequences that a player identified as positive, but that the program established would in fact lead to a negative outcome. We conceptualize confirmation bias as a move sequence that a player evaluates as leading to a positive outcome for them, when in

fact it leads to a negative outcome. Likewise, we were able to distinguish the move sequences that a player identified as leading to a negative outcome and that the program established would be negative if played, from move sequences that the player identified as negative, but that the program established were in fact positive for them. We conceptualize falsification as a move sequence that a player evaluates as leading to a negative outcome for them, when in fact it leads to a negative outcome.

Hypotheses Testing in Chess Masters' Thinking

Masters tended to think about 8 move sequences on average for each board position, and experienced novices tended to think about 6 move sequences. A total of 218 move sequences were generated by the 10 players for the three normal board positions (N = 122, M = 8.1 for each board position for the masters, N= 96, M = 6.4 for novices).

Four types of move sequences were identified from the problem behaviour graphs. (1) 50% were *complete move sequences* where every move for the player and his or her opponent was articulated along the move sequence. (2) 25% were *skipped move sequences* where an essential move was not mentioned somewhere in the move sequence. (3) 19% were *base skip sequences* where the first move or 'base move' of the sequence was not mentioned. (4) 6% were *ambiguous move sequences* where the move sequence could not be interpreted. Only the complete move sequences lend themselves to objective evaluation by *Fritz 8*, so we concentrate our analysis here on these hypothesis tests (see Cowley & Byrne, 2004 for further details).

A complete move sequence is scored using the following criteria: (a) whether it is predicted by the player to lead to a positive, negative, or neutral outcome, and (b) whether it is evaluated objectively by *Fritz 8* as leading to a positive, negative or neutral outcome. Thus there are nine possible hypothesis tests for complete move sequences, as Table 1 illustrates. Confirmation bias corresponds to the '+/-' cell in Table 1, and falsification corresponds to the '-/-' cell. These two types of evaluation accounted for 42% of all evaluations.

Table 1: Objective and subjective evaluations of move sequences ('+' refers to a positive evaluation, '-' to a negative one, '=' to a neutral one; '+/-' means the player's evaluation was positive and the program's evaluation was negative).

	Fritz 8's evaluations		
	Positive	negative	neutral
Player's evaluations			
Positive	+/+	+/-	+/=
Negative	-/+	-/-	-/=
Neutral	=/+	=/-	=/=

Falsification In three of the cells of Table 1 (the three on the diagonal from upper left to lower right), the subjective evaluation of the player matches the objective evaluation of the computer program. One of these matching cells is particularly important for our prediction that experts falsify more than novices: genuine falsification occurs in the situation captured by the '-/-' cell in Table 1, in which the player and the program both evaluated the outcome of the move sequence as negative. Chess masters generated more of these falsifying move sequences than novices (M = 3.2 for masters, M = 1.2 for novices) and this difference was reliable (t (8) =2.02, p = .039).

The result indicates that chess masters are capable of falsifying their plans by identifying opponent moves that would worsen the master's position. People are able to falsify (pace Poletiek, 2000). Domain expertise may facilitate this falsification. Moreover, the moves chosen by chess masters for play at the end of each of the three board positions were evaluated by *Fritz 8* as objectively better moves than novices (the quality of moves is measured in terms of 'pawn advantage' or 'pawn disadvantage', and it was +0.309 pawn advantage for masters compared to –1.2 pawn disadvantage for experienced novices). The result is consistent with the idea that the ability to falsify may contribute to making better moves in chess.

Confirmation Bias Confirmation bias occurs when a move sequence is evaluated subjectively by the participant as leading to a positive outcome, but evaluated objectively by the computer program as leading to a negative outcome (the **'+/-'** cell in Table 1). The results show that novices produced somewhat more instances of confirmation bias than masters (M = 2.6 for novices and M = 1.6 for masters). Although the difference was in the predicted direction it was not reliable (t (8) = 1.443, p = .094).

Positive and Negative Testing The nine test types in Table 1 can be categorized into three groups: (1) Objective tests: the player's positive, negative and neutral evaluations matched *Fritz 8*'s evaluations (the three cells on the diagonal from upper left to lower right mentioned earlier), and this category includes the falsification tests. (2) Positive bias tests: the player's evaluation was more positive than *Fritz 8*'s. The three cells in this category include the second and third cells in the first row (**'+/-'**, **'+/='**), and the middle cell in the third row (**'=/-'**), and this category includes the confirmation bias tests. (3) Negative bias tests: the player's evaluation was more negative than *Fritz 8*'s. The three cells in this category include the second and third cells in the first column (**'-/+'**, **'=/+'**), and the middle cell in the third column (**'-/='**).

Chess masters generated reliably more objective tests than novices (M = 6.6 for masters and M = 2.4 for novices). Novices generated somewhat more positive bias tests than the masters (M = 5 for novices and M = 3.4 for masters), but the difference was not reliable. They generated a similar amount of negative bias tests (M = 1.8 for masters, M = 1.2 for novices), as Figure 3 shows.

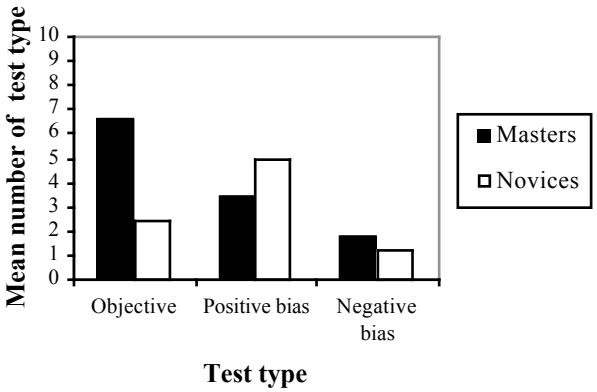


Figure 3: The mean number of objective tests, positive bias tests and negative bias tests generated by masters and novices. (Instances of falsification and confirmation bias are included in these categories).

Conclusions

People are capable of falsifying their hypotheses. Our experimental results show that chess masters falsified their hypotheses: they thought about how their opponent might refute their plan in their move sequences. Chess masters tended to evaluate their moves as good or bad for them more realistically than experienced novices: their judgments matched the objective evaluations of one of the most highly advanced chess computer programs, *Fritz 8*. Experienced novices exhibited something of a confirmation bias: they tended to think about how their opponent would play moves that fit in with their plan, somewhat more than chess masters did. Novices, somewhat more than masters, tended to evaluate their moves as better for them than they were objectively. The evidence that chess masters can falsify suggests that it may be premature to conclude that the normative prescription of falsification is flawed. In this case falsification can be considered a useful and rational strategy.

Hypothesis testing may be influenced by domain expertise. How does domain knowledge affect the ability to falsify by chess experts? We plan to explore this question by examining how masters test their hypotheses for random board positions compared to novices. If falsification relies on domain knowledge, then masters should tend not to falsify their hypotheses about move sequences in the random board positions as often as they do in the normal board positions. Nonetheless, they may attempt to falsify more than experienced novices, if their expertise has helped them to develop a strategy of falsification in this domain.

Acknowledgements

We thank Grandmaster Alexander Baburin, Fintan Costello, Phil Johnson-Laird, Mark Keane and Caren Frosch for helpful comments, and Peter Keating for help with the problem behaviour graphs. Special thanks to Mel O'Cinneide for help with the chess positions. This research was funded by the Irish Research Council for the Humanities and Social Sciences.

References

Aronson, E. (1995). *The Social Animal*. New York: Worth/ W. H. Freeman.

Chabris, C. F., & Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: effects of playing speed and sight of the position on grandmaster chess. *Cognitive Science, 27*, 637-648.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed), *Visual Information Processing*. New York: Academic Press.

Cowley, M., & Byrne, R. M. J. (2004). Hypothesis testing in chess masters' problem solving. *Manuscript in preparation.*

De Groot, A. (1965). *Thought and choice in chess*. The Hague: Mouton.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data.* USA: MIT Press.

Gobet, F. (1998a). Expert memory: A comparison of four theories. *Cognition, 66*, 115-152.

Gorman, M. E. (1995). Confirmation, disconfirmation, and invention: The case of Alexander Graham Bell and the telephone. *Thinking and Reasoning, I (1),* 31-53.

Kuhn, T. S. (1996). *The structure of scientific revolutions*. USA: University of Chicago Press.

Larkin, J. H. Mc Dermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208*, 1335-1342.

Mallie, E. (2001). *Endgame in Ireland*. London: Hodder & Stoughton.

Mitroff, I. (1974). *The subjective side of science.* Amsterdam: Elsevier.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology, 30*, 395-406.

Newell, A., & Simon, H. A. (1972*). Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Poletiek, F. H. (1996). Paradoxes of falsification. *Quarterly Journal of Experimental Psychology, 49A*, 447-462.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Saariluoma, P. (1995). *Chess players' thinking: A cognitive psychological approach.* UK: Routledge.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Synchronization Among Speakers Reduces Macroscopic Temporal Variability

Fred Cummins (fred.cummins@ucd.ie)
Department of Computer Science
University College Dublin
Belfield
Dublin 4
Ireland

Abstract

A recent method for restricting inessential variation in speech is presented. Synchronous Speech is obtained by having two subjects read a prepared text in synchrony. Past results demonstrate that this is easy for subjects to do, and that some prosodic variability is greatly reduced when reading synchronously. Particular advantage has been found in the analysis of pauses and fundamental frequency variation, where synchronous speech has been demonstrated to exhibit markedly less inessential variability, thus furthering analysis and modeling. Here, duration ratios within a phrase are compared across synchronous and solo conditions. Variables associated with global timing and with the relationships between phrases are shown to be more consistent in the synchronous condition, while smaller units are not noticeably affected by the speaking condition. No systematic artifacts are found to be introduced by asking subjects to read in synchrony.

A Method for Restricting Variability

Synchronous Speech is obtained with the simple expedient of having two subjects read a prepared text together, with the minimal instruction to attempt to maintain synchrony (Cummins, 2002). The reason for constraining subjects in this manner is perhaps best appreciated by analogy with the difficult task of attempting to reconstruct a musical score, based only on a recording of a specific musician (Heijink et al., 2000). This task is interestingly similar to the work of the theoretically minded phonetician who attempts to uncover control and timing information, along with combinatorial units, from the continuous stream of speech.

If one were faced with this task, it is worth considering which musician would give one more tractable data: the soloist, or the 14th violin player in the string section. Neither will reproduce the durations (or pitches) specified in their score exactly, of course, due to the inherent underspecification of the score. Both players will overlay some inherent biophysical noise, along with conventional timing variability, such as the predictable decellerando at the end of a phrase. The soloist will add additional complexity, however, in keeping with her role as the expressive focus in performance, making the inverse mapping from the recording to the score considerably more difficult.

Now return to the position of the laboratory phonologist (or theoretical phonetician). An overarching goal is to deduce the units of control which relate to the linguistic message being uttered, and to uncover their mutual relations. This is not so different in kind from the above musical analog, though additional levels of complexity undoubtedly arise. Signal variability which is related to the linguistic content is relevant, while (for many purposes) one might like to find a way to reduce or exclude variability of para- or non-linguistic origin. The approach which I and colleagues have recently been following is to constrain the speaker to speak in time with another co-speaker. For this purpose, speakers read through a given text silently to familiarize themselves with it, and then commence reading together on a signal from the investigator. For many purposes, recording using near field head-mounted microphones onto the left and right channels of a single stereo file is sufficient to separate the two speakers while preserving the relative temporal alignment of speech events.

We call speech collected in this manner Synchronous Speech, and both the task and the product have provided us with much food for thought (Cummins, 2001; Cummins, 2002; Cummins and Roy, 2001; Cummins, 2003). In this paper, I will summarize those findings which have best revealed the advantages of this novel method, then provide some new results which examine the variability of intervals below the whole phrase, and finally provide pointers to areas I believe might benefit from adoption of the method.

Properties of Synchronous Speech

Synchronizing with a co-speaker, without extensive practice, turns out to be simple for subjects to do (Cummins, 2002; Cummins, 2003). After reading through a simple text once, and being given a start signal, subjects typically manage to keep inter-speaker lags to average values of around 60 ms at phrase onsets, and 40 ms or less after the first syllable or so. Rather surprisingly, extensive practice at the task does not improve the degree of synchrony significantly (Cummins, 2003), although with repeated readings of the same text, and with the same co-speaker, a slight improvement may be detected. Visual contact with the co-speaker does seem to have a small beneficial effect on synchrony, even though subjects are typically attending to a read text in front of them (Cummins, 2003).

In experiments done to date, speakers have not been carefully matched for familiarity, intrinsic speaking rate

or volume. Among the heterogeneous pairs of speakers we have studied to date, most appear to be collaborating, producing speech at a relatively slow rate (but faster than some of the slowest speakers' natural reading rate). We have not yet (in over 60 pairs of speakers) found a speaking pair in which one speaker consistently lagged behind the other. Rather, they seem to genuinely speak together, with a high degree of synchrony.

Phrasing and Pauses

One of the first properties of Synchronous Speech we noticed, was that phrasing, i.e. the division of a long stretch of speech into intonational units separated by pauses, appeared to be much more consistent in Synchronous Speech than in control readings done alone. In an initial pilot with 4 speakers, we found that in 48 'solo' readings, pauses occurred at points other than major expected phrase breaks 48 times (Cummins, 2002). By contrast, in Synchronous Speech, there were only 4 such idiosyncratic pauses in 24 paired readings.

These findings have been extended in the studies of pauses in Synchronous Speech (Zvonik and Cummins, 2002; Zvonik and Cummins, 2003), who found that inter-speaker variability in pause duration was greatly reduced in Synchronous Speech, compared with 'solo' speech. The reduced variability allowed the identification of a quantitative relationship between pause duration and the length (in syllables) of the preceding phrase—a relationship which was obscured in the rather more variable solo data. Specifically, we found a restricted distribution of clauses of less than 300 ms length. These pauses were far more likely to occur when the preceding phrase was relatively short (less than 11 syllables long)[1]. We examined pause duration in readings by 6 speakers (3 pairs) of 19 short texts (13 distinct authors). Table 1 shows the proportion of pauses in each environment (preceding phrase long or short, following phrase long or short) which were below 300 ms. The preponderance of short pauses in an environment following a short phrase is clear.

Table 1: Number of pauses of duration less than 300 ms as a function of the length of the surrounding Intonational Phrases. For IPs, 'short' is here taken to mean less than or equal to 10 syllables. Reproduced from Zvonik (2004, unpublished PhD thesis).

Preceding IP	Following IP	Proportion of Short pauses
short	long	0.32
short	short	0.39
long	short	0.11
long	long	0.06

[1]An earlier observation in Zvonick and Cummins (2003) that a similar relationship obtained between pauses and *following* phrases is probably an artifact of the idiosyncratic text used in that study

Fundamental Frequency Variability

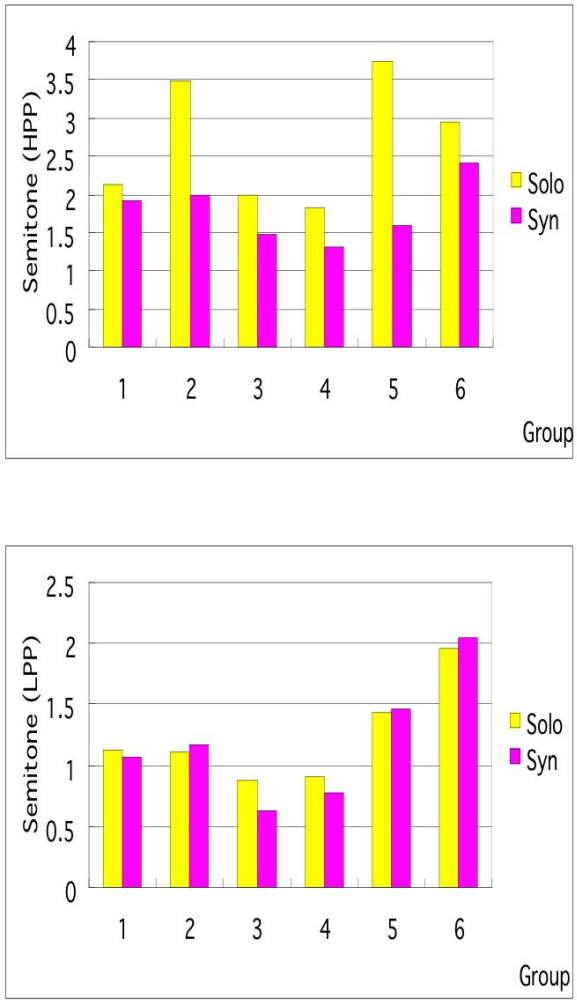


Figure 1: Difference in F0 peaks (HPP, top) and valleys (LPP, bottom) between speakers in a pair in solo and synchronous readings. All F0 values converted to semitones before analysis.

Given the above findings on phrasing and pauses, it seemed natural to examine the effect of speaking synchronously on other prosodic variables. To this end, we recorded six pairs of female speakers reading simple fairy tales (Wang and Cummins, 2003). We identified peaks and valleys in the intonation contour (HPP: High Point of Pitch, LPP: Low Point of Pitch), and looked to see whether these variables were affected by speaking synchronously. We found that the peaks were considerably more highly correlated across speakers within a pair in the synchronous condition (mean $r = 0.72$, s.d = 0.08) than in the solo condition (mean $r = 0.59$, s.d.=0.07). This did not hold for valleys (Synchronous: mean $r = 0.43$, s.d.=0.08; Solo: mean $r = 0.30$, s.d.=0.17). Figure 1 shows the differences in pitch between speakers of a

pair for both conditions. From the peaks (top panel), it can clearly be seen that there is a substantial reduction in inter-speaker differences in the synchronous condition, while the difference is slight or nonexistent for the valleys. This suggests that there might be a difference in the amount of free variability which speakers may employ in the absolute placement of H and L tones, a possibility which has been suggested independently elsewhere (Liberman and Pierrehumbert, 1984).

On Synchronization

How do speakers manage to synchronize so efficiently? One possibility which can be discounted already is that one speaker provides the lead, and the other follows. As mentioned above, we have yet to find a dyad in which a single leader could be identified. The very short lags between speakers also seem to preclude an explanation along these lines, as the typical lag of 40 ms is simply too short to allow perceptually guided correction while speaking.

Most mathematical models we have of synchronization are based on populations of oscillators (Glass and Mackey, 1988; Strogatz and Stewart, 1993). The mathematics of coupling among periodic sources is complex, but by and large tractable. Powerful predictive models have been constructed of such phenomena as juggling (Beek and Lewbel, 1995), heart cells (Mirollo and Strogatz, 1990), etc. Some have chosen to restrict the term 'synchronization' to the "adjustment of rhythms of oscillating objects due to their weak interaction" (Pikovsky et al., 2001, p. 8). Certainly, speech production is not oscillatory or periodic in anything but the loosest sense, and so the known mechanisms of entrainment among periodic sources can not be invoked here.

The answer, it seems to us, must lie in the shared knowledge speakers have of what is essential and what is redundant, or optional, in the modulation of the speech organs. Speakers of the same dialect must have control structures in common that govern the production of, and temporal relations among, the discrete units of speech. Little has yet been ascertained about the degree to which these putative control structures must coincide among such speakers. Certainly, many of the processes of diachronic change in language which have been described suggest that differences among speakers are not particularly rare, e.g. the age-related differences observed among speakers of Brazilian Portuguese by Major (1981). Nonetheless, the efficiency of communication dictates that most such structures must be shared among speakers. Although we do not have privileged access to the units and processes of speech production, speakers do seem to be able to modify their speech in direct response to the task demands, suggesting that the method of synchronous speech elicitation is a promising technique for tapping speakers' unconscious knowledge of the process of speaking.

An acknowledged limitation of the present is that much of the prosodic richness of spontaneous speech, specifically that associated with information management, speaker's attitudes, etc, is clearly not present in Synchronous Speech. The method requires the reading of a prepared text, and the additional constraint of synchrony places strict limits on the degree of personal interpretation and expression which a speaker can employ. Some of what is shorn away can correctly be considered to be meaningful prosodic structure. This limitation has an upside, however, as the timing which remains is still an immensely rich object of study, and those aspects of speech timing which are preserved (very many!) can be seen more clearly in the absence of the other additional sources of variation in speech.

The closest parallels to the demands of the Synchronous Speech task appear to be met in studies of synchronization among ensemble musicians (Rasch, 1979; Rasch, 1988), and the largely unstudied process of synchronization among dancers. In studying ensemble playing, Rasch (1979; 1988) used the standard deviation of differences in onset time of simultaneous notes in many voices as an index of asynchrony and noted typical values of 30 to 50 ms. The more direct measure of mean lag used in our studies of two voices at a time have provided values of approximately 40 ms.

Effect of Synchronization on Proportional Durations

An important question about the process of synchronization is whether it introduces artifacts into the temporal structure of speech, or conversely, whether it merely serves to reduce variability and reveal a shared understanding of temporal structure among speakers. Artifacts might be revealed in the systematic alteration of proportional durations, as would be the case if, e.g., unstressed syllables were found to be less reduced, and hence longer compared with stressed syllables. Any such systematic alteration of the durational properties of speech would severely limit the potential of Synchronous Speech to inform researchers about the properties of speech in the more general case.

As one way to investigate this, we here examine means and variances of a variety of interval ratios. By looking at ratios rather than durations, we better capture the relational properties of speech, and simultaneously avoid the difficult issue of rate normalization.

Methods

Readings of the first paragraph of the Rainbow Text were obtained from 27 pairs of speakers, as described in Cummins (2003). Each subject provided one reading alone and one with a co-speaker, obtained during a larger corpus collection exercise. The second sentence of the passage was chosen for detailed analysis. It reads "The rainbow is a division of white light into many beautiful colors". Reliably identifiable points in the waveform were chosen for measurement (stop releases, V-nasal transitions, etc). Figure 2 illustrates a representative set of measurement points for one recording.

Each variable studied was a ratio of two intervals, and comparisons were made of both mean values (using t-tests) and variability (F-test, one-sided, with the hypothesis of reduced variability in Synchronous Speech). Each

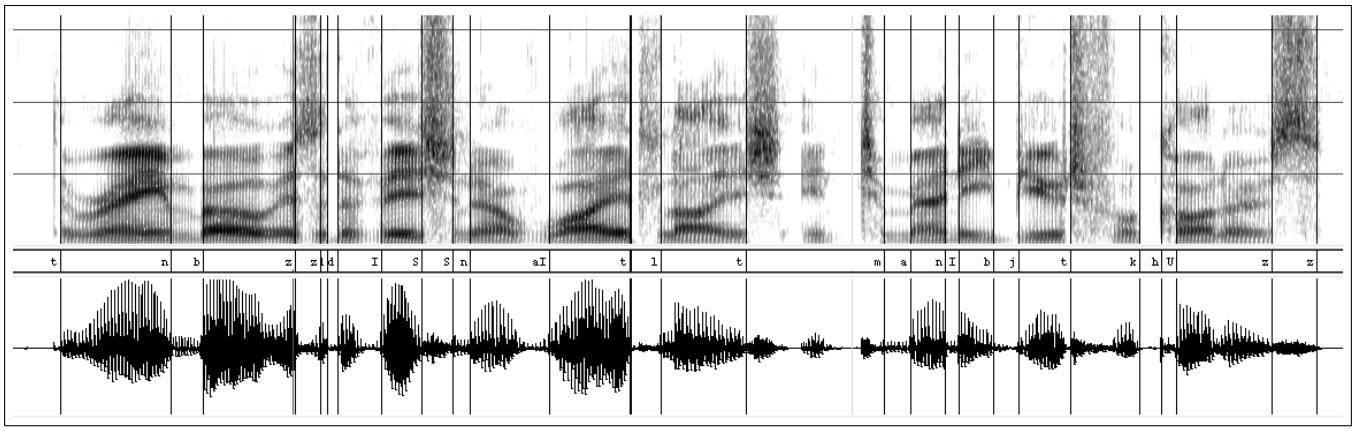


Figure 2: Measurement points for a single recording.

Table 2: Comparison of interval ratios in Synchronous Speech and Solo Speech. Intervals are taken from the sentence "The rainbow is a division of white light into many beautiful colors". Segments at measurement points are capitalized and intervals used are in bold face.

No.	Variable Type	Interval 1	Interval 2	t (df)	$p(t)$	F (df)	$p(F)$
1	juncture/phrase	ligh**T into M**any	rai**Nbow...ighT**	0.15 (93)	n.s.	1.67 (50,45)	< 0.05
2	juncture/phrase[1]	ligh**T into M**any	**Many...colorS**	0.36 (90)	n.s.	2.0 (50,45)	< 0.01
3	phrase/phrase	**Many...colorS**	rai**Nbow...ighT**	-0.27 (89)	n.s.	2.16 (50,45)	< 0.01
4	unstressed/stressed syllables	**NY**	**MA**	1.66 (92)	n.s.	1.06 (47,45)	n.s.
5	onset segment/word	**C**olors	**ColorS**	0.7 (94)	n.s.	1.04 (50,45)	n.s.
6	stressed vowel/word	div**I**sion	**DivisioN**	-0.56 (79)	n.s.	0.56 (43,42)	n.s.

ratio was expressed as the smaller value divided by the larger, and distributions were checked visually for approximate normality. Adjustment to degrees of freedom as appropriate for distributions of unequal variance was made using the Welch approximation.

Previous results had demonstrated that macroscopic phrasing (the division of an utterance into intonation phrases, the placement and duration of pauses) was significantly less variable in Synchronous Speech. No analysis of the durations of shorter intervals had yet been done. The possibility that duration ratios might be significantly different in Synchronous Speech was of interest, as this would suggest that the process of synchronization introduces artifacts into speech timing, and speech so obtained cannot be considered as unproblematically related to conventional speech. On the other hand, it was of interest to see whether the previous indications of reduced inter-speaker variability would be found with shorter, intra-phrasal, units also.

Results

Major Syntactic Juncture: The sentence studied contains one major syntactic juncture, between "white light" and "into many". The most reliably measurable interval spanning this juncture was delimited by the obstruent occlusion at the end of "light" and the first nasal onset of "many". We examined the ratio of this interval to the duration of each of the surrounding phrases (From the onset of "many" to the fricative onset in "colors" and from the nasal of the initial "rainbow" to the obstruent closure in "light"). Rows 1 and 2 of Table 2 show that the relative duration of the interval spanning the juncture is similar across conditions, whether one takes the preceding or the following phrase as a referent, but the variability of this ratio is substantially reduced in Synchronous Speech.

Phrase Length: The durations of the two major phrases ("The rainbow is a division of white light" and "into many beautiful colors") were compared. No difference in the ratios was discernible, but the variability of the ratio was greatly reduced in Synchronous Speech (Row 3, Table 2). This result may be attributable to a greater constancy of speaking rate in the synchronous condition.

Stressed and Unstressed Syllables: The word "many" provides unambiguous measurement points which make a comparison of the duration ratio of an unstressed to a stressed syllable within the same word possible. Row 4 of Table 2 provides the results of the analysis in which no significant differences in either ratios or ratio variability was found across conditions.

Segment 1: Onset. The length of the initial consonant (closure to voicing onset) in "colors" as a propor-

tion of the word length (/k/ closure to /z/ onset) was examined. Row 5 in Table 2 shows that mean ratio was not different across conditions, nor was ratio variability different in Synchronous Speech.

Segment 2: Vowel: The length of the stressed vowel /I/ in "division", expressed as a proportion of the word duration (stop closure to nasal release) was also studied. Again, neither ratio means nor variability was significantly different across conditions.

Discussion

The variables examined in the present study spanned a range of temporal scales and phonological structures. Those variables which were most directly related to macroscopic temporal structure (i.e. phrasing) all showed significant reduction in Synchronous Speech without any discernible change in mean values. The variables which describe smaller intervals showed no effects. In none of these cases was the proportional duration indexed by the ratio found to differ in its mean value between Synchronous Speech and solo speech, nor was the variability affected by speaking condition.

These results accord well with previous findings on Synchronous Speech and suggest areas for further study. No evidence has yet been found that speaking synchronously produced durational artifacts. The only properties of Synchronous Speech which have been reliably identified to date are a demonstrable increase in the consistency of global timing and phrasing (including intonation) across speakers. Those variables which exhibit substantially reduced variability in the Synchronous Speech condition are those most closely tied to timing at a global level, in which whole phrases are coordinated with respect to one another. Neither the unstressed/stressed syllable comparison, nor the segmental variables exhibited any difference in mean value or variability, suggesting that at a finer timescale there is little if any change to speakers' timing when speaking in synchrony with another person.

Some of the reduction in variability which is observed may be due to the forced maintenance of a constant speech rate. The indexing of speech rate is a notoriously difficult problem. Crude indices such as articulation rate, measured in number of syllables or segments per second, do little to match speakers' intuitions of a continuous abstract 'rate' of speaking. The constraint of speaking together with another speaker places severe limits on the freedom of the speaker to continuously modulate this abstract speaking rate, as any modification must be predictable for the co-speaker also.

[2]Alone among the distributions used herein, the ratios of the juncture to the second phrase were not normally distributed in the solo condition, but were skewed right. A Wilcox rank sum test substantiated the findings of the parametric test.

Further Exploitation of Synchronous Speech

The earlier examples of studies of pauses and intonation illustrate two different ways in which Synchronous Speech offers a novel approach to the analysis of variability in speech. In the former case, Synchronous Speech provided cleaner data than solo read speech, allowing the identification of temporal regularities which would otherwise be obscured. Synchronization among speakers is a simple and effective way of obtaining high-quality spoken data which is stripped of inessential sources of variability. This interpretation of the character of Synchronous Speech is supported by the durational measurements reported here for the first time. No difference in the fine structure of speech was observed, but variability associated with macroscopic timing was reduced. Future work should also examine finer gradations in the prosodic hierarchy: are there changes at levels between the intonational phrase and the syllable?

In the intonation study, the difference between solo speech and Synchronous Speech was itself a source of information about essential variability. By comparing Synchronous Speech with solo speech, we obtain a partition of variability into essential and inessential parts. It is tempting to associate the essential variability, preserved in Synchronous Speech, with linguistic sources, and inessential variability, absent in Synchronous Speech, with para- and non-linguistic sources, but this step is probably premature at this stage. To gauge the reliability of this attribution of the source of variation to linguistic or nonlinguistic origins will require further targeted research. However, the prospect of obtaining this partition potentially opens up new avenues of exploration for both kinds of variation.

Important information about the quality of Synchronous Speech will come from testing to see if subjects can perceive artifacts in Synchronous Speech, or indeed distinguish it from normal speech in perception tests. This work is ongoing.

One tantalizing possibility is the identification of parameters of free variability which might be exploited in the synthesis of expressive or characterful voices. Synthetic voices are bland, while carefully tailored voices which convey some sense of personality (emotion, expression) are laborious to construct. Adding random variation to synthesis parameters does nothing but reduce intelligibility. However an analysis of the properties of Synchronous Speech and a comparison with solo speech may inform voice designers about those parameters which they are relatively free to vary for expressive purposes. For example, the above study on intonation strongly suggested that an excitable voice might result from an expanded dynamic range of intonation in which the high targets are modified, but not the low targets.

Obtaining synchronous speech is not difficult. All studies of the properties of Synchronous Speech to date have suggested that the principal effect of the constraint of speaking together is to reduce idiosyncratic variability, leaving the essential quality of the speech untouched. This seems to offer two things to the experimental pho-

netician. Firstly, it provides an easy route to cleaner (less variable) data, for studies in which non-linguistic variability is unwanted. Secondly, it may provide a principled manner of partitioning variability, so that intrinsic variability which cannot be voluntarily avoided is retained, while superfluous variability is removed, thus allowing the differentiation of two kinds of variability in speech.

Acknowledgements

This work has been done with the help of Elena Zvonik and Bei Wang. Funding was provided by an Irish Higher Education Authority grant for collaborative research between Irish Universities and Colleges and Media lab Europe.

References

Beek, P. J. and Lewbel, A. (1995). The science of juggling. *Scientific American*, pages 92–97.

Cummins, F. (2001). Prosodic characteristics of synchronous speech. In Puppel, S. and Demenko, G., editors, *Prosody 2000: Speech Recognition and Synthesis*, pages 45–49, Krakow, Poland. Adam Mickiewicz University.

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, 3(1):7–11.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31(2):139–148.

Cummins, F. and Roy, D. (2001). Using synchronous speech to minimize variability in pause placement. In *Proceedings of the Institute of Acoustics*, volume 23 (3), pages 201–206, Stratford-upon-Avon.

Glass, L. and Mackey, M. C. (1988). *From Clocks to Chaos*. Princeton University Press, Princeton, NJ.

Heijink, H., Desain, P., Honing, H., and Windsor, W. (2000). Music representation–make me a match: An evaluation of different approaches to score-performance matching. *Computer Music Journal*, 24(1):43–56.

Liberman, M. Y. and Pierrehumbert, J. B. (1984). Intonational invariance under changes in pitch range and length. In Aronoff, M. and Oehrle, R. T., editors, *Language sound structure: studies in phonology presented to Morris Halle*, pages 157–233. MIT Press.

Major, R. C. (1981). Stress-timing in Brazilian Portuguese. *Journal of Phonetics*, 9:343–351.

Mirollo, R. E. and Strogatz, S. H. (1990). Synchronization of pulse-coupled biological oscillators. *SIAM Journal of Applied Mathematics*, 50(6):1645–1662.

Pikovsky, A., Rosenblum, M., and Kurths, J. (2001). *Synchronization: A universal concept in nonlinear sciences*. Number 12 in Cambridge Nonlinear Science Series. CUP.

Rasch, R. A. (1979). Synchronization in performed ensemble music. *Acustica*, 43:121–131.

Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In Sloboda, J. A., editor, *Generative Processes in Music*, pages 70–90. Clarendon Press, Oxford.

Strogatz, S. H. and Stewart, I. (1993). Coupled oscillators and biological synchronization. *Scientific American*, pages 102–109.

Wang, B. and Cummins, F. (2003). Intonation contour in synchronous speech. *Journal of the Acoustical Society of America*, 114(4(2)):2397.

Zvonik, E. and Cummins, F. (2002). Pause duration and variability in read texts. In *Proc. ICSLP*, pages 1109–1112, Denver, CO.

Zvonik, E. and Cummins, F. (2003). The effect of surrounding phrase lengths on pause duration. In *Proceedings of EUROSPEECH*, Geneva, CH. to appear.

Active and Passive Statistical Learning: Exploring the Role of Feedback in Artificial Grammar Learning and Language

Rick Dale (rad28@cornell.edu) **Morten H. Christiansen (mhc27@cornell.edu)**

Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

Language is immersed in a rich and active environment. One general dimension of that environment, feedback, may contribute greatly to learning language structure. Artificial-grammar learning offers an experimental means of exploring different kinds of potential feedback. In this paper, two experiments sought to investigate the role of feedback in an artificial-grammar learning task designed to resemble some aspects of language acquisition. An artificial language composed of auditory nonsense syllables and an accompanying visual semantics were created. Participants faced the task of mapping a sample sentence to a visual semantic scene. Results indicated that feedback is highly useful, allows participants to reach a high level of competence in the language, and also helps the acquisition of detailed aspects of the artificial grammar. Implications for language acquisition are discussed, and future directions considered.

Introduction

That humans can learn without any direct feedback has been well established for decades. From basic information extraction in perceptual processes (Gibson & Gibson, 1955), to social facilitation of a choice task (Bandura & Mischel, 1965), it seems that learning can occur passively and observationally across multiple levels of cognitive complexity.

One particular area of research with similar findings has been that of implicit learning or artificial grammar learning (AGL; Reber, 1967), in which subjects become sensitive to the regularities of a simple artificial grammar through passive exposure to sample sentences. A considerable amount of this research has been directed towards uncovering the mechanisms of this learning (e.g., Reber, 1967; Reber & Lewis, 1977; Vokey & Brooks, 1992; Redington & Chater, 1996; Cleeremans, Destrebecqz, & Boyer, 1998; Pathos & Bailey, 2000).

Learning through passive exposure to these grammars, however, is usually defined as performance at above-chance levels. Therefore, to gain further insight into language acquisition, theoretical and empirical bridges are needed between what may be called *passive structural learning* in these cases and the natural world, in which a learner acquires a firm competence with sequential structures in a meaningful, interactive context (e.g., see Berry, 1991, for an investigation of action in learning a probabilistic system). In pursuit of this, some research has been guided by questions about the possible connections between this kind of learning and real-world tasks, in particular language acquisition (e.g., Saffran, Aslin & Newport, 1996; Christiansen & Ellefson, 2002; Lupyan, 2002; Saffran, 2003). AGL can be used for studying the kinds of structural regularities that children discover while learning language (Saffran, 2003). The goal of this work has mostly involved exploring learning under passive observational exposure. Indeed, these experiments have demonstrated the richness of statistical learning under such circumstances.

However, language acquisition does not take place in a social vacuum. Instead, children are acquiring their native language while interacting with both people and things in the environment (e.g., Snow, 1999; Chouinard & Clark, 2003; Moerk, 1992; and even before two-word production; see Tomasello, 2003, for a review). In this context, and others in the natural world, relevant sequential behavioral structure *has a function* or *serves a purpose*, socially or otherwise, and its acquisition is immersed in this interactive context. What kind of information in the environment, and possible mechanisms in the learner, can supplement passive exposure to sequential structure in order to obtain a competence over what is to be learned? This paper presents a first step toward identifying one such dimension of learning. By using an AGL procedure, we explore the role of one kind of feedback that may be present in language acquisition.

We first offer a brief summary and review of this source of feedback in language acquisition. The potential for exploring this dimension is then presented in two experiments, demonstrating how an interactive task can bring a learner to a strong level of competence. In addition, we demonstrate that detailed aspects of an artificial grammar can be acquired in the context of feedback. We end with a discussion of implications, especially in view of language acquisition, and future directions this research may take.

Feedback in Language and AGL

Although the child may not be told explicitly that a given utterance or word is incorrect (also referred to as the lack of "negative feedback"; Saxton, 1997), the child does get other types of evidence or feedback.[1] For example, a mother may

[1] For simplicity, we do not consider the difference between negative feedback and negative evidence, though the distinction is important and may be explored by the experimental means presented here. See Saxton, 1997, for further discussion.

ask her child to pick up a particular toy, say a little plastic pig, from among several other toys. When the child successfully picks up the right toy, the mother may emphatically repeat the name of the target object: *The pig! Yes, that's the pig*. Once the child chooses the right toy, the mother repeats the label (e.g., *the pig*) and thus provides positive feedback on the child's correct mapping of the linguistic label to the appropriate object. Although there is considerable and continuing debate on the cultural variability of such practices (see Lieven, 1994, for a review and discussion), it is nevertheless possible that feedback of this nature may be present and useful in language acquisition (e.g., see Peters and Boggs, 1986, for a discussion of interactional routines across cultures).

Here we take a first step toward assessing the potential usefulness of such feedback in an AGL task meant to model the learning of sequential structure and how it maps to the non-linguistic world – a task not unlike what the child faces.

It should be noted that the role of feedback in language acquisition is highly controversial (see, for example, Morgan, Bonamo & Travis, 1995; Valian, 1999; Moerk, 2000; Saxton, 1997, 2000). It has perhaps for this reason not been extensively investigated in AGL research, where the focus has been on training techniques that largely parallel the kind of passive input considered central during language acquisition. Nevertheless, the role of feedback is widely acknowledged in such areas as skill acquisition (Moerk, 1992), learning theory (Rescorla, 1968), and reinforcement-learning models (Sutton & Barto, 1998).

There are therefore two primary objectives of the following experiments. A basic empirical objective is to consider the influence of feedback on AGL in a training procedure that resembles a natural-world context. To meet this goal, an experimental paradigm has been designed to resemble a kind of task faced by the child during language acquisition, adapted from Lupyan (2002; also, see Billman, 1989 and Morgan, Meier & Newport, 1987 for similar techniques).

Another objective is primarily theoretical: How does learning sequential structure get immersed in an interactive context and lead to competence? These experiments approach one aspect of an answer by considering how interactive feedback in a sequential learning task might bring the learner to a competent level of performance.

Experiment 1

This experiment is a first demonstration of the influence of feedback on learning an artificial grammar. The conditions in this experiment focus on the consistency of forms of feedback, and the extent to which the feedback is a salient, meaningful aspect of the learning task.

Method

Participants 51 college students were recruited for extra credit. Participation required approximately 20 minutes.

Materials A simple artificial grammar was created for the experiment, illustrated in Figure 1. Each category (e.g., N, *noun*) was instantiated by a set of nonsense syllables (e.g., *voop* or *jux;* see Table 1).

$$S \rightarrow N_1 \quad VP$$

$$VP \rightarrow \left\{ \begin{array}{l} \text{intransitive-V} \\ N_2 \quad \text{transitive-V} \\ N_2 \quad N_3 \quad \text{ditransitive-V} \end{array} \right\}$$

Figure 1: The artificial grammar

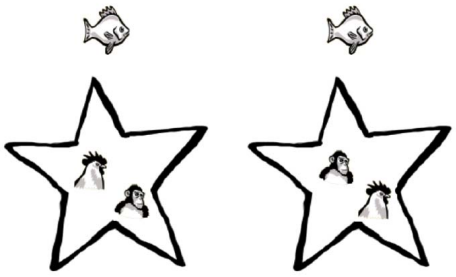


Figure 2: An example stimulus from one trial

An elementary visual "semantics" was created for this language. Each noun was randomly assigned an animal referent, and each verb had as its "meaning" a simple shape. Each nonsense syllable in the language had a referent of this kind in the visual semantics (Fig. 2).

Although the extent to which the visual scene contains a "subject" or "object" or "verb" is abstract, the language and its semantics are meant to capture structure-world correspondences not entirely unlike what might be seen in natural language structure.

Fifty random sentences were constructed for the experiment, and an incorrect visual semantic scene for each sentence was created (see Figs. 3 and 4). This incorrect scene was paired, as a foil, with the correct scene in training, as described below.

Table 1: Classes and assigned syllables

class	sounds	class	sounds
N	*kav*	Intran V	*voop*
	jux		*poox*
	ruj	Tran V	*sook*
	hep		*lem*
	pel	Ditran V	*rud*
	hes		*jove*

Procedure In every trial, participants saw two visual semantic scenes side by side then heard a sample sentence from the grammar. Their task was to select the appropriate visual semantics for the sentence heard. The task therefore involved learning the sequential structure of the grammar, and learning to map each sound to its semantic animal or shape.

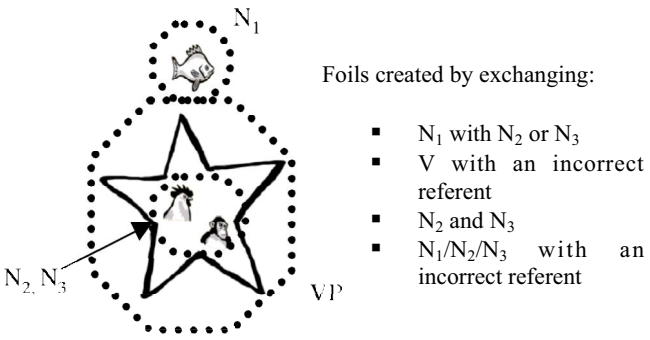


Figure 3: The structure of the visual scene and foils

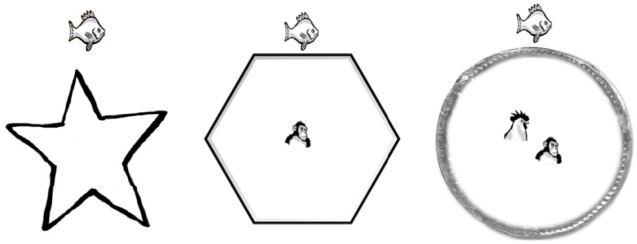


Figure 4: Example sentences

A *positive feedback event* was defined as a repetition of the sentence when the participants selected the appropriate visual semantic scene.

Two feedback conditions were investigated. Some subjects received *only* consistent feedback, occurring with 50% probability on any *correct* trial. Other subjects received 40% *random* repetitions, not contingent on the correctness of their selection (these probabilities were chosen so that all participants heard approximately the same number of repetitions).

Two further conditions were defined in terms of the kinds of instructions provided to subjects. In one condition, subjects were not informed about the meaningfulness of the repetitions (as positive feedback); in a second condition, subjects were explicitly informed that feedback would occur.

Out of the four possible subject groups, three were used in the experiment. One group received no instruction about the feedback but received it consistently. A second received no instruction, but the feedback occurred randomly. A third group of subjects received both consistent feedback and instructions about the presence of feedback during training.

Performance on the final 10 items of training served as the measure of learning. These items were new to the subjects. This permitted observation of performance in a continuous learning task without interruption. There was therefore no distinction between training and testing.

Results

No main effect for condition was found (corrected $F(2, 50) = 1.65$, $p = .21$). However, due to the probabilistic nature of the training phase, an additional planned regression analysis on each condition was conducted (because, by chance, some subjects may experience less consistent positive feedback than others). This was meant to investigate the number of actual feedback events experienced during the first 40 trials of training, and how it might predict performance on the final 10 items.

The only condition that produced a reliable predictive relationship was that in which subjects received information about the presence of feedback ($r = .65$, $p < .05$). Although the consistent feedback without instruction condition had a positive slope, the coefficient was not significant ($r = .28$, $p = .26$).

Discussion

This preliminary experiment offers some important observations. First, inconsistent feedback present in training did indeed stultify learning, even when the subjects were not certain about the significance of the sentence repetitions. We may tentatively contend that even contingent events in

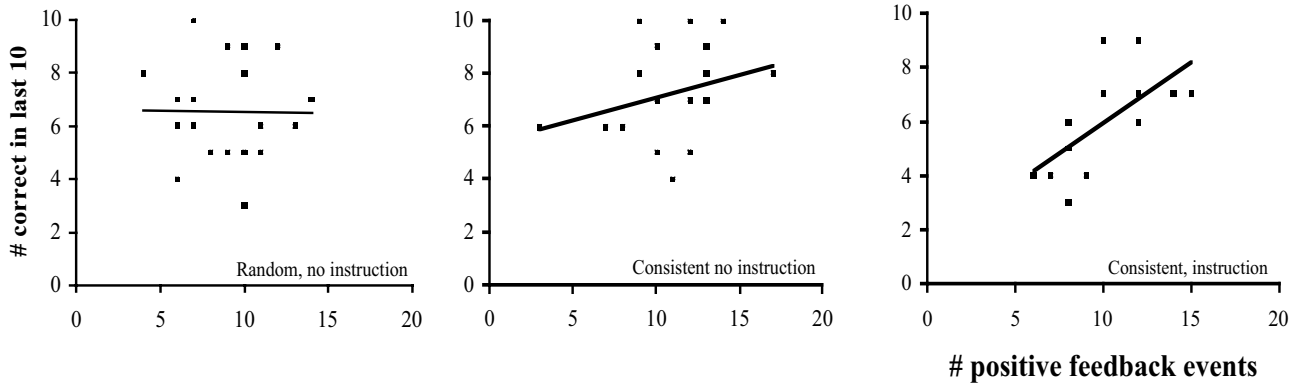


Figure 5: Regression analyses of different conditions. Each point represents a subject.

the learning environment can help or hinder learning sequential structure. It was *not* the case that learners simply ignored the repetitions and extracted the sequential invariants across the training trials.

Second, when participants were informed about the presence of feedback, the repetitions served as *significant* elements of acquiring the structure. It appears that subject performance became highly reliant upon even occasional, contingent repetitions of sentences as positive feedback, especially when the feedback was made meaningful to subjects. There is evidence that perceiving the import of such events may have an important influence on language acquisition (Saxton, 1997; Tomasello, 2003).

Despite these positive results, the learning that took place in the experiment hardly exhibits competence of the kind described in the introductory discussion. We therefore conducted a second experiment to address this and other questions. First, we enhanced the salience of the feedback event by changing the training environment. Second, we devised a separate test phase to explore the learning of specific kinds of structures in the grammar. Finally, we tracked learning of the grammar over time to observe the effect of feedback across training.

Experiment 2

Experiment 2 changed the nature of the feedback event to render it more salient. This involved not merely repeating the sentence, but also changing the visual environment selected by the participants. In addition, we explored how participants learned different aspects of the grammar, such as the abstract verb-argument structure.

Method

Participants 34 college-age participants were recruited for extra credit. Participation required approximately 20 minutes.

Materials The artificial grammar used was the same as in the first experiment. We created an additional 30 sentences to be used in a test phase without feedback. The paired incorrect visual scenes in these test items were constructed so as to sample across all possible grammatical errors. These included exchanging nominal shapes with an incorrect shape, inverting nominal shapes, and exchanging verbal shapes with incorrect shapes (see Fig. 3 for the kinds of foils used).

Once again, 50 sentences were presented randomly in a training phase. Feedback again was defined as a repetition of the spoken sentence.

Procedure Similarly to the first experiment, participants selected one of two visual scenes in response to a heard sentence of the grammar. This training once again consisted of 50 trials. Half the subjects received 60% feedback consistently, the other half hearing random feedback with 50% probability (these were selected once again so that all subjects heard approximately the same number of repetitions).

Given the results of the first experiment, it seems that salience of such feedback is a crucial property of using it in the task. To enhance this effect, we added a feature to the feedback event: When a correct visual scene was selected, the incorrect scene would be removed *and* the sentence would be repeated to the participant. This served to make these events as informative as possible to the participants. Also, this event may bear some resemblance to social interaction between the child and caregiver. When the child correctly interprets a lexical item, the caregiver may emphasize its referent object, thereby focusing the child's attention on it.

Following this training procedure, 30 trials were presented to participants *without* feedback in either condition. Performance on these 30 items served as the basic comparison between groups (consistent vs. random feedback), and item analyses allow us to investigate the role of feedback in acquiring more detailed aspects of the grammar (e.g., verb argument structure).

An additional control condition was conducted in which participants only experienced the test phase of the experiment.

Results

A main effect of condition (positive feedback, random feedback, control) was found ($F(2, 31) = 7.1$, $p < .01$). Subsequent comparisons among the groups indicated that only the positive feedback condition differed significantly from the random feedback and control groups ($p < .05$ in both cases; see Fig. 6). In fact, participants in the random feedback condition did not differ significantly from the control group ($p = .28$).

We further conducted item analyses within the positive and random feedback conditions to find wherein their performance differences lie (see Table 2). A repeated-measures ANOVA was conducted over the different kinds of items within subjects, and found that the primary differences in performance were in verb exchanges, the "subject" shape being exchanged, and a marginally significant result for identifying inversions in the argument structure of the verbal shapes.

By looking at the overall performance of participants, graphed over time, we get an interesting illustration of learning under the condition of consistent feedback (Fig. 7). The final 4 points include performance during the training stage.

Table 2: Number correct on different foils, and significance of the comparisons

Type of error in scene	Pos	Ran	Out of	p
Verbs different	5.2	3.8	6	< .05
Nouns different	3.8	3.8	5	.88
Sub exchanged w/ obj	12	8.8	14	< .01
Objects exchanged	3.2	2.4	5	.07

Discussion

These results further indicate that the salience of positive feedback in a sequential learning task of this kind can strongly influence performance. Participants in this task were performing almost perfectly in the positive feedback condition, even in the test phase, during which feedback was no longer issued.

Moreover, item analyses indicated that even subtle structure-world correspondences as the idealized "verb argument" structure in this artificial grammar was being learned more effectively under the condition of feedback.

General Discussion

Although we feel the current experiments hold considerable promise, they do have limitations. First, although they more closely resemble natural-world contexts than previous research, they are still quite simple. Future experiments will address this issue by incorporating an even more interactive experimental task. Second, the grammar itself is quite simple, and *mere* passive exposure may be sufficient to learn it. Experiments are currently being conducted that directly compare passive exposure to scene-sentence pairs and the active selection task used here.

These limitations notwithstanding, the experiments have provided a first step towards investigating how feedback in an interactive task can bring performance in AGL to a more competent level than typically observed. The language acquisition literature itself has been deeply involved in debate for decades about the nature of feedback and evidence to children. For example, one may argue that the issue of positive and negative feedback has been resolved since Brown and Hanlon (1970), who demonstrated quite clearly that commonplace conceptions of feedback to a language learner are incorrect. Nevertheless, many continue to tease apart the negative and positive function of different types of input to children (e.g., Saxton, 1997; Saxton, 2000; Chouinard & Clark, 2003).

The experiments here can contribute to this endeavor. They may offer empirical means by which different kinds of feedback and their effects can be investigated experimentally, albeit here in college-aged subjects. The technique could be modified for children, and many of its dimensions explored in experiments with both children and adults. Some have pursued similar techniques such as "human simulations" (e.g., Gillette, Gleitman, Gleitman & Lederer 1999; Snedecker, Gleitman & Brent, 1999). For example, Snedecker et al. (1999) used college-aged subjects to explore the role of ambient social and environmental input to support a noun bias during language acquisition. This idea is not unlike what is being argued here (see Snedecker et al. for an interesting exploration and discussion of feedback in word learning).

More importantly, these experiments are intended to support a perspective in "ecological" sequential learning, and particularly language learning, that sees the task facing a learner as an active and interactive one. We would contend that such learning cannot *only* involve passively extracting statistical regularities from different modalities. Instead, sequential structure in the natural world, linguistic or otherwise, is *used* in an *interactive* environment – these uses generate consequences in the environment that impinge upon a learner's expectations and help carry the learner into a world of meaningful sequential structure.

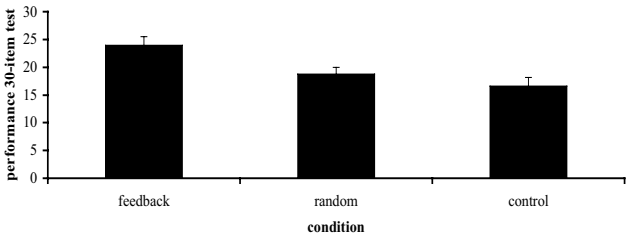


Figure 6: Performance by different groups

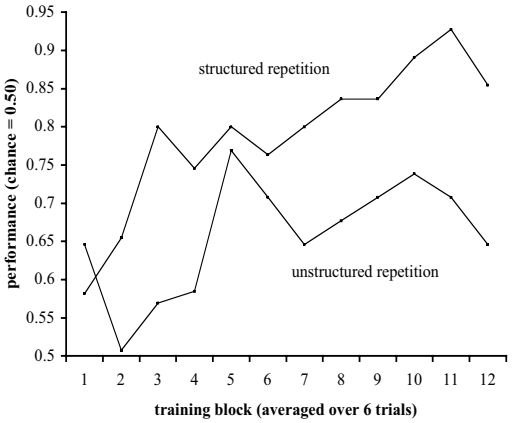


Figure 7: Performance over time averaged over trials

Acknowledgements

Thanks to Michael Spivey, Dima Amso, Emily Balcetis, Kevin Bath, Des Cheung, Joyce Ehrlinger, and Ben Hiles for helpful suggestions throughout this and related projects.

References

Bandura, A. & Mischel, W. (1965). Modification of self-imposed delay of reward through exposure to live and symbolic models. *Journal of Personality and Social Psychology, 2*, 698-705.

Berry, D.C. (1991). The role of action in implicit learning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43A*, 881-906.

Billman, D. (1989). Systems of correlations in rule and category learning: use of structured input in learning syntactic categories. *Language and Cognitive Processes, 4*, 127-155.

Brown, R. & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11-54). New York: Wiley.

Chouinard, M.M. & Clark, E.V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language, 30*, 637-669.

Christiansen, M.H. & Ellefson, M.R. (2002). Linguistic adaptation without linguistic constraints: The role of sequential learning in language evolution. In A. Wray (Ed.), *Transitions to language* (pp. 335-358). Oxford, UK: Oxford University Press.

Cleeremans, A., Destrebecqz, A. & Boyer, M. (1998). Implicit learning: news from the front. *Trends in Cognitive Sciences, 2*, 406-416.

Gibson, J.J. & Gibson, E.J. (1955). Perceptual learning: differentiation or enrichment? *Psychological Review, 62*, 32-41.

Gillette, J., Gleitman, L., Gleitman, H. & Lederer, A. (1999). Human simulation of vocabulary learning. *Cognition, 73*, 135-176.

Lieven, E. (1994). Crosslinguistic and crosscultural aspects of language addressed to children, *Input and interaction in language acquisition* (pp. 56-74): Cambridge University Press.

Lupyan, G. (2002). *Modeling syntactic devices: An exploration of language evolution from connectionist and memetic perspectives*. Cornell University, Ithaca, NY.

Moerk, E. (1992). *A first language taught and learned*. Baltimore, MD: Brookes Publishing.

Moerk, E. (2000). *The guided acquisition of first-language skills*. Westport, CT: Ablex.

Morgan, J.L., Meier, R.P. & Newport, E.L. (1987). Structural packaging in the input to language learning: Contributions of intonational and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19*, 498-550.

Morgan, J.L., Bonamo, K.M. & Travis, L.L. (1995). Negative evidence on negative evidence. *Developmental Psychology, 31*, 180-197.

Peters, A.M. & Boggs, S.T. (1987). Interactional routines as cultural influences upon language acquisition. In B. Schieffelin & E. Ochs (Eds.), *Language socialization across cultures. Studies in the social and cultural foundations of language, 3* (pp. 80-96). New York, NY: Cambridge University Press.

Pothos, E.M. & Bailey, T.D. (2000). The role of similarity in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 847-862.

Reber, A. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Behavior, 6*, 855-863.

Reber, A. & Lewis, S. (1977). Implicit learning: an analysis of the form and structure of tacit knowledge. *Cognition, 5*, 333-361.

Rescola, R.A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1-5.

Saffran, J.R., Aslin, R.N. & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Saffran, J.R. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110-114.

Saxton, M. (1997). The contrast theory of negative input. *Journal of Child Language, 24*, 139-161.

Saxton, M. (2000). Negative evidence and negative feedback: immediate effects on the grammaticality of child speech. *First Language, 20*, 221-252.

Snedecker, J., Gleitman, L. & Brent, M. (1999). The successes and failures of word-to-word mapping. In M. H. A. Greenhill, H. Littlefield & H. Walsh (Ed.), *Proceedings of the Twenty-third Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Snow, C.E. (1999). Social perspectives on the emergence of language. In B. MacWhinney (Ed.), *The emergence of language* (pp. 257-276). Mahwah, NJ: Erlbaum.

Sutton, S.S. & Barto, A.G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tomasello, M. (2003). *Constructing a language*. Cambridge, MA: Harvard University Press.

Valian, V. (1999). Input and language acquisition. In W. C. R. T. K. Bhatia (Ed.), *Handbook of child language acquisition* (pp. 497-530). New York: Academic Press.

Vokey, J.R. & Brooks, L.R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 328-344.

Semantic Inhibition due to Short-Term Retention of Prime Words: The Prime-Retention Effect and a Controlled Center-Surround Hypothesis

Eddy J. Davelaar (e.davelaar@bbk.ac.uk)
School of Psychology, Birkbeck, University of London
Malet Street, WC1E 7HX, London, UK
http://www.geocities.com/ejdavelaar

Abstract

The semantic priming effect is shown to be modulated by the instruction to maintain the prime word while conducting a lexical decision to the target. Specifically, the priming effect is absent in situations of prime-retention. An extended version of Dagenbach and Carr's (1994) Center-Surround hypothesis is proposed, in which the prime-retention effect, the absence of priming under prime-retention, can be accommodated. This extended hypothesis suggests that under prime-retention, activation remains centered on the prime, preventing unwanted spread of activation. This impact of the on-center-off-surround mechanism increases over time, making it sensitive to manipulations of stimulus duration.

Introduction

In the field of memory, semantic priming is a basic paradigm used to investigate the processes that inter-relate conceptual representations in long-term memory. The basic result is reduced lexical decision times or naming latencies and improved accuracy to words (i.e., targets), when they are preceded by a related word (i.e., the prime) relative to control. The semantic priming effect is such a robust phenomenon that even the absence of priming is theoretically relevant, as can be seen by the large literature on the prime-task effect (the absence of priming when attention is allocated away from the semantic level, but is still within the verbal domain) (see for a review, Maxfield, 1997). The implied assumption in the priming literature seems to be that the more attention the prime word receives, the more priming is expected. This paper focuses on the counterintuitive observation that the semantic priming effect is absent when in a standard priming paradigm the prime has to be reported *after* making a lexical decision to a target. This observation will be referred to as the *prime-retention effect*, as it is the active retention of the prime in short-term memory that modulates the priming effect.

Controlled Center-Surround Hypothesis

There are a number of theories and models of priming, but for the present purposes the spreading-activation view of semantic priming will be addressed to highlight the need for auxiliary mechanisms to accommodate the to-be-presented data. In the standard spreading-activation theory, concepts in semantic memory are linked together to form a semantic network, with the strength of the connection between two concepts representing the strength of association. Extensive investigations into the nature of the semantic priming effect has led to the view that the semantic priming effect is due to a fast-acting automatic process and a slow-acting controlled process (Neely, 1976, 1977, 1991). By varying the stimulus-onset-asynchrony (SOA) between the prime and the target, the relative contributions of these processes can be modulated. It is therefore assumed that with short SOAs the priming effect is predominantly due to automatic processes. However, this assumption has been challenged by behavioural and neuroimaging studies that show context effects at short SOAs (e.g., Mummery, Shallice & Price, 1999; Smith, Besner & Miyoshi, 1994).

One theory that specifically addresses the possibility of controlling the spread of activation is the Center-Surround hypothesis by Dagenbach and Carr (1994). In a nutshell, the hypothesis states that there exists a mechanism that facilitates "the semantic code on which it is focused or *centered* while inhibiting *surrounding* codes, codes that are similar to but different from the desired code and are competing with it for retrieval" (p.328, italicised words were between quotes in original). Dagenbach and Carr's work was mainly focused on priming effects found at the threshold of subjective and objective awareness and was applied quite successfully in a model of negative priming (Houghton & Tipper, 1994). However, in the standard supra-threshold priming paradigm, the data does not seem to demand such on-center-off-surround mechanism. This may be because the prime word does not have to be actively maintained. An on-center-off-surround mechanism would be necessary under conditions of prime-retention. For example, when a task requires focusing on a particular word, the increased activation to that word would lead to more spread of activation, which would in turn compromise the attentional focus on the word, due to the now-activated distractors. Intuitively, we are able to focus on the word 'doctor' for several seconds without strongly activating related concepts like 'nurse', 'patient', 'hospital', 'medicine' and so on. Besides preventing a situation where the whole lexicon becomes activated, inhibitory mechanisms seem particularly relevant in situations of short-term retention where a robust focus is necessary (e.g., Grossberg, 1978).

The view that will be pursued here is that there exists a trade-off between prime-activation and activation-spread, of which the balance depends on the task requirements. This hypothesis will be referred to as the Controlled Center-Surround hypothesis, implying that a controlled effort (i.e., deliberate active maintenance) needs to be made in order to observe the on-center-off-surround mechanism at supra-

threshold SOAs. In the prime-retention paradigm used here, the participant is presented with the prime and has to maintain it while making a lexical decision to a target. The Controlled Center-Surround hypothesis would predict that normal priming effects are found when the prime need not be retained (as the on-center-off-surround mechanism is not fully operational), while in the retention condition the off-surround component nullifies (or even reverses) the priming effect.

Before presenting the experiments that were designed to address the Controlled Center-Surround hypothesis, the next section will highlight two earlier reports that presented hints of a *prime-retention effect*.

Earlier reports

A *prime-retention effect*, the absence of a priming effect when the prime is actively maintained during lexical decision, can be observed in reports from at least two research groups (Fischler & Goodman, 1978; Henik, Friedrich, Tzelgov & Tramer, 1994). In a study by Fischler and Goodman (1978), participants were tested on a masked priming paradigm in which the prime was presented very briefly (50 ms) with a visual mask preceding and succeeding it. They asked participants to report the prime word after making a lexical decision to the target string. Participants were able to report the prime in about 50 % of the trials. Semantic priming was only found when the prime could *not* be reported; the priming effect was absent when participants could correctly report the prime word. A second example of the *prime-retention effect* can be found in one of the conditions in a study by Henik, Friedrich, Tzelgov and Tramer (1994). These authors were interested in the time-course of the prime-task effect (the finding that the priming effect is eliminated when the prime word is processed on a non-semantic level). Participants had to read the prime out loud and make a lexical decision to a target word. In one particular condition (in their experiment 3), the SOA was relatively short (240 ms) and therefore participants had to report the prime word *after* the lexical decision was made, thus actively maintaining it during the lexical decision. No priming effect was found in this condition. Henik et. al. (1994) explained the lack of priming in terms of the prime being processed at a shallow (e.g., phonological) level and thereby preventing resources to be allocated to the semantic level. However, this idea was not elaborated further.

The results by Fischler and Goodman (1978) and Henik, et. al. (1994) suggest at the very least that maintaining the prime modulates the priming effect and that the underlying mechanism may be inhibitory in nature. However, the results indicating this were tangential to the main focus of their investigations and did not receive enough attention. Therefore it is possible that their results, indicating a *prime-retention effect*, may have been a chance-finding. For example, in Fischler and Goodman's (1978) study, the results strongly depended on the erroneous recall performance of the participants (50 % error rate), making the data sensitive to participants' idiosyncratic biases (see for discussion, Holender, 1986). In addition, the results obtained by Henik, et. al. (1994) were not replicated without participants completing several other experimental conditions, which could have led to carry-over effects.

Experiments

Here, three experiments are reported that were specifically designed to address the *prime-retention effect*. The experiments involved two blocks of prime-target pairs in a standard lexical decision paradigm. The first block was always the control condition, in which participants did not need to maintain the prime word. In the second block, participants were required to maintain the prime and give a verbal report (in Experiment 1) or recognise it from four alternatives (in Experiment 2 and 3) after the lexical decision. In order to assess whether the prime has been processed on the semantic level, the association strength between prime and target was taken into account in the analyses. Any modulation with associative strength would counter an explanation based purely on shallow non-semantic processing.

According to the Controlled Center-Surround hypothesis, in the control condition, normal priming effects are expected for strongly and weakly related targets at both short and long SOAs, as the activation of the prime is allowed to decay after presentation. However, in the retention condition, it is expected that only strongly related targets show priming effects at both SOAs (controlled on-center), but that weakly related targets show less or even no priming effect (controlled off-surround).

Experiment 1

Participants. Twenty volunteers from the University of London participated in the experiment in exchange for £5. All participants had English as their first language, were right-handed and had normal or corrected-to-normal vision.

Design. The experiment conformed to a 3 x 2 within-subject design, with Relatedness (unrelated, low-related, high-related) and Retention as independent variables. Lexical decision times and accuracy were measured.

Materials. Eighty-four word pairs were selected from the MRC Psycholinguistic Database (Coltheart, 1981). The word pairs had a word frequency ranging from 10 to 660 per million (Kucera & Francis, 1967). The association strengths between the prime and target word ranged from 12.5 to 73.8 (M=34.5; Moss & Older, 1996). A median split divided the targets in the high and low association trials. All words and pseudohomophones were one syllable long. Unrelated trials were formed by rearranging the related pairs. Each participant saw each word only once, but target words rotated across participants in all conditions.

Apparatus. The experiment was run on an IBM-compatible PC using Micro Experimental Laboratory (MEL) Professional software (Schneider, 1995). Letter size was approximately 0.5 cm and average viewing distance was about 50 cm.

Procedure. Participants were given the instructions on the screen as well as verbally by the experimenter. The experiment had a total of 168 trials grouped into two blocks; the 'no retention' block was always followed by the retention block. After the instructions for each block, subjects practised 8 trials. On each trial, a fixation stimulus was presented for one second in the center of a computer screen, followed by a word in lowercase white letters that remained for one second. After a 250 ms interval (blank screen) a target was presented (in uppercase yellow letters) that remained on the screen until a lexical decision was made. In the control condition, the next trial started after a 500 ms delay, whereas in the retention condition a question mark prompted the participant to recall the prime word. The experimenter recorded the recall. Participants got feedback whenever an error was made.

Results and discussion

The mean median reaction times and error rates for all conditions are presented in Table 1. Performance on naming the prime was at ceiling (100% correct), discounting an explanation based on some form of speed-accuracy trade-off. Because of the large differences in standard deviations between the control and the retention condition, log-transformed RTs were used in the analyses with the Retention-variable, while untransformed RTs were used in the pairwise within-block comparisons.

An overall ANOVA on the lexical decision times revealed a main effect of Retention [$F(1,19)=40.35$, $MSe=0.04$, $p<.001$] and a marginal effect of Relatedness [$F(2,38)=2.55$, $MSe=0.02$, $p=.091$]. The interaction did not reach significance. A similar ANOVA on the response accuracy only revealed a marginal effect of Relatedness [$F(2,38)=2.64$, $MSe=0.001$, $p=.084$].

Pairwise comparisons were conducted to specifically address the predictions made by the Controlled Center-Surround hypothesis. This analysis revealed that priming effects were only obtained for the strongly related prime-target pairs in the no-retention control condition both for RTs [$t(19)=2.48$, $p<.05$] and error rates [$t(19)=2.42$, $p<.05$].

Experiment 1 replicates the failure to obtain a priming effect when the prime word needs to be retained. Although several methodology-related explanations could be given for the absence of priming, the mere observation of a lack of priming due to an experimental manipulation begs further inquiry. One possibility for the lack of priming in the retention condition could be that participants were preparing the articulatory response while making the lexical decision. This could lead to a form of response interference, where both verbal and manual responses are prepared and executed. Therefore, Experiments 2 and 3 employed a recognition task on the prime word instead of a verbal response. It was hoped that this would 'clean up' the processes during the lexical decision.

Experiment 2

Participants. Twenty-four volunteers from the University of London participated in the experiment. All participants had English as their first language, were right-handed and had normal or corrected-to-normal vision.

Design. The experiment conformed to a 3 x 2 within-subject design, with Relatedness (unrelated, low-related, high-related) and Retention as independent variables. Lexical decision times and accuracy were measured.

Materials. 112 word pairs were selected from the MRC Psycholinguistic Database (Coltheart, 1981). The word pairs had a word frequency ranging from 10 to 660 per million (Kucera & Francis, 1967). The association strengths between the prime and target word ranged from 5.4 to 66.7 ($M=34.5$; Moss & Older, 1996). A median split divided the words in high and low association-trials. All words and pseudohomophones were one syllable long. Unrelated trials were formed by rearranging the related pairs. Each participant saw each word only once, but target words rotated across participants in all conditions.

Apparatus. The apparatus as in Experiment 1 was used.

Procedure. Participants were given the instructions on the screen as well as verbally by the experimenter. The experiment had a total of 224 trials grouped into two blocks; the 'no retention' block was always followed by the retention block. After the instructions for each block, subjects practised 8 trials. On each trial, a fixation stimulus was presented for one second in the centre of a computer screen, followed by a word in lowercase white letters that remained for one second. After a 250 ms interval (blank screen) a target was presented (in uppercase yellow letters) that remained on the screen until a lexical decision was made. In the control condition, the next trial started after a 500 ms delay, whereas in the retention condition a list of four words appeared (the prime and three distractors) and the participant had to indicate by pressing one of four keys which one was the prime word. Participants got feedback whenever an error was made.

Results and discussion

The mean median reaction times and error rates for all conditions are presented in Table 1. Performance on recognising the prime was at ceiling (99% correct) and did not show an effect of Relatedness.

An overall ANOVA on the (log-transformed) lexical decision times revealed a main effect of Retention [$F(1,23)=43.34$, $MSe=0.0056$, $p<.001$] and a main effect of Relatedness [$F(2,46)=11.61$, $MSe=0.004$, $p<.001$]. The interaction did not reach significance. A similar ANOVA on the response accuracy only revealed a main effect of Relatedness [$F(2,46)=3.97$, $MSe=0.001$, $p<.05$] and a marginal Retention x Relatedness interaction [$F(2,46)=2.59$, $MSe=0.001$, $p=.086$].

Pairwise comparisons revealed that priming effects were only obtained for the strongly related prime-target pairs in the no-retention control condition for RTs [$t(23)=3.90$, $p=.001$]. Priming effects in the error rates were

found for strongly related prime-target pairs in both Retention conditions [control: $t(23)=2.10$, $p<.05$; retention: $t(23)=2.40$, $p<.05$] and for weakly related prime-target pairs in the retention condition [$t(23)=2.81$, $p=.01$].

Experiment 2 replicates the findings of Experiment 1 in showing priming effects only in the control condition and only for the strongly related prime-target pairs. A between-experiment analysis further revealed that the reaction times between the two groups did not differ (all $ps>.15$), suggesting that the type of memory task did not have a noticeable impact on performance.

Given the possibility that the amount of priming is affected by the increased attentional focus, a Controlled Center-Surround hypothesis would have to predict that the off-surround component exerts more influence the more attention is paid to the prime word. It is therefore expected that at shorter SOAs, priming will be observed in the retention condition, even when the prime can be reported after the lexical decision. Experiment 3 tests this assumption.

Experiment 3

Thirty-two volunteers from the University of London participated in the experiment. All participants had English as their first language, were right-handed and had normal or corrected-to-normal vision. The design, materials and procedure were the same as in experiment 2, with the difference that the prime was presented for 250 ms and was immediately followed by the target.

Results and discussion

The mean median reaction times and error rates for both conditions are presented in Table 1. Performance on recognising the prime was at ceiling (98% correct) and did not show an effect of Relatedness. The data from one participant were excluded from the analysis due to extreme long RTs.

An overall ANOVA on the (log-transformed) lexical decision times revealed a main effect of Retention [$F(1,30)=75.14$, $MSe=0.06$, $p<.001$] and a main effect of Relatedness [$F(2,60)=8.15$, $MSe=0.005$, $p=.001$]. The interaction did not reach significance. A similar ANOVA on the response accuracy also revealed a marginal effect of Retention [$F(1,30)=3.51$, $MSe=0.002$, $p=.071$] and a marginal effect of Relatedness [$F(2,60)=2.62$, $MSe=0.001$, $p=.081$].

Pairwise comparisons revealed that, for RTs, priming effects were obtained for the strongly related prime-target pairs in both the control [$t(30)=3.46$, $p<.005$] and the retention [$t(30)=2.20$, $p<.05$] condition. Weakly related prime-target pairs showed a priming effect only in the retention condition [$t(30)=2.15$, $p<.05$]. Priming effects in the error rates were found only for strongly related prime-target pairs in the no-retention control condition [$t(30)=2.43$, $p<.05$].

Experiment 3 confirms the assumption that the mechanism responsible for the absence of priming in Experiments 1 and 2 develops over time. Interestingly, in contrast to the findings with long SOA, with short SOA, the numerical values of the RT-priming effect are larger in the retention than in the control condition.

Table 1: Results of Experiments 1, 2 and 3. RTs in ms and proportion correct within brackets.

	Unrelated	Weak-related	Strong-related
Experiment 1 (N=20): recall + long SOA			
Control	606 (.95)	587 (.98)	576 (.98)
Retention	759 (.96)	769 (.98)	736 (.96)
Experiment 2 (N=24): recognition + long SOA			
Control	633 (.96)	631 (.99)	597 (1.0)
Retention	816 (.98)	828 (1.0)	788 (1.0)
Experiment 3 (N=31): recognition + short SOA			
Control	755 (.97)	741 (.96)	722 (.99)
Retention	1070 (.98)	1015 (.98)	1012 (.99)

General Discussion

The three experiments provided further insight into the observation that priming effects are absent when the prime word is actively maintained during lexical decision to the target. In all three experiments, reaction times in the retention condition were slower than in the standard control condition, which merely reflects the increase in cognitive demand in this dual-task situation. The effect of relatedness was only significant in Experiments 2 and 3, which employed a recognition task on the prime word. Although the interaction between Retention and Relatedness was not significant in any of the experiments, based on previous reports, pilot studies and the predictions from the Controlled Center-Surround hypothesis, pairwise comparisons revealed an interesting picture. With long SOA (1250ms), priming was only found for strongly related targets and only in the control condition, the *prime-retention effect*. With short SOA (250 ms), priming was found for weakly and strongly related targets in the retention condition and for strongly related targets in the control condition.

Although further studies are required to investigate this pattern in more depth, the present set of experiments already rules out two alternative explanations for the absence of a priming effect in the retention condition. First, in Fischler and Goodman (1978), Henik et. al. (1994) and Experiment 1, a verbal report had to be given after the lexical decision task. It is possible that the absence of semantic priming does not originate at a memory level, but instead may be due to some form of interference between executing the lexical decision and preparing the articulatory response for reporting the prime. However, the fact that the main results did not change when a recognition task was used instead of

a recall task, suggest that the *retention* of the prime and not the articulatory preparation of the prime was crucial. Second, Henik, et. al. (1994) suggested that the absence of priming in their experiment was due to the prime word being held at a shallow level of processing (e.g., phonological code) preventing "the needed attentional resources from being allocated at the semantic level" (p. 165). However, in the experiments reported here, the strength of the prime-target association modulated the effect. This also indicates that the *prime-retention effect* and the *prime-task effect* are different phenomena. The former requires full processing and active maintenance of the prime word, whereas the latter requires allocating attention away from the semantic level (but remaining within the same processing domain; Chiappe, Smith & Besner, 1996).

In awaiting more conclusive evidence, the current results support the proposal that the center-surround mechanism, which is assumed to be a structural component of the semantic memory system, dominates when more attention is directed to the prime word. In such situations prime-activation and activation-spread may trade off. To illustrate the Controlled Center-Surround hypothesis, consider Figure 1. In this figure, the strength of association between prime and target are set on the abscissa with the weakest strength to the right. On the ordinate the priming effect (RTunrelated – RTrelated) is set out for the long SOA (averaged over Experiments 1 and 2) and for the control and retention condition. The 'priming effect' for the 'prime' is a linear extrapolation of the values for the strong and weak associates. This figure makes two points. First, it makes the intuitive prediction that when the prime is in short-term memory, a decision on the prime itself is speeded up. Second, the overall pattern resembles the textbook example of an attentional on-center-off-surround 'Mexican hat' receptive field in the visual domain.

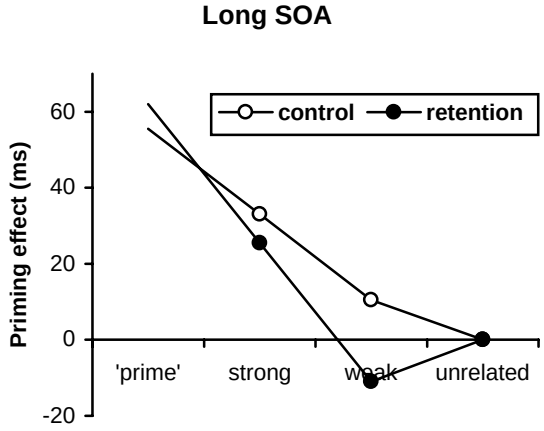


Figure 1: Priming effect as function of the strength of the prime-target association for *long* SOA. The values for the 'prime' are linear extrapolations from the values at strong and weak strength.

The figure for the short SOA is more complex (see Figure 2). The same prediction for identity priming is made, with larger 'priming' in the retention condition compared to control. However, the priming effect at short SOA seems larger in the retention condition than in the control condition. This pattern was replicated in a follow-up study (not reported here) and if this holds true in future studies, it would mirror the two-stage activation process proposed in the literature using homographs (words that have multiple meanings). In this two-stage process, an initial (automatic) activation of all meanings of a word is followed by a stage in which non-dominant or incongruent meanings are suppressed (Simpson & Burgess, 1985; Simpson & Kang, 1994). According to the hypothesis proposed here, the suppression in the second stage is a result of biased competition, where a deliberate directed attention to relevant word meanings makes them win this competition.

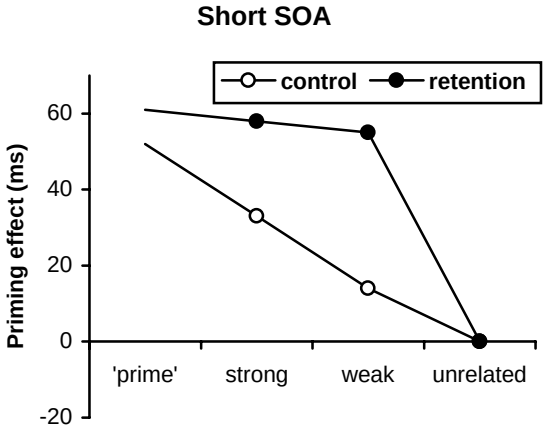


Figure 2: Priming effect as function of the strength of the prime-target association for *short* SOA. The values for the 'prime' are linear extrapolations from the values at strong and weak strength.

Although not predicted initially, the numerically larger priming effect with short SOA is not inconsistent with the Controlled Center-Surround hypothesis. The initial activation of the prime facilitates strong and weak related targets, but the inhibitory influence is only felt after the prime has received a large amount of activation (as is the case with long SOA). When the prime needs to be retained, the prime is activated very strongly, leading to larger priming effects for both weak- and strong-related targets at short SOA, but at long SOAs the off-surround component depresses both targets below the point where priming effects are obtained (or even a trend for a negative priming is observed). This attentional tuning on semantic concepts is only suggested by the presented dataset. A series of experiments are being prepared to address other methodological and theoretical issues. Nevertheless, the mere observation that the priming effect is modulated by short-term retention of the prime word poses interesting

constraints on existing and future computational models of priming.

The finding of a *prime-retention effect* motivates taking a closer look at the structure of semantic memory and the influence of controlled attention on its internal dynamics. Understanding these characteristics may provide valuable contributions to debates on the automaticity assumption of the spread of activation and resource limitations in language/cognitive processing. For example, an initial step in modelling the prime-retention effect (Davelaar, 2004), suggests ways to account for a variety of empirical findings on the interaction between attention and memory, such as hyperpriming in thought-disordered schizophrenic patients (e.g., Spitzer, et. al., 1993), individual differences in negative priming and presentation rate effects in false memory (McDermott & Watson, 2001).

Dagenbach and Carr (1994) proposed the Center-Surround hypothesis to account for the strategic carry-over effects in masked priming experiments. Here, the *prime-retention effect* suggests that (1) the center-surround mechanism can be observed in the behavioural data when attention is allocated to parts of the semantic system and (2) has a specific time-course. Future research, using the prime-retention paradigm, could provide detailed information on the structure of semantic memory and the temporal dynamics of the processes that control the spread of activation.

Acknowledgments

During the writing process of this paper, the author was supported by a fellowship (T026271312) and subsequently by a research grant (RES-000-22-0655) both from the Economic and Social Research Council. The author would like to thank Jennifer Aydelott and Marius Usher for comments on a previous version.

References

Chiappe, P. R., Smith, M. C., & Besner, D. (1996). Semantic priming in visual word recognition: activation blocking and domains of processing. *Psychonomic Bulletin and Review, 3*, 249-253.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33*, 497-505.

Dagenbach, D., & Carr, T. H. (1994). Inhibitory processes in perceptual recognition: Evidence for a center-surround attentional mechanism. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory Processes in Attention, Memory, and Language*. San Diego, CA: Academic Press.

Davelaar, E. J. (2004). Towards a model of the prime-retention effect. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Fischler, I., & Goodman, G. O. (1978). Latency of associative activation in memory. *Journal of Experimental Psychology: Human Perception and Performance, 4*(3), 455-470.

Grossberg, S. (1978). A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in Theoretical Biology*, Vol. 5. New York: Academic Press.

Henik, A., Friedrich, F. J., Tzelgov, J., & Tramer, S. (1994). Capacity demands of automatic processing in semantic priming. *Memory and Cognition, 22*, 157-168.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision and visual masking: a survey and appraisal. *Behavioral and Brain Sciences, 9*, 1-66.

Houghton, G., & Tipper, S. P. (1994). A model of inhibitory mechanisms in selective attention. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory Processes in Attention, Memory, and Language.* San Diego, CA: Academic Press.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Maxfield, L. (1997). Attention and semantic priming: a review of prime task effects. *Consciousness and Cognition, 6*, 204-218.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: the impact of presentation duration. *Journal of Memory and Language, 45*, 160-176.

Moss, H., & Older, L. (1996). *Birkbeck word association norms*. Lawrence Erlbaum Associates: Hove, UK.

Mummery, C. J., Shallice, T., & Price, C. J. (1999). Dual-process model in semantic priming: a functional imaging perspective. *NeuroImage, 9*, 516-525.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: evidence for facilitatory and inhibitory processes. *Memory and Cognition, 4*, 648-654.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226-254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: visual word recognition*. Hillsdale, NJ: Erlbaum.

Schneider, W. (1995). *MEL Professional version 2.0*. Psychology Software Tools, Inc: Pittsburgh.

Simpson, G. B., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 28-39.

Simpson, G. B., & Kang, H. (1994). Inhibitory processes in the recognition of homograph meanings. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory Processes in Attention, Memory, and Language.* San Diego, CA: Academic Press.

Smith, M. C., Besner, D., & Miyoshi, H. (1994). New limits to automaticity: context modulates semantic priming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 104-115.

Spitzer, M., Braun, U., Hermle, L., & Maier, S. (1993). Associative semantic network dysfunction in thought-disordered schizophrenic patients: direct evidence form indirect semantic priming. *Biological Psychiatry, 34*, 864-877.

Temporal Distance, Event Representation, and Similarity

Samuel B. Day (s-day2@northwestern.edu)
Department of Psychology, Northwestern University
2029 N. Sheridan Drive Evanston, IL 60208-2710 USA

Daniel M. Bartels (d-bartels@northwestern.edu)
Department of Psychology, Northwestern University
2029 N. Sheridan Drive Evanston, IL 60208-2710 USA

Abstract

A recent line of research suggests that an event's temporal distance from the present has an effect on the way in which it is likely to be construed. Specifically, more distant events are proposed to be represented primarily in terms of abstract, decontextualized information, while events in the near future tend to produce relatively more concrete, situation-specific construals. In the current study, we examine the extent to which this sort of effect can differentially emphasize *commonalities* between two events. Similarity ratings were collected for pairs of events sharing either high-level or low-level commonalities, and described as occurring in either the near or distant future. Consistent with predictions, an interaction was observed between temporal distance and commonality level. Broader implications for cognitive processing are discussed.

Introduction

One of the remarkable strengths of the human cognitive system is its flexibility. Not only are we able to store vast amounts of information about our world, organized into categories, scripts and schemas, we also seem particularly proficient at tailoring that information to fit the current demands of our environment. In particular, we seem capable of representing the same entity or event in a wide variety of ways. In addition to taxonomic organizations that can differentially emphasize various aspects of the same individual (e.g., *animal* / *mammal* / *dog* / *collie* / *Rover*), we appear to fluently cross-classify things based on goals, scripts, and evaluations (Barsalou, 1983, 1985; Ross & Murphy, 1999). Further, activation and retrieval of specific information can be highly subject to effects of general context (e.g., Tulving, 1972; Godden & Baddeley, 1975). Similarly, contextual or top-down effects may have a substantial impact on how new experiences are perceived and encoded (Bower, Black, & Turner, 1979; Bransford & Johnson, 1972). An important implication of this variation in representation concerns the impact that it may have on the processes that operate this activated information.

An intriguing recent body of research demonstrates an additional factor which may broadly affect event representations: temporal distance. Given the importance of our perceptions of the future for our ability to plan, to set and pursue goals, and to generally make long-term judgments, predictions, and choices, these effects have received surprisingly little attention in cognitive science on the whole. In the current paper, we examine the impact of such temporally-based construals on the ubiquitous process of similarity judgment. A demonstration that these construal effects may influence similarity—which is widely implicated in such fundamental cognitive processes as retrieval, categorization and inference—could serve to emphasize the expansive role that this kind of context-based effect plays throughout cognition.

Temporal Construal Theory

A recent set of studies in the judgment and decision making literature suggests that the way a person construes an event can be influenced by the temporal context in which that event takes place. According to Temporal Construal Theory (TCT) (Liberman & Trope, 1998; Sagristano, Trope, & Liberman, 2002; Trope & Liberman, 2000), events in the near future are likely to be represented largely in terms of concrete, circumstantial, and goal-irrelevant features, while the representations of events in the distant future tend to emphasize features that are more abstract, central, and goal-relevant. Trope and Liberman refer to these sparser, less contextualized representations characteristic of the distant future as *high-level* construals, and to the more enriched and highly contextualized representations more common in near future events as *low-level* construals.

These differences in representation can have behavioral consequences. For instance, one might agree to give a talk at a conference on some date in the distant future, perhaps focusing almost exclusively on the event's abstract, positive aspects: the opportunity to receive feedback on one's work, the opportunity for public exposure, and so on. As the date of the conference approaches, however, one's focus may begin to shift to some of the contextualized details that were absent in the initial representation. For instance, the time demands one faces in preparing for the presentation might become more salient, making the whole experience seem more effortful. The net effect, in this case, would be that the presentation loses some of the positive valence it once had.

These proposed differences in event representations can lead to changes in preference, depending on whether an event is described as being in the near or distant future. In one set of studies (Liberman & Trope, 1998), participants were given descriptions of events which, like the conference example, had opposite evaluative valences for high- versus low-level construals. For instance, one set of participants was asked to judge their likelihood of attending a lecture that was described as relevant and interesting but scheduled at an inconvenient time of day, while another set considered the case of a less interesting lecture, but one which was scheduled for a more convenient time. In the first group, the high-level construal was assumed to be positive

(relevant, pleasant experience) and the low-level construal negative (the logistics of fitting it into one's schedule). The second group was expected to have the opposite pattern of appraisal: a more negative impression of the abstract details of the situation, but a more positive sense of the low-level procedures involved in participating. While participants were more likely to attend the interesting lecture overall, this difference was more pronounced when the lecture was described as taking place a year from now rather than tomorrow. In other words, the decision pertaining to the near future seemed to give significantly more weight to the concrete, contextual aspects of the situation. The overall preference for the interesting lecture is also relevant, since it is consistent with the theory's suggestion that more proximal events are represented by some combination of contextual and abstract information, while distant events are primarily abstractly represented.

Note, however, that in these studies, the focus is on shifts in preference; representation is just a tool for the demonstrations. As such, this and all of their evidence concerning how we represent events is second-order, in that we have to infer differences in event representation from people's choices, or from people's preferred descriptions of events (see Liberman & Trope, 1998). One purpose of the present study was to seek more direct evidence for differences in event representation as a function of temporal context.

Similarity and Representational Level

Similarity is widely held to be one of the most critical concepts in cognition, and there are few aspects of mental life that do not seem to depend on it in one way or another. Similarity is seen as playing a vital role in recall through reminding (Hintzman, 1984; Ross, 1984); most theories of categorization rely heavily on the concept of similarity to determine category membership of a new item (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984); similarity is proposed to be fundamental in making generalizations, inferences and knowledge transfer (e.g., Novick, 1988; Osherson, et al, 1990; Ross, 1984).

It has been demonstrated that judgments of similarity are not simply a function of low-level featural overlap, but also depend on structural relationships between representations (Gentner & Markman, 1997; Markman & Gentner, 1993). This is true not only in terms of the impact of commonalities in the relations and relational systems themselves, but also in the way that attributes in corresponding roles that are defined by those structures are emphasized.

These "deeper" commonalities of relational systems are extraordinarily useful for generating new knowledge and for applying existing knowledge to new situations. For example, recognition of a common causal structure in two ostensibly different systems may lead to a deeper understanding of one system via analogical inferences from the other (*see* Gentner, 1983; Hummel & Holyoak, 1997).

Thus, the similarity rating task seems particularly apt in the current context, first because it acts as a connection to deeper aspects of cognition in general, and also because it may provide added insight into situations in which high-level, relational commonalities may be highlighted in comparison.

Experiment

The primary goal of this study was to determine whether temporal distance could affect event representations in such a way as to alter the perceived similarity between events. There are two significant motivations for this approach. First, it would provide a fairly direct way of assessing representation that would not need to rely on complex secondary tasks. Second, and perhaps more important, it would provide a bridge linking temporal effects such as these to a much broader set of cognitive issues. The critical role that similarity is proposed to play in so many mental activities suggests that observed systematic changes in similarity should have a relevant and far-reaching impact on cognition generally.

Just as any entity or event may be represented in a wide variety of ways, so may any pair of things share a great number of commonalities. Some have gone so far as to suggest that this robs similarity of any explanatory power (Goodman, 1972), since all pairs of items are potentially infinitely similar to one another (*e.g*., two things may both have mass, may both be smaller than the sun, etc.). A more measured and practical approach has been to emphasize the relative salience of the various pieces of information in each entity's representation. This salience could vary as a function of things such as prior knowledge, recent exposure or priming, and even the nature of the comparison context itself (Medin, Goldstone, & Gentner, 1993; Tversky, 1977; Gentner & Markman, 1997; Gentner, Rattermann & Forbus, 1993). Information that is more salient in a representation is assumed to be given more weight in judgments of similarity. Importantly, as previously noted, the properties that contribute to similarity judgments are not limited to concrete perceptual features, but also include the more abstract, relational concepts that structure and bind those features together (Gentner & Markman, 1997).

The predictions for the current study are straightforward. If the representations of two events are composed primarily of high-level, abstract descriptions of those events, then commonalities (or lack of commonalities) at that level of analysis should play a major role in their perceived similarity. That is, we would expect similarity ratings to be driven significantly by abstract, structuring information such as goals, causes and relationships. If, on the other hand, the representations also contain information involving low-level concrete and perceptual aspects of the situations, an impact of the commonalities at that situation-specific level should also be observed.

Consider for a moment your representation of visiting a dentist's office. This representation could include fairly high-level information pertaining to conscientiousness and long-term health benefits, as well as more concrete situational information about the particular setting and sensations involved. Now consider two different events to which this situation could be compared: the act of joining a health club, or the act of getting a tattoo. The health club event seems to share a number of abstract characteristics with the dentist visit (the goal of health benefits, etc.), but

appears quite different in terms of the situation-specific details. The tattoo, on the other hand, shares a surprising number of low-level, concrete features (reclining chair, needles, physical pain), but little in the way of high-level commonalities. The important outcome of this is that the abstraction level at which the dentist event is construed should have significant (and opposite) effects on the outcomes of these two comparisons.

This is exactly the situation that we created in our study. Participants were asked to give similarity ratings for pairs of events that shared primarily either high-level or low-level commonalities. Further, these events could be described as taking place in either the near future or the distant future. If, in fact, the temporal relationship of the events to the present time has an impact on their level of construal, then we should expect to see an interaction between temporal distance and level of commonality, such that pairs with high-level commonalities should be perceived as *more* similar in the distant than the near future, while pairs sharing low-level features should become *less* similar in the distant future relative to the near future.

Participants

Twenty-three Northwestern University undergraduates participated in this study for partial course credit.

Materials and Procedure

The materials for this experiment consisted of sentence pairs describing two events that a fictitious character was planning to undertake in the future. Each test item included a *standard* sentence, and one of two *comparison* sentences. These comparison sentences were constructed to share either high-level or low-level commonalities with the standard, but not both. In addition to these test items, the material set included several filler sentence pairs, which were either *literally similar*, sharing both high- and low-level features, or *non-similar*, sharing neither.

Additionally, these events were described as taking place either in the near future ("this week") or the distance future ("next year"). This distinction acted as a between-subjects factor, with all events for a particular participant being described at the same temporal distance. Commonality level served as a within-subjects factor, with half of the standards randomly being paired with high-level comparison sentences and the other half with low. Thus, the experiment was a 2 (temporal distance: near vs. distant future) × 2 (commonality level: high vs. low pairing) mixed design.

In total, 10 test items (five at each commonality level) and 13 filler items were presented in a completely randomized order (different for each participant), with the exception that all participants were given the same two initial items (one literally similar, and one non-similar) to help "anchor" their rating range and reduce variability. Within each item, sentence order was randomized, with the standard appearing first in approximately half of the pairs. A typical test item might read as follows: "Tomorrow, Karen will go to the dentist. Tomorrow she also will join a health club." Sample materials are given in Table 1.

The experiment was implemented as a computer-based task. After instructions, the first sentence pair appeared on the screen, followed by the prompt "How similar do you think these activities are to each other?" Beneath this prompt was a horizontal bar, with endpoints labeled "very dissimilar" and "very similar". Participants were instructed to click a location on this bar to indicate their perception of the similarity of the two events. This response was normalized to a value between 0 and 1, for the "dissimilar" and "similar" endpoints, respectively. To ensure that participants were attending to the task, response latencies of less than 3 seconds for any item resulted in the warning "Too Fast" appearing on the screen, followed by a delay of several seconds before proceeding to the subsequent item.

Table 1. Sample events. Low-level comparison sentences were designed to share concrete features and procedures with the standard, while High-level comparisons share more abstract commonalities.

Event Standard	Low-level comparison	High-level comparison
Reading and coding completed research questionnaires	Doing taxes	Conducting telephone surveys
Going door-to-door distributing leaflets about the environment	Going trick-or-treating with daughter	Writing letters to congressmen and local council members
Going to the dentist	Getting a tattoo	Joining a health club
Buying diamond necklace for wife	Buying expensive watch for self	Taking wife out for gourmet meal
Calling colleges requesting information packets	Calling hotels to arrange Summer trip to Mexico	Taking the SAT

Results

Consistent with predictions, a 2 × 2 ANOVA revealed an interaction between temporal distance and commonality level, F (1, 21) = 8.60, $p < .01$. Participants rated high-level pairs as more similar in the distant future condition (M = 0.63, SD = 0.11) than in the near future condition (M = 0.58, SD = 0.10). Conversely, low-level pairs were rated as more similar in the near future condition (M = 0.54, SD = 0.10) than in the distant future condition (M = 0.40, SD = 0.09).

A main effect of commonality level was also observed. Participants rated high-level pairs as more similar (M = 0.60, SD = 0.11) than low-level pairs (M = 0.47, SD = 0.12) overall, F (1, 21) = 18.84, $p < .001$. Additionally, there was a marginal trend for participants to rate near future events as more similar than distant future events (p = .092). This latter effect is more pronounced in the across-item analyses: a 2 × 2 ANOVA run across items revealed the same interaction between temporal distance and commonality level, F (1, 18) = 7.77, p = .012, the same main effect of commonality level, F (1, 18) = 9.96, p = .005, and a significant effect of temporal distance, F (1, 18) = 4.73, $p <$ .05. This final effect reflects the fact that the low-level comparisons showed a more dramatic decrease with temporal distance than the corresponding increase in the high-level comparisons. In fact, post-hoc t-tests indicated that only the low-level comparisons changed significantly across distances (t (1,22) = 3.67, $p < .01$) (see discussion below).

Discussion

The primary predictions of the experiment were confirmed. Participants judged event pairs with abstract, high-level commonalities to be more similar in the distant future than the near future, while pairs sharing more concrete and low-level procedural features showed the opposite pattern. This supports the proposal that temporal distance is in fact influencing the level at which events are construed, and that these representational differences are stable enough to be reflected in perceived similarity.

The two observed main effects, while not predicted, are in retrospect completely consistent with the assumptions of temporal construal theory. While distant events are assumed to be represented primarily in terms of their abstract characteristics, proximal events are suggested to have more "enriched" representations that combine some contextual and some abstract information. Consistent with this characterization, the greatest observed effect was the drop in the similarity of low-level pairs between close and distant conditions, contributing to both of these additional effects. As noted above, although both low- and high-level comparisons changed in the predicted direction with temporal distance, this change was only statistically significant for the low-level pairs.

The second main effect—the overall preference for high-level commonalities—has been replicated in more recent pilot data, and is consistent with prior research showing a preference for relational over attributional similarity (Gentner & Clement, 1988; Goldstone, Medin, & Gentner, 1991).

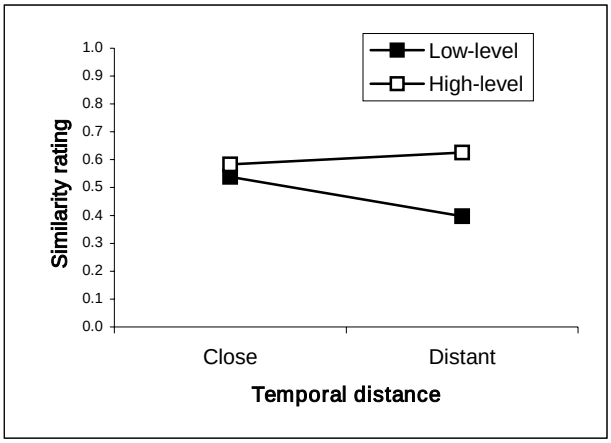


Figure 1. Interaction between commonality level and temporal distance.

The most immediate implication of these findings is that this sort of temporally-based context effect should now be predicted to influence many of the other cognitive processes in which similarity participates. For instance, it is possible that categorization of events and entities considered as being in the distant future may be based on somewhat more abstract dimensions than those considered in the immediate present. This categorization would in turn affect the inferences that an individual is likely to make in the absence of explicit knowledge, and their confidence in the accuracy of those inferences. One interesting prediction is that these perceived similarities may influence the extent to which knowledge is successfully transferred from one domain to another. Moreover, this knowledge transfer—which is seen as relying on the mapping of structural commonalities—might benefit more generally from the abstract representations characteristic of temporal distance.

Preliminary pilot data collected by the authors suggest that temporal context may have an effect on cued retrieval. That is, the perceived increase in similarity associated with an "appropriate" encoding situation (temporally close for low-level commonalities, temporally distant for high-level commonalities) may improve the probability of retrieving an event from memory when cued with the previously compared event.

The results suggest a number of avenues for future research. One important direction would involve varying the kind of high-level commonalities involved in the comparisons to include abstract characteristics other than those emphasizing individual plans and goals. While this proved to be a useful way of describing future events for our experimental purposes, it could potentially lead to confounds such as attribution of particular personality traits to the characters (e.g., planning to do x is consistent with planning to do y), and the use of a broader set of abstract event commonalities would help to address this.

Another interesting approach would be the examination of temporal distance in the opposite direction, seeing whether similar results could be obtained with events that occurred

in the recent or distant past. Taking this a step further, it may be the case that the important dimension in these effects is *psychological distance* rather than simply temporal distance. If this were the case, we might expect to find similar effects by varying dimensions such as similarity of the character to the participant, or the probability of a future event occurring.

Conclusions

We spend a great deal of time thinking about the future. In fact, this capacity seems to be a defining and distinctive characteristic of human cognition. We consider possible outcomes, evaluate potential alternatives, and pursue distant goals that may take years or even decades to achieve. Because our mental focus is so often situated in the future, it seems particularly relevant to consider the influence of temporal distance on cognition. The current study, though modest, highlights just how far-reaching these effects may be.

Acknowledgements

This work was supported by ONR award N00014-02-1-0078. Thanks to Jason Jameson, Jennifer Asmuth, Wisconsin Badger hockey, and the Similarity and Analogy group.

References

Barsalou, L. W. (1983). Ad hoc categories. Memory & Cognition, 11, 211–227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*, 629–649.

Bransford, J. D, & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717–726.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177–220.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*, 45–56.

Gentner, D., Rattermann, M. J., and Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology 25*(4), 524–575.

Godden, D. R., & Baddeley, A. D. (1975). Context dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*(3), 325–331.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgments. *Cognitive Psychology, 23*, 222–264.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects* (pp. 437–447). New York: Bobbs-Merrill.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*(3), 427–466.

Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology, 75*, 5–18.

Markman, A. B., & Gentner, D. (1993b). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32*, 517–535.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254–278.

Medin, D.L., and Schaffer, M.M. (1978). Context Theory of Classification Learning, *Psychological Review, 85*, 207–238.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104–114.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 510–520.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Ross, B.H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology, 16*(3), 371–416.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38*, 495–553.

Sagristano, M. D., Trope, Y., & Liberman, N. (2002). Time-dependent gambling: Odds now, money later. *Journal of Experimental Psychology: General, 131*, 364–376.

Trope, Y., & Liberman, N. (2000). Temporal construal and time-dependent changes in preference. *Journal of Personality and Social Psychology, 79*, 876–889.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory*. New York: Academic Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

The Time Course of Verb Processing in Dutch Sentences

Dieuwke de Goede (d.de.goede@let.rug.nl)
Femke Wester (f.wester@let.rug.nl)
Dirk-Bart den Ouden(d.b.den.ouden@let.rug.nl)
Roelien Bastiaanse (y.r.m.bastiaanse@let.rug.nl)
University of Groningen
Graduate School of Behavioral and Cognitive Neurosciences
Department of Linguistics, PO Box 716
9700 AS Groningen, The Netherlands

Lewis P. Shapiro (shapiro@mail.sdsu.edu)
San Diego State University
Department of Communicative Disorders, 5500 Campanile Drive
San Diego, CA 92182-1518

David A. Swinney (dswinney@psy.ucsd.edu)
University of California, San Diego
Department of Psychology, 9500 Gilman Drive
La Jolla, CA 92093-0109

Abstract

In Dutch matrix clauses the verb is not in its base position, but has been moved from the end of the clause to second position. Three Cross-Modal Priming experiments showed that the on-line activation pattern for moved verbs in Dutch differs significantly from the pattern for moved nouns in English. Whereas in *wh*-movement reactivation of moved nouns is found at their base position, the current results suggest that moved verbs are maintained active during the entire clause. The results are discussed in light of a gap-filling account, and three proposals are given to explain the long-lasting activation of the verb.

Introduction

Cross-Modal Priming (CMP) studies (e.g. Love & Swinney, 1996; Swinney, Ford, Bresnan, & Frauenfelder, 1988) have shown that in complex sentences where the object is not in its base position the meaning of this moved constituent is reactivated at its original position, directly after the verb. Swinney et al. (1988) tested reactivation of moved *wh*-phrases in sentences such as:

(1) The cop saw the boy_i who_i the crowd at the party accused t_i of the crime.[1]

Since *boy* is the direct object of *accused* in base structure[2], and English is an SVO language, the base position of *boy* is to the right of the verb *accused*. Therefore, it is assumed that a trace (or gap) is postulated after *accused*. Using the CMP task[3] priming effects were found for probes related to the antecedent *boy* when presented at the gap position (directly after the verb). Importantly, no priming was found at a control position before the verb. In other words, it appears that listeners *re*activate the meaning of the moved constituent when they encounter the gap. This finding has been replicated many times (for an overview see Featherston, 2001; Love & Swinney, 1996).

Several explanations have been given for this phenomenon. Swinney and colleagues (1988) provide a structural account: the meaning of the moved object is recovered in its base position to regain the canonical sentence structure. This would suggest a close correspondence between linguistic theory (or at least the generative approach) and psychological reality (the functioning of the human parser). Others (e.g. Pickering & Barry, 1991) came up with a verb-centered, semantic, account: the meaning of the moved object is reactivated because after processing the verb a dependency relation is established between the verb and its dependents, that is, its arguments (see also Nicol, 1993).

The current paper is an attempt to broaden this research area by extending the topic of research to the activation

[1] Following the traditions in processing research, the *t* refers to the gap position (*trace)* and co-indexation indicates the relationship between filler and gap.

[2] Strictly speaking, 'who' is the direct object of 'accused' and not 'boy', but it is assumed that 'who', because of its coreference with 'boy', has inherited the semantic characteristics of 'boy'.

[3] The CMP task is a dual task in which participants listen to sentences and make a lexical decision to a visual probe presented at a particular point during each sentence. Faster reaction times to a probe that is associatively related to a particular word in the sentence as compared to reaction times to a probe that is unrelated (but otherwise comparable to the related probe) is attributed to priming effects. If priming is found at a certain point during the sentence, this is taken as evidence that the meaning of the relevant word in the sentence is activated.

pattern of verbs. It is unknown whether *verbs* that are not in their base position are reactivated at this position. If moved verbs show reactivation at their base position, then a strong case can be made for the possibility that listeners attempt to recover base word order whenever they encounter a structure that is non-canonical. However, some characteristics of verbs suggest that moved verbs might behave differently from moved noun constituents.

Linguistic Background

Whereas in *wh*- and NP-movement whole sentence constituents (XPs) are moved, in verb movement it is only the verb itself (the head, X^0) that moves. Linguists (e.g. Chomsky, 1995) stress that movement of syntactic heads is radically different from movement of syntactic phrases, having for example no effect on interpretation.

In English, the language in which most studies on gap-filling have been performed, verb movement only occurs in negative inversion structures and in questions, where the auxiliary moves. In Dutch, however, verb movement is an omnipresent phenomenon. Dutch is generally agreed to be an SOV language (Koster, 1975, but see Zwart, 1997), but in a matrix clause the finite verb moves from its basic, clause-final, position to the second position in the clause (Verb Second or V2).

Psycholinguistic Background

Verbs play a central and binding role in the sentence: they not only determine the event of the sentence, but they are also linked to all other main constituents of the sentence (the arguments) and assign thematic roles to these constituents. Psycholinguistic studies suggest that these differences between nouns and verbs may matter in sentence processing: influences of argument structure have been found at different levels of sentence processing (for interference effects see Shapiro, Zurif, & Grimshaw, 1987; for syntactic priming effects, see Trueswell & Kim, 1998).

It is unknown what role the characteristics of verbs play in on-line processing of moved verbs, although two unpublished studies suggest that the patterns for verb movement might differ from those found in *wh*-movement. Muckel, Urban and Heartl (p.c.) examined split particle verbs in German SVO sentences in which the matrix verb had been moved to V2 and the particle remained in the base position of the verb (German is an SOV language). Responses to identical probes were faster than responses to control probes at the control probe point (which was placed in front of the word preceding the particle, so two words before the gap) as well as at the experimental probe point. No interaction between probe point and probe type was found, so no evidence was found for *re*activation.

Basilico, Piñar and Antón-Méndez (1995) ran a CMP-study that focused on moved verbs in Spanish. They used declarative sentences with two different word orders (VSO and VOS) and concluded that in the sentences where the verb was moved (VSO) the verb was activated at the gap. Conclusions about *re*activation cannot be drawn, however, because no pre-gap control probe-point was included in the design.

Experiment 1 & 2

In two consecutive CMP experiments, the question was addressed whether or not verb movement has processing consequences similar to *wh*-movement. The experimental sentences in these experiments are Dutch matrix clauses in which the verb has been moved from its base clause-final position to V2 position, leaving behind a gap. If verb movement and *wh*-movement are processed similarly, we expect to find activation of the verb directly after the verb (direct priming), deactivation of the verb in between the overt verb position and the gap, and finally reactivation at the gap (gap-filling).

Method

Participants 44 Participants were tested in experiment 1 and 60 in experiment 2.

Materials Sentences consisting of a matrix clause (SVO) followed by an embedded clause were auditorily presented. In both experiments the matrix clause ended after the direct object. In experiment 1 the direct object occurred directly after the verb (see example sentence 2), in experiment 2 an adjunct preceded the direct object (see example sentence 3).

(2) De kleine jongens imiteren$_i$ [1] hun fanatieke [2] rood-aangelopen voetbaltrainer t_i, omdat [3] ze later allemaal profvoetballer willen worden.
The little boys imitate$_i$ [1] their fanatical [2] red-faced soccer coach t_i, because [3] they all want to be professional soccer players when they grow up

(3) De trouwe volgelingen wijzigen$_i$ [1] eens in de zoveel tijd hun altijd [2] controversiële mening t_i [3], want [4] hun leider is een wispelturig man.
The faithful followers change$_i$ [1] once in a while their always [2] controversial opinion t_i [3], because [4] their leader is a rather fickle man.

The probes that were presented during the experimental sentences were verbs that were either associatively related to the finite verb or unrelated but matched to the related probe for baseline lexical decision time, frequency, length and argument structure. Both probe types were pre-tested off-line for any possible inadvertent source of priming. The same prime - unrelated probe - related probe triads were used in both experiments.[4]
Probes were presented at four different positions (see example sentences 2 and 3):

1. *verb* probe point: indicated as [1], placed directly after the verb

[4] Two triads were excluded in experiment 2 for counterbalancing reasons (we used 4 probe points instead of 3, so the number of experimental sentences had to be dividable by 4).

2. *control* probe point: indicated as [2], presented at 700 ms after [1] in experiment 1 and at 1500 ms after [1] in experiment 2
3. *end-of-clause* probe point: measured in experiment 2 only and indicated as [3], presented at the end of the clause (offset object head noun)
4. *conjunction* probe point: indicated as [3] in experiment 1 and as [4] in experiment 2, presented at the offset of the conjunction.

In experiment 1 42 experimental sentences were used, in experiment 2 there were 40 experimental sentences. In each experiment, an equal number of pseudo-experimental sentences (sentences with the same structure as the experimental sentences) were added and combined with non-words, to prevent any correlation between sentence type and response type. In addition, 20 filler sentences of different structures (10 words, 10 non-words) and 15 yes/no comprehension questions were added (to encourage participants to pay attention to the spoken sentences).

Probe point and probe type were both within-participants factors. All sentences were ordered pseudo-randomly. A completely counterbalanced design was created to assure that all participants saw both related and control probes, and saw probes at all three probe points. Each participant was tested twice, on the same list, but with related and control probes shifted.

Procedures The participants were tested individually in a sound-proof room with no visual distractors. The sentences were presented over headphones with an interval of 1500 ms. The probes were presented on a standard computer screen. The experimental software Tempo (designed at the University of California, San Diego, for running CMP-studies), combined with a response box with two buttons, was used to present the items and register the accuracy and RTs of the responses. Each probe was presented for 300 ms and a response could be given within a 2000 ms interval from stimulus onset. Importantly, the sentences continued without interruption during visual presentation of the probe.

Participants were instructed to listen carefully to the sentences and to expect comprehension questions about some sentences, but only about the sentence immediately prior to the question. Questions were answered and lexical decisions were made by pressing the left button on the button box for *no* and *non-word* and the right button for *yes* and *word*. Participants were instructed to answer as quickly and accurately as possible.

Results

Participants were excluded from further analysis 1) if their error score on the lexical decision task was greater than 10%, 2) if their mean or SD RT deviated from the overall mean or SD by more than 2.5 SD, or 3) if less than 67% of the comprehension questions were answered correctly. Data from three participants were excluded in both experiments.

Error rates were low (1.4% and 1.8%, respectively) and equally distributed across related and control probes and across probe points. The exclusion of errors and outliers (all values deviating from the participants and item mean for the particular data point with more than 2.5 SD were excluded) resulted in 2.7 and 3.1 percent data loss, respectively.

The mean RTs for all probe points and probe types are presented in Table 1 (the values that are presented here and in the following tables are derived from the subject-analyses; the item-analysis revealed very similar data).

The subject-based ANOVAs revealed a significant main effect of probe type in both experiments; overall, the related probes generated shorter RTs than the control probes (exp 1: F1 (1,40) = 7.91, p = .008; exp 2: F1 (1,56) = 4.61, p = .036). The item-based ANOVA was marginally significant in experiment 1 (F2 (1,41) = 3.43, p = .07), but did not reach significance in experiment 2 (F2 (1,39) = .98, p > .3).

Paired t-tests showed significant[5] faster responses to related than to control probes (priming) at the *verb* probe point in experiment 1 (t1 (40) = 2.53, p = .008; t2 (41) = 1.75, p = .044), but not in experiment 2 (t1(56) = .15, p > .4; t2 (39) = .29, p > .3). At the *control* probe point priming was found in both experiments (this effect was significant in the subject analysis (exp 1: t1 (40) = 2.64, p = .006; exp 2: t1 (56) = 2.49, p = .008), but in the item-analysis only a trend was found (exp 1: t2 (41) = 1.40, p = .085; exp 2: t2 (39) = 1.37, p = .09). Furthermore, priming was found at the *end-of-clause* probe point in experiment 2 (t1 (56) = 2.08, p = .021; t2 (39) = 1.76, p = .043). Neither of the experiments, however, showed a priming effect at the *conjunction* probe point (exp 1: t1 (40) = .81 p > .2; t2 (41) = .82, p > .2; exp 2: t1 (56) = -.98, p > .15; t2 (39) = -.95, p > .15).

Table 1: Mean RTs (and SDs) to probe type as a function of probe position in experiment 1 and 2.

probe type	verb	control	end-of-clause	conjunction
Experiment 1				
control	633 (68)	635 (61)	-	626 (72)
related	617 (65)	621 (66)	-	620 (73)
difference	*16 (41)*	*14 (33)*	-	*6 (47)*
Experiment 2				
control	663 (94)	671 (99)	668 (95)	666 (88)
related	662 (91)	657 (84)	654 (101)	672 (103)
difference	*1 (49)*	*15 (45)*	*14 (51)*	*-6 (42)*

Conclusion and Discussion

These experiments did not provide evidence for *re*activation of the verb at its base position. Both experiments converge on a pattern of activation of the verb at the *control* probe points (700 and 1500 ms after the actual occurrence of the verb in the sentence) and deactivation of the verb immediately following the conjunction linking the matrix to the second clause. The second experiment further shows that the verb is active at the *end of the clause*. The results for the

[5] As no inhibition effects were expected all t-tests are 1-tailed.

verb probe point, where priming of the verb was expected directly after its occurrence, are less clear. Although significant facilitation of the related probe compared to the control probe was found in experiment 1, experiment 2 surprisingly showed a null-effect at this probe position. [6]

An important question that remains to be answered after these experiments is: Why do verbs remain active for such an extended period of time? One possible reason is that verbs stay active to 'find' their arguments in order to theta-mark them (Argument Structure Hypothesis).[7] According to this hypothesis continued activation is predicted up to the final argument. As all verbs that were used in the current experiments were two-place verbs, the final argument was always the direct object, which occurred at the end of the matrix clause in both experiments.

Experiment 3

To test the Argument Structure Hypothesis, an experiment was run that employed an adjunct immediately after the second argument. This allowed investigation of whether saturation of the argument structure of the verb is the basis for discontinued activation of the verb, or whether the verb always remains active up till the end of the clause.

Method

Materials The same primes, related probe and control probes were used in experiment 3, but the sentences were slightly altered. The experimental sentences still consisted of a matrix clause followed by an embedded clause, but after the Object Noun Phrase an adjunct was inserted (4). The adjuncts that were used were Adverbial Phrases of Time.

(4) De domme gedetineerden beroven [1] vijftien rijke bejaarden tijdens hun [2] eerste proefverlof [3], dus vrijlating zit er voorlopig niet in.
The stupid detainees rob [1] fifteen rich seniors during their [2] first parole [3], so release seems to be out of the question for now.

The *verb* probe point (see [1]) was presented slightly later than in experiment 1 and 2: at the onset of the first word following the verb, and if this point could not be measured adequately, at the onset of the first vowel of this word. The *control* probe point [2] was now at 700 ms after the onset of the adjunct. The final probe point was at the *end of the clause*, directly at the offset of the final word of the adjunct [3].

[6] Phonological assimilation made it impossible to place all probe points exactly at the onset of the next word. Post hoc analyses showed that probes that were placed too early had a low probability to show faster RTs to related probes than to control probes, whereas the majority of probes that were presented exactly at the onset of the word following the verb showed facilitation for the related probe (χ^2 (1) = 17.4, p < .001). Part of the third experiment focuses on this observation.

[7] More accounts will be discussed in the General Discussion.

Participants and Procedures 48 Participants were tested following the same procedure as in experiment 1 and 2.

Results

The data were handled in the same way as in experiment 1 and 2. Three participants were excluded from further analysis and exclusion of errors and outliers resulted in 3.0 percent data loss.

The RTs for this experiment are presented in table 2 and show faster responses to related probes than control probes at all probe points (F1 (1,44) = 24.35, p < .001; F2 (1,41) = 6.38, p = .016). So, first of all, directly after the *verb*, a significant priming effect is obtained again, which indicates that a small adaptation in probe placement (consistently at the onset of the first word following the verb or slightly later) resulted in stable priming effects at this probe point (t1 (44) = 3.08, p = .002; t2 (41) = 1.75, p = .044). But more interestingly, activation of the verb was still evident after all arguments were processed, 700 ms into the adjunct (t1 (44) = 2.35, p = .012; t2 (41) = 2.39, p = .011), as well as at the end of the clause (t1 (44) = 2.59, p = .007; t2 (41) = 1.77, p = .042).

Table 2: Mean RTs (and SDs) to probe type as a function of probe position in experiment 3.

probe type	verb	control	end-of-clause	conjunction
control	721 (86)	723 (95)	712 (89)	-
related	697 (89)	706 (96)	693 (88)	-
difference	*24 (53)*	*17 (48)*	*19 (48)*	-

Conclusion

The current experiment shows that verbs in Dutch declarative matrix clauses are maintained active throughout their entire clause, even after all arguments have been encountered.

General Discussion

The aim of the first two experiments was to evaluate whether moved verbs (in Dutch) behave similarly to moved nouns, which show reactivation of the moved constituent at the location of the gap. The results show that, at least in Dutch, processing of moved verbs is different from that of moved noun constituents in English. No evidence was provided for *re*activation of the verb at its base (clause-final) position. Instead, the experiments demonstrated that the verb remains active from the point where it is first encountered up to the offset of the object head noun, clause-finally. So even though activation of the verb at the site of the gap was found in the second experiment, the verb was also active at the control probe points, unlike in the gap-filling studies in English (Love & Swinney, 1996).

The question whether verb movement is reflected in psychological reality cannot be answered on the basis of these data. Although we did not find evidence for reactivation, a syntactically based *Gap-Filling Account* can

still hold for the data; the verb was indeed active at the gap. It is possible that, instead of being reactivated, verbs are maintained active until the gap (and because of the gap).

The studies by Basilico et al. (1995) and Muckel et al. (p.c.) show that the present results do not stand alone. The results from both studies can be interpreted in different ways, but their findings might well be in line with ours. Muckel et al. found activation of a split particle verb at a control probe point and at the particle. Since they did not test for priming directly after the verb and do not present baseline reaction times for the two probe types the results are suggestive of either continued activation of the verb or no activation at all[8]. Basilico et al. found activation of the verb at the site of the assumed gap (VS*O) in Spanish sentences with non-basic word order. However, they did not test directly after the verb, neither did they use a control probe point to check whether the activation of the verb had faded in between. It is possible, therefore, that a moved verb in Spanish also remains active.

As far as the third experiment is concerned, current linguistics theories disagree about the position of adjuncts and therefore it remains unclear whether the base position of the verb should be postulated at the end of the clause (after the adjunct) or after the direct object (in front of the adjunct). If the verb gap is posited in front of the adjunct in sentences like the ones used in experiment 3, the results of this particular experiment show activation of the verb even after the gap and thus provide evidence against the Gap-Filling Account as an explanation for our data.

It is not inconceivable that movement of verbs is not reflected in psychological reality. In *wh*-extraction, all linguistic theories accept some kind of relation between the *wh*-constituent and the verb later in the sentence. In movement of noun constituents, gap-filling is necessary to assign thematic roles to constituents that are detached from their subcategorizer. In contrast, the assumption of verb movement is very theory-internal to generative linguistics (Chomsky, 1995). Also, where verb movement is concerned, gap-filling is not necessary for sentence interpretation but only for structural, syntactic reasons: to fulfill requirements formulated by formal syntactic theory.

The Gap-Filling Account can definitely be rejected if non-moved verbs are also found to remain active in English. A study in English will be performed by Lewis Shapiro and David Swinney. In this study sentences similar to the ones used in the present experiments will be used. In English, the finite verb is in base position, and there is no gap at the end of the clause. If the verb does not remain active in English, the present results are most likely to be due to restoration of the base word order (gap-filling). If, however, the results for the study on English are similar to the results of the experiments discussed in this paper, other accounts will gain in credibility.

[8] In this case, the priming effects can be explained by the materials: identical probes are generally more 'sensitive to priming' than associatively related primes.

Three possible alternative accounts will be discussed in the final section of this paper.[9] The first account is the *Argument Structure Account* which was tested in experiment 3. According to this explanation, verbs remain active to be able to assign theta roles to their arguments. This hypothesis was falsified in experiment 3 where activation of the verb was found during an adjunct phrase placed after the final argument.

An alternative syntactic explanation for the present data is the *VP-shell Theory* (Larson, 1988). According to the variant of Van Zonneveld & Bastiaanse (2000), both arguments and adjuncts are in SpecVP positions. This is only possible when the VP is recursive: it takes a VP as a complement. The VP-shell theory thus predicts a sentence structure with as many VPs as there are specifiers (arguments and adjuncts). The VP-shell in itself is a complement of IP. An example of sentences as used in experiment 2 can illustrate this (figure 1).

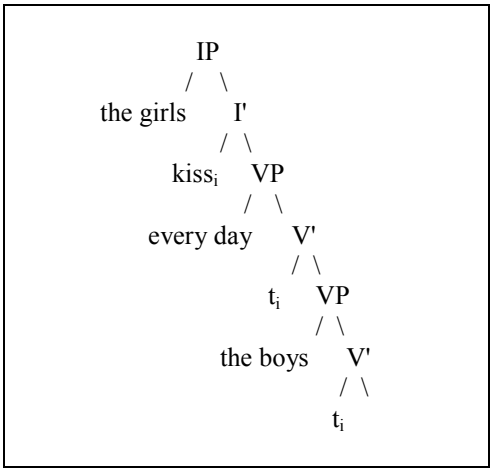


Figure 1: Syntactic representation for Dutch, according to VP-shell theory (simplified version of an experimental sentence from experiment 2).

As this figure shows, there is an empty V^0 position in the first and second VP. According to Van Zonneveld & Bastiaanse (2000) the verb is not 'moved' to the head of I', but it is 'lexicalized' or 'activated' again in each head position of V'. This means that the verb is active during the entire clause and, unlike nouns, does not need to be *re*activated at the gap position.

[9] These accounts are all based on the assumption that only verbs show long-lasting activation. Although no continued activation for *wh*-constituents was found in CMP experiments within the standard *wh*-movement paradigms, Nicol (1993) could not exclude a pattern of maintained activation of PPs in sentences with ditransitive verbs, where the PP was fronted, and of subjects in S - relative clause - V - O sentences, where priming of the subject was found after the verb. Nevertheless, other explanations could apply to these data as well (or even better), suggesting that detailed studies on the activation pattern of nouns need to be done as well. We are currently preparing a study on the activation pattern of nouns.

The verb thus remains active within its own clause boundary. Only when a new CP or IP is encountered, with a new VP-shell structure as its complement, the activation of the verb is no longer necessary and will therefore disappear. This is exactly what was found in the present experiments. Interestingly, unlike the Gap-filling Account, VP-shell theory predicts similar verb-activation patterns for both English and Dutch matrix clauses, because in both cases the verb c-commands the entire VP-shell.

A third theory (*Semantic Account)* is deduced from studies that show that verbs are more polysemous than nouns (Fellbaum, 1993; Gentner & France, 1988) and are more adjustable and mutable. In a paper-and-pencil task, Gentner & France (1988) found that if the meaning of a verb does not fit with the noun that it co-occurs with, participants are more eager to change the meaning of the verb than the meaning of the noun. Interestingly, these mutability effects are seen in on-line processing as well: an eye-tracking study by Pickering and Frisson (Pickering & Frisson, 2001) showed that lexical ambiguity resolution for verbs is delayed compared to nouns. The suggestion of these authors is that the interpretation of a verb is highly dependent on the arguments with which it combines in a particular sentence. Interestingly, this is also the case for non-ambiguous verbs that have multiple senses. To understand the full meaning of, for example, the verb *open*, one needs to know whether it concerns *opening a door*, or *a file*. According to the "underspecification model" (Frisson & Pickering, 1999) "the processor activates a single underspecified meaning for a verb with multiple senses and uses evidence from context to home in on the appropriate sense" (Pickering & Frisson, 2001, p. 564). Therefore, also in the case of unambiguous verbs, delaying the interpretation process until the arguments are processed seems to make sense.

Although Pickering and Frisson focus on the role of arguments, it is possible that adjuncts play a role in the interpretation of the verb, too. Actually, one should notice that verb-interpretation and sentence-interpretation are intermingled and can be seen as ongoing processes, which possibly only stop at clause boundaries. This also suggests that the VP-shell Account and the Semantic Account might be difficult to tease apart and should perhaps be interpreted as accounts that explain the same phenomena, but do so at a different linguistic level.

Acknowledgments

The first author is funded by the Dutch Organization for Scientific Research (NWO) under grant #360-70-091. Edwin Maas has been of invaluable help in setting up the first experiment. We are grateful to John Hoeks, Roel Jonkers and three anonymous reviewers for their helpful comments on earlier versions of this paper.

References

Basilico, D., Pinar, P., & Anton-Mendez, I. (1995). *Canonical word order and the processing of verbal traces in Spanish*. Poster presented at the Eighth Annual CUNY Conference on Human Sentence Processing, Tucson, AZ.

Chomsky, N. (1995). *The minimalist program*. Cambrigde, MA: MIT Press.

Featherston, S. (2001). *Empty categories in sentence processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Fellbaum, C. (1993). English verbs as a semantic net. Obtained from: www.cogsci.princeton.edu/~wn, *Five papers on Wordnet* (pp. 40-61).

Frisson, S. & Pickering, M. J. (1999). The Processing of Metonymy: Evidence From Eye Movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1366-1383.

Gentner, D. & France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.). *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence* (pp. 343-382). San Mateo, CA: Kaufmann.

Koster, J. (1975). Dutch as an SOV Language. *Linguistic Analysis, 1,* 111-136.

Larson, R. K. (1988). On the double object construction. *Linguistic Inquiry, 19,* 335-391.

Love, T. & Swinney, D. (1996). Coreference processing and levels of analysis in object-relative constructions; demonstration of antecedent reactivation with the cross-modal priming paradigm. *Journal of Psycholinguistic Research, 25,* 5-24.

Pickering, M. J. & Frisson, S. (2001). Processing ambiguous verbs: evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 556-573.

Pickering, M. & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes, 6,* 229-259.

Shapiro, L. P., Zurif, E., & Grimshaw, J. (1987). Sentence processing and the mental representation of verbs. *Cognition: International Journal of Cognitive Science, 27,* 219-246.

Swinney, D. A., Ford, M., Bresnan, J., & Frauenfelder, U. L. (1988). On the temporal course of gap-filling and antecedent assignment during sentence processing. In B. Grosz, R. Kaplan, M. Macken, & I. Sag (Eds.), *Language Structure and Processing* (Stanford, CA: CSLI.

Trueswell, J. C. & Kim, A. E. (1998). How to prune a Garden Path by nipping it in the bud: fast priming of verb argument structure. *Journal of Memory and Language, 39,* 102-123.

Zonneveld, R.van & Bastiaanse, R. (2000). Frasale recursie: De syntaxis van de gelaagde VP. *TABU, 30,* 143-173.

Zwart, J. W. (1997). *Morphosyntax of Verb Movement: a minimalist approach to the syntax of Dutch*. Dordrecht: Kluwer Academic Publishers.

Smarter and Richer?: Executive Processing and the Monty Hall Dilemma

Wim De Neys (Wim.Deneys@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Abstract

The Monty Hall Dilemma (MHD) is a striking example of the human tendency to base probability judgment on intuitive, erroneous heuristics instead of an analytic, normative reasoning process. Two experiments tested the claim (e.g., Stanovich & West, 2000) that correct, normative reasoning draws on executive, working memory resources (WM) whereas heuristic reasoning would be purely automatic. Experiment 1A examined the link between MHD-reasoning and WM-capacity. Participants that solved the MHD correctly had a significantly higher WM-capacity. Experiment 1B presents a new approach to test the role of the WM-resources experimentally. Participants solved the MHD while WM-resources were burdened by a secondary task. Correct responses decreased under load. The results provide new evidence for the differential role of executive resources in heuristic and analytic reasoning.

Introduction

A main theme of cognitive reasoning research over the last decades is that human judgment frequently violates traditional normative standards. In a wide range of reasoning tasks most people do not give the answer that is correct according to logic or probability theory. The discrepancy between normative models and peoples actual performance has been labeled the "normative/descriptive gap" (Stanovich, 1999). The present study focuses on one of the most striking examples of this discrepancy: The Monty Hall Dilemma.

The notorious, counterintuitive Monty Hall Dilemma was adapted from a popular TV game show (Friedman, 1998). Host Monty Hall asks his final guest to choose one of three doors. One of the doors conceals a valuable prize and the other two contain worthless prizes such as goats or a bunch of toilet paper. After the guest makes a selection, the host, who knows where the prize is, opens one of the non-chosen doors to show that it contains a dud. The guests are then asked if they want to stay with their first choice or switch to the other unopened door.

Most people have the strong intuition that whether they switch or not the probability of winning remains 50% either way. However, from a normative point of view, the best strategy is to switch to the other door. Indeed, switching yields a 2/3 chance of winning. The solution hinges on the crucial fact that the host will never open the door concealing the prize, and obviously, he will not open the guest's door either. Taking into account that two thirds of the times the prize will be in one of the non-chosen doors, the non-chosen door that is still closed will hide the prize in two thirds of the trials (Tubau & Alonso, 2003).

Empirical studies of the Monty Hall Dilemma consistently showed that the vast majority of college students fails to give the correct response (switching rates ranging from 9% to 21%, e.g., Burns & Wieth, 2000, 2003; Friedman, 1998; Granberg & Brown, 1995; Krauss & Wang, 2003; Tubau & Alonso, 2003). Likewise, vos Savant (1997) reports that after she discussed the problem in a weekly magazine column she received up to 10,000 letters in response. Ninety-two percent of the writers from the general public disagreed with the switching answer. To paraphrase Friedman (1998), it seems that because of peoples poor MHD reasoning "millions of dollars were left on Monty's table".

Research indicates that the typical MHD response can be attributed to the operation of erroneous but very powerful intuitions or heuristics. For example, Shimojo and Ichikawa (1989) found that most people base their answer on the so called number-of-cases heuristic ("if the number of alternatives is N, then the probability of each one is 1/N"). Thus, since only two doors remain people will automatically assign a 50% chance to each door and fail to take the "knowledgeable host" information into account. Similar claims can be found in Falk (1992) and Johnson-Laird, Legrenzi, Girotto, Legrenzi, and Caverni (1999).

It has been argued that human thinking in general typically relies on the operation of intuitive, prepotent heuristics instead of a deliberate, controlled reasoning process. The primacy of these heuristics has been called the fundamental computational bias in human cognition (Stanovich, 1999). Whereas the fast and undemanding heuristics provide us with useful responses in many situations they can bias reasoning in tasks that require more elaborate, analytic processing (e.g., Evans & Over, 1996; Kahneman, Slovic, & Tversky, 1982;. Sloman, 1996; Stanovich, 1999; Stanovich & West, 2000; Tversky & Kahneman, 1983).

Stanovich and West (e.g., 2000) stressed that although the modal response is often erroneous in many reasoning tasks, a small proportion of

the participants does give responses that are in line with the normative standards. Their research on individual differences showed that participants that gave the normative response on classic reasoning tasks such as the conjunction fallacy (Tversky & Kahneman, 1983) and the Wason (1966) selection task were disproportionally those highest in cognitive (working memory) capacity. According to Stanovich and West's dual process framework (see also Evans & Over, 1996; Sloman, 1996) correct normative responding requires that an analytic, controlled reasoning process overrides the prepotent heuristics. The inhibition of the heuristic system and the computations of the analytic system would draw on limited, executive working memory resources. The more resources that are available, the more likely that the analytic system will be successfully engaged and the correct response calculated.

The Stanovich and West (2000) findings suggest that a possible antidote to erroneous MHD reasoning might be a high working memory span. If correct normative reasoning requires executive working memory (WM) resources, then participants with a higher WM-span should be more likely to select the switching response. Bluntly put, the guests that win the prize in the game show will not only be richer but also "smarter"[1]. The link between MHD-reasoning and WM-capacity was examined in Experiment 1A.

The Stanovich and West framework and related dual process theories have been severely criticized (e.g., Stanovich & West, 2000). One important issue concerns the central assumption about the role of controlled, executive resources. Both the claim that correct, normative reasoning depends on the executive system and the characterization of the heuristic system as automatic and independent from executive control have been questioned (e.g., Handley, Feeney, Harper, 2002; Klaczynski, 2000, 2001; Osman, 2002)

Experiment 1B presents a new approach in the dual process field. The experiment adopted secondary task methodology to burden the executive WM-resources while participants were solving the MHD. If correct responding in the MHD draws on WM-resources, performance should decrease under load since less resources will be available for inhibition of the prepotent "50%-heuristic" and subsequent analytic computations. On the other hand, if the heuristic processing would not be automatic and would draw on WM, it will also become harder for people to come up with the "equal probability" answer. The procedure thereby allows a direct, experimental test of the basic executive processing assumptions.

[1] The term « smarter » refers of course to the tight connection between executive WM-capacity and general cognitive ability (e.g., Engle, Tuholski, Laughlin, & Conway, 1999).

Experiment 1A

Method

Participants

A total of 236 first-year psychology students from the Unversity of Leuven, Belgium, participated in return for psychology course credit.

Material

Working memory measure. Participants' working memory capacity was measured using a version of the Operation Span task (Ospan, La Pointe & Engle, 1990) adapted for group testing (Gospan, for details see De Neys, d'Ydewalle, Schaeken, & Vos, 2002). In the Ospan-task participants solve series of simple mathematical operations while attempting to remember a list of unrelated words The main adaptation in the Gospan is that the operation from an operation-word pair is first presented separately on screen (e.g., 'IS (4/2) – 1 = 5 ?'). Participants read the operation silently and press a key to indicate whether the answer is correct or not. Responses and response latencies are recorded. After the participant has typed down the response, the corresponding word (e.g., 'BALL') from the operation-word string is presented for 800 ms. As in the standard Ospan three sets of each length (from two to six operation-word pairs) are tested and set size varies in the same randomly chosen order for each participant. The Gospan-score is the sum of the recalled words for all sets recalled completely and in correct order.

Participants who made more than 15% math errors or whose mean operation response latencies deviated by more than 2.5 standard deviations of the sample mean were discarded (participants already in the bottom quartile of the Gospan-score distribution were not discarded based on the latency criterion). De Neys et al. (2002) reported an internal reliability coefficient alpha of .74 for the Gospan. The corrected correlation between standard Ospan and Gospan-score reached .70.

Monty Hall Dilemma. Participants were presented a standard version of the MHD taken from Krauss and Wang (2003). The formulation tried to avoid possible ambiguities (e.g., the random placement of the prize and duds behind the doors and the knowledge of the host were explicitly mentioned). The text stated (translated from Dutch):

> *Suppose you're on a game show and you're given the choice of three doors. Behind one door*

is the main prize (a car) and behind the other two doors there are dud prizes (a bunch of toilet paper). The car and the dud prizes are placed randomly behind the doors before the show. The rules of the game are as follows: After you have chosen a door, the door remains closed for the time being. The game show host, Monty Hall, who knows what is behind the doors, then opens one of the two remaining doors which always reveals a dud. After he has opened one of the doors with a dud, Monty Hall asks the participant whether he/she wants to stay with his/her first choice or to switch to the last remaining door. Suppose that you chose door 1 and the host opens door 3, which has a dud.

The host now asks you whether you want to switch to door 2. What should you do to have most chance of winning the main prize?

a. Stick with your first choice, door 1.
b. Switch to door 2.
c. It does not matter. Chances are even.

The MHD was presented on computer. Participants were instructed to carefully read the basic problem information (text in italics), first. When they were finished reading they pressed the ENTER-key and then the question and answer-alternatives (underlined text) appeared on the screen (text in italics remained on the screen). Participants typed their response (a, b, or c) on the keyboard. Instructions stated there were no time limits.

Procedure

The experiment was run on computer. Participants were tested in groups of 21 to 48. Participants completed the Gospan and MHD in a one-hour session, in which they also completed some other tasks not part of the present investigation. The MHD was presented after the Gospan task.

Results and discussion

Six participants were discarded because they did not meet the operation correctness or latency requirements of the WM-measure. Mean Gospan-score of the remaining 230 participants was 32.26 (SD = 10.45).

Consistent with previous MHD-studies only a small minority of the participants (5.2%) gave the correct switching answer. The vast majority (85.7%) believed that switching and sticking were equally good strategies. However, the crucial finding is that the participants that did give the correct response had a significantly larger WM-capacity. Mean Gospan-score of the participants that gave the correct response was 38.08 vs. only 31.94 for the incorrect responders, $t(228) = 2$, $n1 = 12$, $n2 = 218$, $p < .05$. In terms of effect sizes, Cohen's d reached .59. Such an effect is classified as "moderate" (Rosenthal & Rosnow, 1991) and corresponds to the effect sizes reported by Stanovich and West (1998a, 1998b) for the impact of executive capacity on the reasoning tasks in their studies.

The present association between MHD-performance and WM-capacity supports Stanovich and West's basic claim concerning the involvement of executive resources in normative reasoning. However, the evidence remains purely correlational. More direct evidence for the mediating role of the executive resources is needed (Klaczynski, 2000). Experiment 1B introduces secondary task methodology to test the basic processing claims experimentally.

Experiment 1B

A major problem for Stanovich and West (2000) and related dual processing frameworks is that the basic processing assumption, the different involvement of controlled, executive resources in heuristic and analytic reasoning, is disputed. On one hand, available (correlational) evidence for the role of executive resources in analytic, normative reasoning has been questioned (e.g., Klaczynski, 2000). On the other hand, the proposed characterization of the heuristic system as automatic and independent from executive control has been challenged (e.g., Handley, Feeney, Harper, 2002; Klaczynski, 2001; Osman, 2002). Experiment 1B presents a new approach to test the basic processing claims.

Participants solved the MHD while they performed a secondary task that burdened the executive WM-resources. If correct responding in the MHD draws on WM-resources, performance should decrease under load since less resources will be available for inhibition of the dominant "50%-heuristic" and subsequent analytic computations. On the other hand, if heuristic processing would not be automatic and would draw on WM, it will also become harder for people to come up with the "equal probability" answer and we would expect a decrease in "50%" responses. In case both the heuristic and normative response would draw on executive resources the computation of any single response should be hindered and we might expect a random guessing pattern under secondary task load.

The secondary task was adopted from Kane and Engle (2000). Participants were requested to continuously tap a novel, complex finger pattern (e.g., index finger/ring finger/ middle finger/ pinkie) with their non-dominant hand while reasoning. The task was selected because previous studies (e.g., Kane & Engle, 2000; Moscovitch, 1994) consistently showed that it put a premium on efficient executive WM-functioning.

Method

Participants
Forty-one first-year psychology students from the Unversity of Leuven, Belgium, participated in return for psychology course credit. None of the participants had participated in Experiment 1A. All participants had taken the Gospan-test prior to Experiment 1B. Between 13 to 46 days intervened between participation in the Gospan-test and Experiment 1B.

Materials
Monty Hall Dilemma. Participants were presented the same version of the MHD as in Experiment 1A.

WM-load task. A program executed by a second computer collected the finger-tapping data. All participants tapped on the "V", "B", "N", and "M" keys on the QUERTY-keyboard of the second computer.

Procedure
All participants were tested individually. Participants were instructed to tap the index-ring-middle-pinkie pattern with their non-dominant hand. The experiment started with five 30s practice tapping trials. Participants always received on-line accuracy feedback: Whenever a wrong finger (key) was tapped the computer emitted a 300 ms, low pitch tone. During the first three practice trials the program also calculated the mean tapping speed for each participant. If any one intertap interval in the subsequent trials was more than 150 ms slower than the established mean, the computer emitted a 600 ms, high pitch tone. The online monitoring served to assure that the tapping task was properly performed.

After the tapping practice, the experimenter explained that the practice tapping speed had to be maintained in the upcoming reasoning task. Participants then read the basic MHD problem information (underlined text) on the screen. When they were finished reading they pressed the ENTER-key and started tapping. Then the question and answer-alternatives (text in bold) were presented and participants continuously tapped the finger pattern (with online response time and accuracy feedback) until they gave their response. Participants said out loud the letter (a, b, or c) corresponding to their answer. Instructions stated there were no time limits.

Results and discussion

Performance of the participants in Experiment 1A was used as a baseline to evaluate the impact of the WM-load. A control analysis established that the WM-capacity of the participants in Experiment 1A (Mean Gospan-score = 32.26, SD = 10.45) and 1B (Mean Gospan-score = 33.1, SD = 10.88) did not differ, $F(1, 269) < 1$.

Burdening the executive resources with the tapping task affected participants' performance. As Table 1 shows the response pattern was clearly not random. The switching rate under the secondary task load decreased to 0%. This decrease in the proportion of correct responses reached marginal significance, p1 = 5.22%, p2 = 0%, n1 = 230, n2 = 41, $t(269) = 1.50$, $p < .07$, one-tailed. The finding supports the claim that correct normative reasoning in the MHD draws on executive WM-resources. Burdening the executive resources did not decrease the rate of "equal probability" answers. Indeed, there was a slight tendency in the opposite direction. This suggests that the central MHD intuition to assign a 50% chance to the two remaining doors is an automatic, heuristic response that does not involve executive processing.

Table 1: Percentage of Different Responses in Experiment 1A and 1B

	Experiment	
Answer	1A: No load	1B: Load
Stick	9.1 (21)	7.3 (3)
Switch	5.2 (12)	0.0 (0)
Equal	85.7 (197)	92.7 (38)

Note. Raw frequencies in parentheses.

General Discussion

The present study focused on the Monty Hall Dilemma because it is one of the most striking examples of the "normative/descriptive" gap in the literature (Friedman, 1998). As in previous MHD studies only a small proportion of participants gave the correct, normative switching response. However, Experiment 1A established that the participants that did give the correct response had a significantly larger working memory capacity. This finding complements the work of Stanovich and West (2000) on individual differences in executive resources with related reasoning tasks.

According to the Stanovich and West (2000) framework and associated dual process theories (e.g., Evans & Over, 1996; Sloman, 1996) correct normative responding requires that an analytic, controlled reasoning process overrides prepotent heuristics. The inhibition of the heuristic system and the computations of the analytic system would draw on limited executive, working memory resources. The more resources that are available, the more likely that the analytic system will be successfully engaged and the correct response calculated. Experiment 1B provided experimental evidence for this view. Burdening the executive resources with a secondary task while participants were solving the MHD tended to decrease the rate of correct,

switching responses: Although the participants in Experiment 1A and 1B had comparable span sizes, non of the participants in Experiment 1B managed to solve the MHD correctly under WM-load. Indeed, more people tended to commit the intuitive tendency to assign a 50% chance to the two remaining doors. These findings support the basic claim of dual process theories concerning the differential involvement of executive resources in analytic and heuristic reasoning.

As in most MHD-studies, the proportion of correct responses under "standard" conditions in the present study was very low. A consequence of this low figure is that the study inevitably suffers from a floor-effect. The decrease in correct performance under executive load in Experiment 1B still reached marginal significance but the decrease could never be large. One possible solution for the floor-effect is adopting some of the manipulations known to increase MHD performance. Previous studies indicated that practice with the task, training procedures and simple clarifications of the causal structure of the task (e.g., Burns & Wieth, 2000, 2003; Krauss & Wang, 2003; Tubau & Alonso, 2003) increase performance. Thus, testing the impact of the WM-load with such modified MHD versions might suffer less from a floor-effect. In addition, it might be especially enlightening to examine how different span groups benefit from the increased performance manipulations.

The present findings indicate that executive resources are *necessary* for correct, normative reasoning. However, by no means this implies that a large resource pool is also *sufficient* for correct reasoning. The relation between WM-capacity and reasoning performance is not absolute. Although participants that solved the MHD correctly in Experiment 1A had a larger WM-span, numerous "high spans" nevertheless answered erroneously. To illustrate this point MHD-performance of participants in the top and bottom quartile of the WM-capacity distribution in Experiment 1A was compared. Consistent with the previous findings high spans gave significantly more switching responses[2]. But even among the 25% most cognitively gifted college students only 10% gave the correct response. Clearly, factors outside the cognitive WM-ability spectrum will also affect performance (e.g., "epistemic thinking dispositions", see Stanovich, 1999). Thus, in pointing out the necessary role of executive, WM-resources for correct reasoning the present study does not minimize the role of other mediating factors.

The present study demonstrated the potential of a dual task approach to test the central processing claims of dual process theories. In principle, future studies could adopt this procedure with all the classic tasks studied in the field (e.g., conjunction fallacy, base-rate neglect, selection task). Of course, the final empirical evaluation of the executive processing claims of dual process theories will depend on the generalization of the present findings.

[2] p1 = 10%, p2 = 1%, n1 =70, n2= 70, $t(138) = 2.21$, $p < .03$.

Acknowledgements

Preparation of this manuscript was supported by a grant from the Fund for Scientific Research-Flanders (FWO).

References

Burns, B. D., & Wieth, M. (2000). *The Monty Hall dilemma: A causal explanation of a cognitive illusion*. Paper presented at the Forty-First Annual Meeting of the Psychonomic Society, New Orleans, LA.

Burns, B. D., & Wieth, M. (2003). *Causality and Reasoning: The Monty Hall dilemma*. Paper presented at the Twenty-Fifth Annual Meeting of the Cognitive Science Society, Boston, MA.

De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica, 42,* 177-190.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*, 309-331.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

Falk, R. (1992). A closer look at the probabilities of the notorious three prisoners. *Cognition, 77*, 197-213.

Friedman, D. (1998). Monty Hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review, 88*, 933-946.

Granberg, D., & Brown, T. A. (1995). The Monty Hall dilemma. *Personality and Social Psychology Bulletin, 21*, 182-190.

Handley, S. J., Feeney, A., & Harper, C. (2002). Alternative antecedents, probabilities, and the suppression of fallacies in Wason's selection task. *Quarterly Journal of Experimental Psychology, 55A*, 799-818.

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., Legrenzi, M. S., & Caverni, J-P. (1999). Naïve probability: A mental model theory of extensional reasoning. *Psychological Review, 106*, 62-88.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and*

biases. Cambridge, MA: Cambridge University Press.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 336-358.

Klaczynski, P. A. (2000). Is rationality really "bounded" by information processing constraints? *Behavioral and Brain Sciences, 23*, 683-684.

Klaczynski, P. A. (2001). Framing effects on adolescent task representation, analytic and heuristic processing, and decision making: Implications for the normative/descriptive gap. *Applied Developmental Psychology, 22*, 289-309.

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall Problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General, 132*, 3-22.

La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1118-1133.

Moscovitch, M. (1994). Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. *Neuropsychology, 8,* 524-534.

Osman, M. (2002). *Is there evidence for unconscious reasoning processes?*. Paper presented at the Twenty-fourth Annual Meeting of the Cognitive Science Society, Fairfax, VA.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd Edition). New York: McGraw-Hill.

Shimojo, S., & Ichikawa, S. (1989). Intuitive reasoning about probability: Theoretical and experimental analysis of the "problem of three prisoners". *Cognition, 32*, 1-24.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E., & West, R. F. (1998a). Individual differences in framing and conjunction effects. *Thinking and Reasoning, 4*, 289-317.

Stanovich, K. E., & West, R. F. (1998b). Cognitive ability and variation in selection task performance. *Thinking and Reasoning, 4*, 193-230.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

Tubau, E., & Alonso, D. (2003). Overcoming illusory inferences in a probabilistic counterintuitive problem : The role of explicit representations. *Memory & Cognition, 31*, 596-607.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293-315.

vos Savant, M. (1997). *The power of logical thinking*. New York: St. Martin's Press.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology: I* (pp. 106-137). Harmandsworth, England: Penguin.

Inference Suppression and Working Memory Capacity: Inhibition of the Disabler Search

Wim De Neys (Wim.Deneys@psy.kuleuven.ac.be)
Kristien Dieussaert (Kristien.Dieussaert@psy.kuleuven.ac.be)
Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)
Géry d'Ydewalle (Géry.dYdewalle@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Abstract

We examined whether individual differences in WM-capacity affected the acceptance of an extended MP problem that explicitly mentioned a possible disabler. The explicit disabler presentation was assumed to stimulate the spontaneous disabler search process. Two experiments showed that the acceptance ratings of the extended MP problems followed a U-shaped, quadratic trend with low and high spans showing the highest MP acceptance. Contrasting performance with extended and standard MP problems indicated that all span groups showed the standard suppression effect. Findings support the claim that high spans manage to inhibit the spontaneous disabler search and underline the generality and robustness of this inhibition phenomenon.

Introduction

Suppose that on a hot, summer day you hear someone claiming "If Jenny turns on the air conditioner, then she will feel cool". Next, you hear that Jenny did turn the air conditioner on. It is likely that you will conclude that Jenny will feel cool. This inference ('If P then Q. P. Therefore, Q') is known as the Modus Ponens (MP). The MP inference is considered valid in standard logic. Now, suppose that you would also have been reminded of the fact that the air conditioner might be broken or that Jenny might have a fever. In this case you would probably have been rather reluctant to accept the standard MP inference that turning on the air conditioner will make Jenny feeling cool. Thereby, the additional information would have tempted you to commit a fallacy.

Cognitive scientists have spent a great deal of research to establish how people reason with 'if, then' sentences or conditionals. One of the main findings is that additional, 'background' knowledge about the conditional relation affects the inferences people are willing to draw (Evans, Newstead, & Byrne, 1993; Manktelow, 1999). The crucial kind of background knowledge for the evaluation of the MP inference is referred to as 'disabling conditions'. A disabling condition (also 'disabler' or 'additional requirement') is a condition that prevents the antecedent specified in the conditional (e.g., turning on the air conditioner or the P part) from bringing about the consequent (e.g., feeling cool or the Q part). In the introductory example a broken air conditioner or Jenny having a fever will both function as disablers.

In a pioneering study Byrne (1989) showed that when a possible disabler was explicitly presented to participants (e.g., If she has an essay to write, then she will study late in the library. If the library is open, then she will study late in the library. She has an essay to write. Thus, she will study late in the library?) the MP inference was less frequently accepted compared to the standard MP condition without presented disabler. Further studies established that during conditional reasoning people spontaneously search their long-term memory for stored disablers. Cummins (1995) used causal conditionals for which a pilot group could retrieve many (e.g., If you put fertilizer on plants, then they grow well) or only few (e.g., If Tom grasps the glass with his bare hands, then his fingerprints are on it) disablers. Cummins reasoned that for conditionals with many (vs. few) disablers spontaneous retrieval of a disabler would be more likely. Although no specific disablers were explicitly presented, the results indeed showed that MP inferences based on conditionals with many disablers were rejected more frequently. Numerous studies confirmed these findings (e.g., Bonnefon & Hilton, 2002; Byrne, Espino, & Santamaria, 1999; Thompson, 1994; De Neys, Schaeken, & d'Ydewalle, 2002; Stevenson & Over, 1995; George, 1997; see Politzer & Bourmaud, 2002 for a review). Thus, it is well established that finding a disabler (either spontaneously or presented by the experimenter) will result in a decreased MP acceptance. This impact of disablers on the MP inference acceptance is known as the suppression effect.

In the present study we present the first experiments that look at individual differences in the suppression impact. More precisely, we will examine whether differences in working memory (WM) capacity affect the acceptance of MP problems when a possible disabler is explicitly presented. Such WM-mediation could be expected on the basis of recent findings pointing to the role of WM in the retrieval and inhibition of stored disablers (e.g., De Neys, Schaeken, & d'Ydewalle, 2003a, 2003b; Markovits, Doyon, & Simoneau,

2002; Simoneau & Markovits, 2003; Verschueren, De Neys, Schaeken, & d'Ydewalle, 2002).

De Neys et al. (2003a) and Verschueren et al. (2002) established that the efficiency of the disabler retrieval process is mediated by WM-capacity. In a task where people were asked to generate disablers for a set of conditionals in limited time, participants higher in WM-capacity retrieved more disablers. Putting a load on WM also reduced the efficiency of the retrieval process. In a further experiment, De Neys et al. (2003b) tested a group of low, medium, and high spans (participants in the bottom, middle and top quintile of first-year psychology students' WM-capacity distribution, respectively) in an everyday conditional reasoning task. Consistent with the more efficient disabler retrieval, medium spans were more likely to reject the MP inference than low spans. On the other hand, despite the intrinsic superior retrieval capacity high spans showed nevertheless higher MP acceptance ratings than the medium spans (see also Markovits et al., 2002).

Based on findings of Stanovich and West (2000), it was assumed that a basic decontextualization ability would allow high spans to put background knowledge aside when it conflicts with the logical standards. Remember that in standard logic MP is a valid inference. Since disabler retrieval will result in the rejection of MP, a basic validity notion will conflict with the disabler retrieval process. De Neys et al. reasoned that high spans would therefore use their WM-resources for an active inhibition of the disabler search.

Simoneau and Markovits (2003) showed that more efficient inhibitory processing (as measured by a negative priming task) was indeed linked with higher MP acceptance. In a related dual-task study De Neys et al. found additional support for the inhibition hypothesis. The basic assumption states that lower spans allocate WM-capacity to the disabler retrieval, while high spans allocate WM-capacity primordially to the retrieval inhibition. Consistent with the hypothesis, the dual-task study showed that a less efficient disabler retrieval under WM-load resulted in higher MP acceptance ratings under load (vs. no load) for low spans, while the less efficient inhibition resulted in lower MP ratings under load (vs. no load) for high spans.

In sum, there is evidence for the claim that high spans are inhibiting the disabler retrieval process during conditional reasoning. Inhibition of cognitive processes deemed inappropriate is indeed one of the key executive working memory functions (e.g., Baddeley, 1996; Levy & Anderson, 2002; Miyake & Shah, 1999).

The present study will allow a further test of the disabler inhibition hypothesis. Presenting extended MP problems where a possible disabler is explicitly mentioned will push the inhibition demands to the limit. The explicit disabler presentation will stimulate the search process. Note that one of the difficulties of retrieving disablers in a reasoning task is that there is no explicit retrieval cue (e.g., Markovits & Barrouillet, 2002; Markovits & Quinn, 2002). The standard MP premises do not tell you what kind of information you should look for. If a possible disabler is added, it will be incorporated in the elementary mental representation of the inference problem. It is assumed that this representation is held in working memory. As suggested by many authors, activation will automatically start to spread from the information stored in WM (or "the focus of attention" see Cowan, 1995) to related long-term memory elements (Anderson, 1993; Cowan, 1995; see also Markovits & Barrouillet, 2002 for an integrated account). Therefore, stored disablers will receive more activation by the explicit presentation of a possible disabler with the MP inference. Consequently, it will be more likely that additional stored disablers will be automatically retrieved.

In Experiment 1 participants were given a measure of WM-capacity and extended MP problems that mentioned a possible disabler. If the high spans still manage to inhibit the stimulated disabler search we expect to see a U-shaped, quadratic trend in the acceptance ratings in function of WM-capacity. Low spans were expected to show high MP acceptance ratings because of the inefficient disabler retrieval. Medium spans should show lower MP ratings because the search will be more efficient. If high spans still manage to inhibit the disabler retrieval, acceptance ratings should increase again for the high spans.

In Experiment 2 we compared the acceptance ratings of standard and extended MP problems for different WM-span groups.

Experiment 1

Method

Participants

A total of 105 first-year psychology students from the University of Leuven (Belgium) participated in the experiment in return for course credit. None of the students had had any training in formal logic.

Material

Working memory task. Participants' working memory capacity was measured using a version of the Operation span task (Ospan, La Pointe & Engle, 1990) adapted for group testing (Gospan, for details see De Neys, d'Ydewalle, Schaeken, & Vos, 2002). In the Ospan-task participants solve series of simple mathematical operations while attempting to remember a list of unrelated words The main adaptation in the Gospan is that the operation from an operation-word pair is first presented separately on screen (e.g., 'IS (4/2) – 1 = 5 ?'). Participants read the operation silently and press a key to indicate whether the answer is correct or not. Responses and

response latencies are recorded. After the participant has typed down the response, the corresponding word (e.g., 'BALL') from the operation-word string is presented for 800 ms. As in the standard Ospan three sets of each length (from two to six operation-word pairs) are tested and set size varies in the same randomly chosen order for each participant. The Gospan-score is the sum of the recalled words for all sets recalled completely and in correct order.

Participants were tested in groups of 21 to 48 at the same time. Participants who made more than 15% math errors or whose mean operation response latencies deviated by more than 2.5 standard deviations of the sample mean were discarded (participants already in the bottom quartile of the Gospan-score distribution were not discarded based on the latency criterion). De Neys, d'Ydewalle et al. (2002) reported an internal reliability coefficient alpha of .74 for the Gospan. The corrected correlation between standard Ospan and Gospan-score reached .70.

Reasoning task. Participants received three extended MP problems. These were Dutch translations of the three Byrne (1989) MP problems (see Dieussaert, Schaeken, Schroyens, & d'Ydewalle, 2000). The following item format was used:

Rule: If she has an essay to write, then she will stay late in the library.
If the library stays open, then she will stay late in the library.
Fact: She has an essay to write
Conclusion: She will stay late in the library.

All three MP problems were presented on a separate page of a booklet together with a 7-point rating scale ranging from 1 (*Very certain that I cannot draw this conclusion*) to 7 (*Very certain that I can draw this conclusion*) with 4 representing *can't tell*. Participants placed a mark on the number of the scale that best reflected their evaluation of the conclusion.

Procedure

Participants were tested in groups of 21 to 42 at the same time in a large computer room with an individual booth for every participant. All participants started with the Gospan task that was run on computer. After all participants of a group had finished the Gospan-task the extended MP evaluation task was presented. The three items were presented on separate pages of a booklet. The first page of the booklet included the task instructions. They showed an example item that explained the specific task format. Participants were told that the task was to decide whether or not they could accept the conclusions. Care was taken to make sure participants understood the precise nature of the rating scale. The task instructions did not mention to accept the premises as true or to endorse conclusions that follow necessarily. Instead participants were told they could evaluate the conclusions by the criteria they personally judged relevant (see Cummins, 1995).

Results and discussion

Rejection probability for all reported statistical analyses was .05. For completeness, we always report the individual estimated p-values.

Three participants were discarded because they did not meet the operation correctness or latency requirements of the WM-task (see De Neys, d'Ydewalle et al., 2002). The remaining 102 participants were split in three span groups of equal n based on the boundaries of the Gospan-score distribution. Mean Gospan-score for the three successive span groups was 23.27 (SD = 4.34, low span), 35.15 (SD = 2.79, medium span), and 45.89 (SD = 4.97, high span).

For every participant we calculated the mean acceptance rating for the three extended MP problems. The means were subjected to an ANOVA with span group as between-subject variable. There was a significant effect of span group, $F(2, 99) = 5.55$, $MSE = 1.23$, $p < .01$. The acceptance rating showed the expected pattern: Medium spans ($M = 3.84$, $SD = 4.67$) showed lower MP acceptance ratings than the low ($M = 4.56$, $SD = 1.11$) and high spans ($M = 4.67$, $SD = .99$). A trend analysis confirmed that there was a significant U-shaped, quadratic trend, $F(1, 99) = 10.85$, $MSE = 1.23$, $p < .005$ without mediation of a linear trend, $F(1, 99) < 1$.

Thus, even when the disabler search was specifically stimulated high spans showed the highest levels of MP acceptance. This is consistent with the claim that high spans are inhibiting the disabler search and underlines the generality and robustness of the inhibition phenomenon.

Experiment 2

The first experiment showed that the acceptance ratings for the extended MP problems differed for participants of different WM-capacity. In Experiment 2 we compare the acceptance ratings of standard vs. extended MP problems in function of WM-span. This allows us to establish the impact of the explicit disabler presentation per se. For the validity of our framework it is crucial that the acceptance ratings decrease when a disabler is explicitly presented.

First, for low spans it is assumed that the disabler search with standard MP problems will not be very successful. Although low spans' limited resources will restrict the impact of the extra search stimulation, the extended disabler manipulation does present low spans a disabler they will probably not retrieve in the standard condition. Therefore, low spans' inference acceptance should decrease for the

extended MP problems. Second, because of the more efficient retrieval, medium spans in the standard condition will probably retrieve the disabler presented in the extended MP condition themselves. Hence, the mere presentation of the disabler should not affect medium spans. Nevertheless, if we are right that the search process is stimulated by the disabler presentation one should expect that additional disablers will be retrieved in the extended condition and this should further decrease the MP acceptance (see De Neys, Schaeken, & d'Ydewalle, 2003c, for a study on the effect of the number of retrieved disablers on MP acceptance). High spans are expected to inhibit the search both for the standard and extended problems. However, it is explicitly assumed that the inhibition is not automatic, but draws on WM-resources. Therefore, the inhibition should be less successful when the process is more demanding. De Neys et al. (2003a) already observed that the increasing inhibition demands caused by an increasing number of available disablers resulted in a less efficient disabler inhibition. Hence, although the high spans should overall show a high MP acceptance, their acceptance level should nevertheless be affected by the stimulated disabler search.

In sum, we expected a standard suppression effect for all span groups: Acceptance ratings should be lower for the extended (vs. standard) MP problems. In addition, overall MP acceptance ratings should be affected by WM-span: Extended and standard MP acceptance in the successive span groups should follow the U-shaped trend observed in Experiment 1.

Method

Design

As standard condition or baseline we used the MP evaluations of the 282 participants in the study of De Neys et al. (2003b). In this study participants were presented a standard conditional inference task with causal conditionals and a measure of WM-capacity. We calculated the mean MP acceptance for different span groups and used this as a baseline to compare the MP acceptance of matched span groups with similar extended causal MP problems.

Participants

All 105 participants of Experiment 1 evaluated the extended MP inferences in the present experiment. The data for the standard MP condition were taken from the study of De Neys et al. (2003b) where 282 first-year psychology students evaluated standard conditional inferences.

Material

Working memory task. All participants' working memory capacity was measured with the Gospan-task (see De Neys, d'Ydewalle et al., 2002).

Reasoning tasks. All conditionals were selected from the generation studies of De Neys et al. (2002) and Verschueren et al. (2002). Eight causal conditionals were used for the standard condition and six causal conditionals for the extended condition. Half of the conditionals in each condition were previously classified as having many possible disablers, while the other half had only few possible disablers. The number of possible alternative causes (see Cummins, 1995) of the selected conditionals with few and many disablers was kept constant. The item format for the extended and standard task was similar to the format used in Experiment 1, except that for the extended items a possible disabler was mentioned. We always presented the disabler that was most frequently generated for that conditional in the generation task (e.g., see De Neys et al., 2002). As in Byrne (1989) the disablers (e.g., engine broken) were always presented as an additional requirement, embedded in a conditional (e.g., If the engine works, then the car starts). This resulted in the following format:

Rule: If the ignition key is turned, then the car starts.
If the engine works, then the car starts.
Fact: The ignition key is turned.
Conclusion: The car starts.

It should be noted that the set of conditionals in the standard and extended condition was not completely similar. Although both conditions used causal conditionals with a comparable number of possible disablers, the standard condition should therefore not be conceived as a control condition per se. Rather, the standard condition serves as a baseline against which the performance of the different WM-span groups can be compared.

Procedure

Participants were tested in groups of 21 to 48 at the same time in a large computer room with an individual booth for every participant. All participants started with the Gospan task that was run on computer. After all participants of the group had finished the Gospan-task the extended MP evaluation task or the standard conditional inference task was presented. The standard task was run on computer. Participants evaluated eight standard MP inferences mixed with other conditional inferences. The six items of the extended MP task were presented on separate pages of a booklet. This booklet was presented before the booklet with the items of Experiment 1. Task instructions for the standard and extended MP task were similar to the instructions given in Experiment 1.

Results and discussion

In order to match the span groups in the extended and standard conditions as closely as possible we decided to split both samples up in five span groups

each, based on the quintile boundaries of the Gospan-score distribution of the 282 participants in the standard condition. A 5 (span group, between-subjects) x 2 (MP task, between-subjects) ANOVA on the Gospan-scores established that there were no WM-capacity differences for participants in both task conditions [effect of MP task, $F(1, 374) < 1$; interaction MP task x Span-group, $F(4, 374) = 1.84$, $MSE = 10.91$].

Each participant evaluated eight or six MP evaluations. The mean of these ratings was calculated and subjected to a 5 (WM-span, between-subjects) x 2 (MP task, between-subjects) ANOVA.

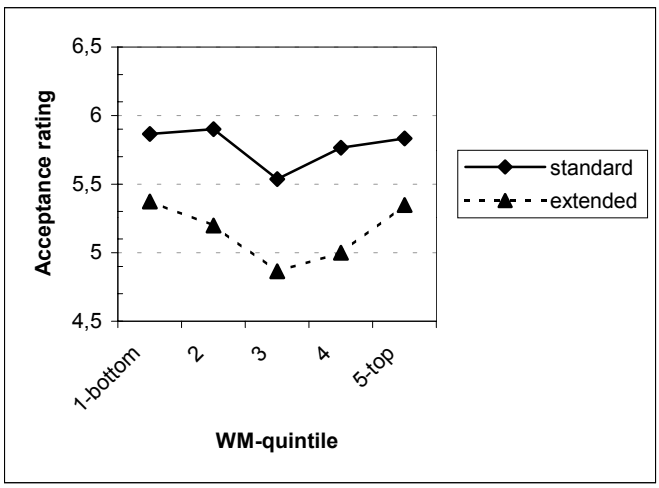


Figure 1. Mean MP acceptance rating in function of WM-capacity with (extended) and without (standard) explicitly presented disabler. The rating scale ranged from 1 (very sure cannot draw this conclusion) to 7 (very sure can draw this conclusion).

Explicitly presenting a disabler clearly decreased the MP acceptance, $F(4, 374) = 37.56$, $MSE = .72$, $p < .0001$. Figure 1 shows that, as expected, this effect was present for all WM-span groups, span group x MP task interaction, $F(4, 374) < 1$. There was also a marginal main effect of WM-span, $F(4, 374) = 2.28$, $MSE = .72$, $p < .06$. As Figure 1 indicates, a trend analysis clearly established that the MP ratings followed a U-shaped, quadratic trend in function of WM-span, $F(1, 374) = 6.77$, $MSE = .72$, $p < .01$. There was no sign of a linear trend, $F(1, 374) < 1$, and the quadratic trend did not differ for the standard and extended MP problems, $F(1, 374) = 1.07$, $MSE = .71$, $p > .35$. Thus, as expected, all span groups showed an impact of the explicit disabler presentation, but both on the standard and extended problems the MP acceptance ratings were affected by WM-capacity.

Finally, one might note that the number of disablers of the adopted conditionals in the present experiment varied systematically (e.g., half of the conditionals had few vs. many possible disablers). We had no specific hypotheses concerning the impact of this factor on the manipulations. For completeness, the variable was entered as a within-subjects factor in the ANOVA. We replicated the traditional (e.g., Cummins, 1995) main effect of the number factor: MP acceptance was always lower for conditionals with many disablers than for conditionals with few disablers, $F(1, 374) = 139.40$, $MSE = .51$, $p < .0001$. However, none of the interactions with the other factors reached significance. Thus, the crucial effects of span group and the explicit disabler presentation were not affected by the number factor.

General Discussion

The two experiments clearly established that even when a disabler is explicitly presented MP acceptance ratings of the successive working memory (WM)-span groups follow a U-shaped trend. Previous studies already suggested that the higher MP acceptance ratings of high vs. medium spans, despite high spans' superior retrieval capacities, result from an active inhibition of the disabler search. The fact that in the present study the same pattern is found under conditions that can be assumed to stimulate the search process points to the robustness and generality of the inhibition phenomenon.

As De Neys et al. (2003a, 2003b) we hypothesized that the disabler inhibition is not occurring in a cognitive vacuum but draws on working memory resources. Therefore, higher inhibition requirements were expected to result in a less efficient inhibition process. The goal of the search stimulation by the explicit disabler presentation was precisely to increase the inhibition demands. Disablers that would be inhibited under less demanding inhibition conditions could 'slip through' the filter and decrease the MP acceptance. Consistent with these hypotheses Experiment 2 clearly showed that even high spans' MP acceptance decreased for the extended MP problems.

The present findings have implications for traditional suppression studies. The results indicate that Byrne's (1989) findings can be generalized over the whole WM-capacity distribution: For all WM-span groups MP acceptance decreased when a possible disabler was explicitly presented. Thus, all WM-span groups show the basic suppression effect. However, it is important to note that the final acceptance level is systematically affected by WM-capacity: Reasoners with an inefficient disabler retrieval and reasoners that inhibit the retrieval show the highest levels of MP acceptance. These people will be typically situated in the bottom and top levels of the WM-capacity distribution, respectively. Reasoner that can allocate sufficient resources to the retrieval and do not inhibit the search process, typically people with medium sized WM-span, will be most likely to reject MP. These findings further

emphasize the role of WM-capacity in the retrieval and inhibition of disablers during conditional reasoning.

Acknowledgements

Preparation of this manuscript was supported by grants from the Fund for Scientific Research-Flanders (FWO).

References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology, 49A*, 5-28.

Bonnefon, J.-F., & Hilton D. J. (2002). The suppression of modus ponens as a case of pragmatic preconditional reasoning. *Thinking & Reasoning, 8*, 21-40.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.

Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40*, 347-373.

Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition, 23*, 646-658.

De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica, 42,* 177-190.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & Cognition, 30,* 908-920.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003a). *Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples*. (Tech. Rep. No. 295). Leuven, Belgium: University of Leuven, Laboratory of Experimental Psychology.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003b). *Working memory span and everyday conditional reasoning: A trend analysis*. Paper presented at the Twenty-Fifth Annual Meeting of the Cognitive Science Society, Boston, MA.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003c). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition, 31,* 581-595.

Dieussaert, K., Schaeken, W., Schroyens, W., & d'Ydewalle, G. (2000). Strategies during complex conditional inferences. *Thinking & Reasoning, 6*, 125-160.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Erlbaum.

George, C. (1997). Reasoning from uncertain premises. *Thinking & Reasoning, 3*, 161-189.

La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1118-1133.

Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences, 6*, 299-305.

Manktelow, K. I. (1999). *Reasoning and thinking*. Hove, UK: Psychology Press.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review, 22*, 5-36.

Markovits, H., Doyon, C., & Simoneau, M. (2002). Individual differences in working memory and conditional reasoning with concrete and abstract content. *Thinking & Reasoning, 8*, 97-107.

Markovits, H., & Quinn, S. (2002). Efficiency of retrieval correlates with logical reasoning from causal conditional premises. *Memory & Cognition, 30*, 696-706.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, MA: Cambridge University Press.

Politzer, G., & Bourmaud, G. (2002). Deductive reasoning from uncertain conditionals. *British Journal of Psychology, 93*, 345-381.

Simoneau, M., & Markovits, H. (2003). Reasoning with premises that are not empirically true : Evidence for the role of inhibition and retrieval. *Developmental Psychology, 39*, 964-975.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology, 48A*, 613-643.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22*, 742-758.

Verschueren, N., De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). *Working memory capacity and the nature of generated counterexamples*. Paper presented at the Twenty-Fourth Annual Meeting of the Cognitive Science Society, Fairfax, VA.

A Computational Model of Children's Semantic Memory

Guy Denhière (denhiere@up.univ-mrs.fr)
L.P.C & C.N.R.S. Université de Provence
Case 66, 3 place Victor Hugo
13331 Marseille Cedex, France

Benoît Lemaire (Benoit.Lemaire@upmf-grenoble.fr)
L.S.E., University of Grenoble 2, BP 47
38040 Grenoble Cedex 9, France

Abstract

A computational model of children's semantic memory is built from the Latent Semantic Analysis (LSA) of a multisource child corpus. Three tests of the model are described, simulating a vocabulary test, an association test and a recall task. For each one, results from experiments with children are presented and compared to the model data. Adequacy is correct, which means that this simulation of children's semantic memory can be used to simulate a variety of children's cognitive processes.

Introduction

Models of human language processing are usually based on a layer of basic semantic representations on top of which cognitive processes are described. For instance, the construction-integration model (Kintsch, 1998) describes processes that operate on a network of propositions. These basic representations can just be descriptions of what the human memory looks like, in order for the upper models to be explicitly stated, but they can also be operationalized so that the model can be tested on a computer. In the first case, these representations are usually designed by hand, but this method prevents large-scale simulations.

This was the case with Kintsch's construction-integration model until 1998. Before that, researchers had to code propositions by hand and guess relevant values to code the strength of links between nodes. Then Kintsch (1998) used the Latent Semantic Analysis (LSA) model (Deerwester et al., 1990; Landauer et al., 1998) which provides a way to automatically build these basic representations. This was a major step since the construction-integration processes could then be tested on a large variety of inputs, while being less dependent on idiosyncratic codings. Such a mechanism for automatically constructing basic semantic representations should be carefully designed and tested in order to simulate as good as possible human semantic memory.

LSA is nowadays considered as a good candidate for modeling an adult semantic memory based on a large corpora of representative texts: Bellissens et al. (2002), Kintsch (2000) and Lemaire & Bianco (2003) used it for modeling metaphor comprehension; Pariollaud et al. (2002) used it for modeling the comprehension of idiomatic expressions; Howard & Kahana (2002) relied on it to model free recall and episodic memory retrieval; Laham (1997) did the same for modeling categorization processes; Landauer & Dumais (1997) designed a model of vocabulary acquisition based on LSA; Lemaire & Dessus (2001), Rehder et al. (1998) and Wolfe et al. (1998) used it for modeling knowledge assessment; Quesada et al. (2001) modeled complex problem solving by means of LSA basic representations; Wolfe & Goldman (2003) worked on a model of reasoning about historical accounts based on LSA. However, to our knowledge, no computational basic representations were made that mimic full children's semantic memory.

This paper aims at presenting such a model. First, we present LSA. We then describe our corpus, which is supposed to mimic the kind of texts children are exposed to. Finally, we present three experiments which aim at validating the model.

Latent Semantic Analysis

Basic semantic representations

There are many ways of constructing basic semantic representations that can be processed by a computer. The first one is to build them by hand. Powerful formalisms like description logic (Borgida, 1996) or semantic networks (Sowa, 1991) have been designed to accurately represent concepts, properties and relations. However, in spite of huge efforts (Lenat, 1995), no full set of symbolic representations has been made that can be considered a reasonable model of human semantic memory. Hand-coding semantic information is tedious and, as we mention later, symbolic representations might not be the best formalism for that.

Another strategy is to rely on corpora to get the semantic information. Artificial intelligence researchers have designed sophisticated syntactic processing tools for automatically describing the knowledge using the kind of symbolic formalisms mentioned earlier. They usually refer to them as ontologies or knowledge bases (Vossen, 2003). However, in spite of great strides, this approach still cannot be the means to form the basic semantic representations that cognitive researchers need. First, it cannot be fully automatized, except for specific domains, thus preventing complete descriptions of the language. Second and quite paradoxically, since the descriptions are quite elaborated, it is very hard to design reasoning processes on top of them. For instance, a simple process like estimating the degree of semantic association is very hard to operationalize on complex structures like semantic networks.

Instead of relying on symbolic representations, a third approach consists in (1) analyzing the co-occurrence of words in large corpora in order to draw semantic similarities and (2) relying on very simple structures, namely high-dimensional vectors, to represent meanings. In this approach, the unit is the word. The meaning of a word is not defined per se, but rather determined by its relationships with all others. For instance, instead of defining the meaning of bicycle in an absolute manner (by its properties, function, role, etc.), it is defined by its degree of association to other words (i.e., very close to bike, close to pedals, ride, wheel, but far from duck, eat, etc.). This semantic information can be established from raw texts, provided that enough input is available. This is exactly what human people do: it seems that most of the words we know, we learn by reading (Landauer & Dumais, 1997). The reason is that most words appear almost only in written form and that direct instruction seems to play a limited role. Therefore, we would learn the meaning of words mainly from raw texts, by mentally constructing their meaning through repeated exposure to appropriate contexts.

Relying on direct co-occurrence

One way to mimic this powerful mechanism would be to rely on direct co-occurrences within a given context unit. A usual unit is the paragraph which is both computationally easy to identify and of reasonable size. We would say that:

R1: words are similar if they occur in the same paragraphs.

Therefore, we would count the number of occurrences of each word in each paragraph. Suppose we use a 5,000-paragraph corpus. Each word would be represented by 5,000 values, that is by a 5,000 dimension vector. For instance:

avalanche: (0,1,0,0,0,0,1,0,2,0,0,0,0,0,0,1,1,0,1,0,1,0,0,0,0,0,0…)
snow: (0,2,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,2,1,1,0,1,0,0,0,0,0,0…)

This means that the word avalanche appears once in the 2nd paragraph, once in 7th, twice in the 9th, etc. One could see that, given the previous rule, both words are quite similar: they co-occur quite often. A simple cosine between the two vectors can measure the degree of similarity. However, this rule does not work well (Perfetti, 1998; Landauer, 2002): two words should be considered similar even if they do not co-occur. French & Labiouse (2002) think that this rule might still work for synonyms because writers tend not to repeat words, but use synonyms instead. However, defining semantic similarity only from direct co-occurrence is probably a serious restriction.

Relying on higher-order co-occurrence

Therefore, another rule would be:

R1: words are similar if they occur in similar paragraphs.*

This is a much better rule. Consider the following two paragraphs:

Bicycling is a very pleasant sport. It helps keeping a good health.

For your fitness, you can practice bike. It is very nice and good to your body.

Bicycling and *bike* appear in similar paragraphs. If this is repeated over a large corpus, it would be reasonable to consider them similar, even if they never co-occur within the same paragraph. Now we need to define paragraph similarity. We could say that two paragraphs would be similar if they share words, but that would be restrictive: as illustrated in the previous example, two paragraphs should be considered similar although they do not have words in common (functional words are usually not taken into account). Therefore, the rule is:

R2: paragraphs are similar if they contain similar words.

Rules 1* and 2 constitute a circularity, but this can be solved by a specific mathematical procedure called singular value decomposition, which is applied to the occurrence matrix. This is exactly what LSA does. To state it in other words, LSA is not only based on direct co-occurrence, but rather on higher-order co-occurrence. Kontostahis & Pottenger (2002) have shown that these higher-order co-occurrences do appear in large corpora.

LSA consists in reducing the huge dimensionality of direct word co-occurrences to its best N dimensions. All words are then represented as N-dimensional vectors. Empirical tests have shown that performance is maximal for N around 300 for the whole general English language (Landauer et al., 1998; Bellegarda, 2000) but this value can be smaller for specific domains (Dumais, 2003). We will not describe the mathematical procedure which is presented in details elsewhere (Deerwester, 1990; Landauer et al., 1998). The fact that word meanings are represented as vectors leads to two consequences. First, it is straight-forward to compute the semantic similarity between words, which is usually the cosine between the corresponding vectors, although others similarity measures can be used. Examples of semantic similarities between words from a 12.6 million word corpus are (Landauer, 2002):

cosine(doctor, physician) = .61
cosine(red, orange) = .64

Second, sentences or texts can be assigned a vector, by a simple weighted linear combination of their word vectors. This is a powerful feature of a semantic representation to be able to go easily from words to texts. An example of semantic similarity between sentences is:

cosine(the cat was lost in the forest, my little feline disappeared in the trees) = .66

Modeling children's semantic memory

Semantic space

As we mentioned before, our goal was to rely on LSA to define a reasonable approximation of children's semantic memory. This is a necessary step for simulating a variety of children cognitive processes.

LSA itself obviously cannot form such a model: it needs to be applied to a corpus. We gathered French texts that approximately correspond to what a child is exposed to:

stories and tales for children (~1,6 million words), children's productions (~800,000 words), reading textbooks (~400,000 words) and children's encyclopedia (~400,000 words). This corpus is composed of 57,878 paragraphs for a total of 3.2 million word occurrences. All punctuation signs were ruled out, capital letters were transformed to lower cases, dashes were ruled out except when forming a composed word (like *tire-bouchon*). This corpus was analyzed by means of LSA and the occurrence matrix reduced to 400 dimensions, which appears to be an optimal value as we will see later. The resulting semantic space contains 40,588 different words. This step took 15 minutes on a 2.4 Ghz computer with 2 Gb RAM.

Tests

In order to test whether this semantic space can be an acceptable approximation of the semantic memory of children, we tested three features: its *extent*, its *organization* and its *use*. For each one, we relied on a specific task and compared the data from the simulation of the task to data obtained from children on the exact same task.

The *extent* feature has to do with the size of lexical knowledge. Does our semantic space *knows* the kind of words that a child knows? We used a vocabulary task for that: given a word, the goal is to find the correct definition from four of them. By comparing the model data with children's data at various ages, our goal is to approximately identify the kind of children we are mimicking.

The *organization* feature concerns the way words are associated to others in memory. Do we correctly mimic the semantic neighborhood of words? The task we used for testing that feature is an association task :given a word, the goal is to provide the most associated one. We will compare children's association norms to association measures in the semantic space.

The *use* feature has to do with the way semantic memory is used. Is our semantic space adequate enough so that it can account for a process that uses it? We used a recall task for studying the text comprehension process which obviously largely relies on semantic representations.

These three experiments cover different tasks and different grain sizes of language entities, from words to texts: the first one consists of word comparisons, the second one compares a word and a sentence and the third one compares texts. We expect a good match between human data and model data. In addition, we hypothesize that results will be higher with our children corpus than with adult corpora.

Experiment 1

The first experiment, which aims at validating the model, involves a vocabulary task. The design of the material as well as the experiments with children were realized by Denhière et al. (in preparation). Material consists of 120 questions, each one composed of a word and four definitions: the correct one, a close definition, a far definition and an unrelated definition. For instance, given the word *nourriture* (*food*), translations of the four definitions are:

- *what is used to feed the body (correct);*
- *what can be eaten (close)*;
- *matter which is being spoiled (far)*;
- *letter exchange (unrelated).*

Participants were asked to select what they thought was the correct definition. This task was performed by four groups of children: 2nd grade, 3rd grade, 4th grade and 5th grade. These data were compared with the cosines between the given word and each of the four definitions. For instance, the four cosines on the previous examples were: .38 (correct), .24 (close), .16 (far) and .04 (unrelated). 116 questions were used because the semantic space did not contain four rare words.

The first measure we used was the percentage of correct answers. Figure 1 displays the results. The percentage of correct answers is .53 for the model, which is exactly the same value as the 2nd grade children. Except for unrelated answers, the model data globally follow the same pattern as the children's data.

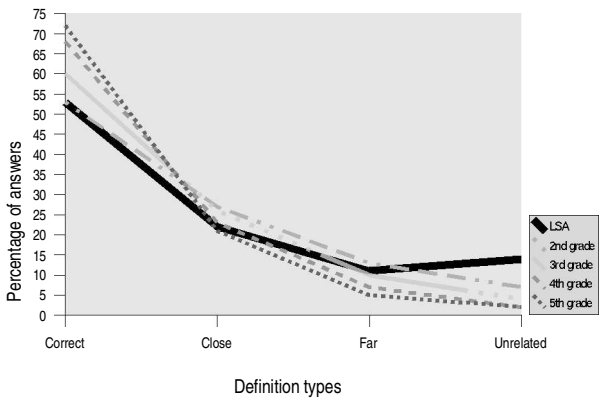


Figure 1: Percentage of answers for different types of definitions

In order to compare our semantic spaces with adult semantic spaces, we defined a measure which integrates the four values. We used a d measure, which is a normalized difference between the cosines for correct and close definitions together and the cosines for far and unrelated definitions together. The higher this measure, the better the result. Given a word W, four definitions (correct, close, far and unrelated) and a global standard deviation S, the formula is the following:

$$d = \frac{\frac{\cos(W, correct) + \cos(W, close)}{2} - \frac{\cos(W, far) + \cos(W, unrelated)}{2}}{S}$$

We also compared these results with several adult corpora, in order to test whether our semantic space was specific to children. We used five corpora: a literature corpus, composed of novels from the XIXth and XXth centuries and four corpora from the French daily newspaper *Le Monde*, of the years 1993, 1995, 1997 and 1999. Table 1 shows the results.

Table 1: Comparison between children's semantic space and adult semantic spaces

Semantic space	Size (in million words)	Percentage of correct answers	***d***
Children	3.2	.53	.69
Literature	14.1	.38	.52
Le Monde 1993	19.3	.44	.23
Le Monde 1995	20.6	.37	.21
Le Monde 1997	24.7	.40	.28
Le Monde 1999	24.2	.34	.25

In accordance with the previous experiment, the children's semantic space has the better results, although its size is much smaller. Student tests have shown that the children semantic space is significantly different from others ($p < .05$) except for the percentage of correct answers when compared to the *Le Monde* 1993 corpus ($p < .1$).

Experiment 2

This second experiment is based on verbal association norms published by de La Haye (2003). Two-hundred inducing words (144 nouns, 28 verbs and 28 adjectives) were proposed to 9 to 11-year-old children. For each word, participants had to provide the first word that came to their mind. This resulted in a list of words, ranked by frequency. For instance, given the word *cartable* (*satchel*), results are the following for 9-year-old children:

- école (*school*): 51%
- sac (*bag*): 12%
- affaires (*stuff*): 6%
...
- classe (*class*): 1%
- sacoche (*satchel*): 1%
- vieux (*old*): 1%

This means that 51% of the children answered the word *école* (*school*) when given the word *cartable* (*satchel*). The two words are therefore strongly associated for 9-year-old children. These association values were compared with the LSA cosine between word vectors: we selected the three best-ranked words as well as the three worst-ranked (like in the previous example). We then measured the cosines between the inducing word and the best ranked, the 2nd best-ranked, the 3rd best ranked, and the mean cosine between the inducing word and the three worst-ranked. Results are presented in Table 2.

Table 2: Mean cosine between inducing word and various associated words for 9-years-old children

Words	Mean cosine with inducing word
Best-ranked words	.26
2nd best-ranked words	.23
3rd best ranked-words	.19
3 worst-ranked words	.11

Student tests show that all differences are significant ($p < .03$). This means that our semantic space is not only able to distinguish between the strong and weak associates, but can also discriminate the first-ranked from the second-ranked and the latter from the third-ranked.

Measure of correlation with human data is also significant ($r(1184 =.39,\ p<.001$). Actually, two factors might have lowered this result. First, although we tried to mimic what a child has been exposed to, we could not control all word frequencies within the corpus. Therefore, some words might have occurred with a low frequency in the corpus, leading to an inaccurate semantic representation. When the previous comparison was performed on the 20% most frequent words, the correlation was much higher ($r(234 =.57$, $p<.001$).

The second factor is the participant agreement: when most children provide the same answer to an inducing word, there is a high agreement, which means that both words are very strongly associated. However, there are cases when there is almost no agreement: for instance the three first answers to the word *bruit* (*noise*) are *crier* (*to shout*) (9%), *entendre* (*to hear*) (7%) and *silence* (*silence*) (6%). It is not surprising that the model corresponds better to the children data in case of a high agreement, since this denotes a strong association that should be reflected in the corpus. In order to select answers whose agreement was higher, we measured their entropy. The formula is the following:

$$entropy(item) = \sum_{answer} freq(answer).\log(\frac{1}{freq(answer)})$$

A low entropy corresponds to a high agreement and vice versa. When we selected the 20% items with the lowest entropy, the correlation also raises ($r(234)=.48$, $p<.001$).

All these results show that the association degree between words defined by the cosine measure within the semantic space seems to correspond quite well to children's judgement of association.

We also compared these results with the previous adult semantic spaces. Results are presented in Table 3.

Table 3: Correlations between participant child data and different kinds of semantic spaces

Semantic space	Size (in million words)	Correlation with child data
Children	3.2	.39
Literature	14.1	.34
Le Monde 1993	19.3	.31
Le Monde 1995	20.6	.26
Le Monde 1997	24.7	.26
Le Monde 1999	24.2	.24

In spite of much larger sizes, all adult semantic spaces correlate worse than the children's semantic space with the data of the participants in the study. Statistical tests show that all differences between the child model and the other semantic spaces are significant ($p<.03$).

Experiment 3

The third experiment is based on recall or summary tasks. Children were asked to read a text and write out as much as they could recall, immediately after reading or after a fixed

delay. We used 7 texts. We tested the ability of the semantic representations to estimate the amount of knowledge recalled. This amount is classically estimated by means of a propositional analysis: first, the text as well as the participant production are coded as propositions. Then, the number of text propositions that occur in the production is calculated. This measure is a good estimate of the knowledge recalled. Using our semantic memory model, this is accounted for by the cosine between the vector representing the text and the vector representing the participant production.

Table 4 displays all correlations between these two measures. They range from .45 to .92, which means that the LSA cosine applied to our children's semantic space is a good estimate of the knowledge recalled.

Table 4: Correlations between LSA cosines and number of propositions recalled for different texts.

Story	Task	Number of participants	Correlations
Poule	Immediate recall	52	.45
Dragon	Delayed recall	44	.55
Dragon	Summary	56	.71
Araignée	Immediate recall	41	.65
Clown	Immediate recall	56	.67
Clown	Summary	24	.92
Ourson	Immediate recall	44	.62
Taureau	Delayed recall	23	.69
Géant	Summary	105	.58

In an experiment with adults, Foltz et al. (1996) have shown that LSA measures can be used to predict comprehension. Besides validating our model of semantic memory, this experiment shows that an appropriate semantic space can be used to assess text comprehension in a much faster way than propositional analysis, which is a very tedious task.

Conclusion

A model of the development of children's semantic memory

Our model is not only a computational model of children's semantic memory, but of its *development*. Other computational models of human memory have been developed but some of them are based on inputs that do not correspond to what humans are exposed to. They are good models of the memory itself, but not of the way it is mentally constructed. In order to be cognitively plausible, models of the construction of semantic memory need to be approximately based on the kind of input to humans.

LSA is such model. Its performance is similar to those of human people. It needs an input of a few million words, which is comparable to what humans are exposed to (Landauer & Dumais, 1997). On the contrary, PMI-IR (Turney, 2001) is a good model of semantic similarities, even better than LSA in modeling human judgement of synonymy, but it is based on an input of thousands of millions of words, since it relies on all the texts published on the web. This is of course cognitively unplausible. HAL (Burgess, 1998) is another model of human memory. It is quite similar to LSA except that it does not rely on a dimension reduction step. It is currently based on a corpus of 300 million words, which is closer to the human inputs than PMI-IR, although this could be considered quite overestimated.

Further investigations

Our semantic space provides a means for researchers studying children's cognitive processes to design and simulate computational models on top of these basic representations. In particular, computational models of text comprehension could be tested using the basic semantic similarities that the space provides. It would also be possible to investigate the development of semantic memory by looking at the evolution of various semantic similarities according to the size of the corpus in detail. In particular, Landauer & Dumais (1997) claim that we learn the meaning of a word through the exposition to texts that do not contain it. Our semantic space gives the opportunity to test this assertion by checking the kind of paragraphs that cause an increase of similarity through incremental exposure to the corpus.

Improvements

Our semantic space could be improved in many ways. Its composition (50% stories, 25% production, 12.5% reading textbooks, 12.5% encyclopedia) is very rough and work has to be done to better know the amount and nature of texts that children are exposed to. Several studies led us to think that lemmatization could significantly improve the results, especially for the French language that has so many forms for some verbs. We did perform the previous experiments on a lemmatized version of the corpus (using the Brill tagger on the French files developed by ATILF, and the Flemm lemmatizer written by Fiametta Namer). Results were worse than with the non-lemmatized version. In order to know more about this surprising result, we distinguished between verbs and nouns. We found that the overall decrease is mainly due to a decrease for the nouns. One reason could be that the singular and plural forms of a noun are not arguments of the same predicates. For instance, the word *vague* (*wave*) is generally used in its plural form in the context of the *sea*, but more frequently in the singular form in its metaphorical meaning (*a wave of success*). Therefore, if both forms are grouped into the same one, this affects the co-occurrence relations and modifies the semantic representations.

Another way of improvement would have to deal with syntax. LSA does not take any syntactic information into account: all paragraphs are just bags of words. A slight improvement would consist in considering a more precise unit of context than a whole paragraph. A sliding context window (like in the HAL model for instance) would take into account the local context of each word. This might improve the semantic representations, while being cognitively more plausible. We are working in that direction.

For the moment, our model is an estimation. We cannot precisely identify to which age it corresponds. Our goal is to stratify it so that we would have a model for each age. Developmental models would then be able to be simulated.

Acknowledgements

This work was done while the second author was in sabbatical at the university of Aix-Marseille. We would like to thank D. Chesnet, E. Lambert and M.-A. Schelstraete, for providing us with parts of the corpus, F. de la Haye for the association data as well as M. Bourguet and H. Thomas for the design of the vocabulary test. We also thank P. Dessus and E. de Vries for their comments on a previous version.

References

Bellegarda, J. (2000) Exploiting Latent Semantic Information in statistical language modeling. *Proceedings of IEEE*, 88 (8), 1279-1296.

Bellissens, C., Thiesbonenkamp, J. & Denhière, G. (2002). Property attribution in metaphor comprehension: simulations of topic and vehicle contribution within the LSA-CI framework. *Twelfth Annual Meeting of the Society for Text and Discourse*, June 2002.

Borgida, A. (1996). On the relative expressive power of description logics and predicate calculus, *Artificial Intelligence, 82*, 353-367.

Burgess, C. (1998). From simple associations to the building blocks of language: modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30, 188-198.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by Latent Semantic Indexing. *Journal of the American Society for Information Science* , 41-6, 391-407.

de la Haye, F. (2003). Normes d'associations verbales chez des enfants de 9, 10 et 11 ans et des adultes. *L'Année Psychologique*, 103, 109-130.

Dumais, S. D. (2003). Data-driven approaches to information access. *Cognitive Science 27* (3), 491-524.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28-2, 197-202.

French, R.M. & Labiouse, C (2002). Four problems with extracting human semantics from large text corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, NJ:LEA.

Howard, M.W. & Kahanna, M.J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language* 46, 85-98.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review,* 95, 163-182.

Kintsch, W. (2000). Metaphor comprehension: a computational theory. *Psychonomic Bulletin & Review,* 7-2, 257-266.

Kontostathis, A. & Pottenger, W.M. (2002). Detecting patterns in the LSI term-term matrix. *Workshop on the Foundation of Data Mining and Discovery, IEEE International Conference on Data Mining*.

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (p. 979). Mawhwah, NJ: Erlbaum.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes* , 25, 259-284.

Landauer (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.), *The psychology of Learning and Motivation, 41*, 43-84.

Lemaire, B & Dessus, P. (2001). A system to assess the semantic content of student essay. *Journal of Educational Computing Research,* 24-3, 305-320.

Lemaire, B. & Bianco, M. (2003). Contextual effects on metaphor comprehension: experiment and simulation. *Proc. of the 5th International Conference on Cognitive Modeling (ICCM'2003)*, Bamberg, Germany.

Lenat, D.B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Comm. of the ACM*, 11, 32-38.

Pariollaud, F., Denhière, D. & Verstiggel, J.C. (2002) Comprehension of idiomatic expressions: effect of meaning salience. *Proc. of the 9th Int. Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Annecy, July 1-5.*

Perfetti, C. A. (1998). The limits of co-occurrence: tools and theories in language research. *Discourse Processes, 25,* 363-377.

Quesada, J., Kintsch, W., & Gomez E. (2002) A theory of complex problem solving using Latent Semantic Analysis. In W. D. Gray & C. D. Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates, Mahwah, NJ.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K. & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: some technical considerations. *Discourse Processes*, 25, 337-354.

Sowa, J.F. (1991). *Principles of Semantic Networks: Exploration in the Representation of Knowledge*, Morgan Kaufmann.

Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In De Raedt, L. and Flach, P., (Eds). *Proceedings of the 12th European Conference on Machine Learning (ECML-2001)* , 491-502, Freiburg.

Vossen, P. (2003). Ontologies. In R. Mitkov (Ed.) *The Oxford Handboof of Computational Linguistics.* Oxford, 464-482.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch & W., Landauer, T. K. (1998). Learning from text: matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, 25, 309-336.

Wolfe, M.B.W. & Goldman, S.R. (2003) Use of Latent Semantic Analysis for predicting psychological phenomena: two issues and proposed solutions. *Behavior Research Methods Instruments & Computer*, 35(1), 22-31.

Learning relations between concepts: classification and conceptual combination

Barry Devereux (Barry.Devereux@ucd.ie), Fintan Costello (Fintan.Costello@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland.

Abstract

People interpret noun-noun compounds like "wind power" by inferring a relational link between the compound's two constituent concepts. Various studies have examined how people select the best relation for a compound from a set of candidate relations. However, few studies have investigated how people learn such relations in the first place. This paper describes an experiment examining how people learn which relations are possible between concepts. Participants in this experiment learned artificial, laboratory controlled relations between pairs of items and then judged how likely those relations were for new pairs of items. The results showed that people's judgement of relation likelihood was reliably influenced by the presence of facilitating features for relations and by the diagnosticity of features for relations. A simple exemplar-based model of classification, using both diagnostic and facilitating features, was applied to people's judgements of relation likelihood. This model accurately predicted people's judgements of relation likelihood in the experiment, using no free parameters to fit the data.

Introduction

When, in everyday discourse, people encounter noun-noun compounds such as "mountain stream" or "lake boat", they interpret those compounds by inferring a relation that can be used to combine the two constituent concepts (inferring that a "mountain stream" is a stream that flows down a mountain, that a "lake boat" is a boat that sails on a lake). In theoretical accounts of conceptual combination, this process involves selecting the best relation for a compound from a set of candidate relations. Some theories give a standard set of candidate relations to be used in all compounds (Gagné & Shoben, 1997), while others derive candidate relations from the internal structure of the concepts being combined (Costello & Keane, 2000; Wisniewski, 1997). Many studies have investigated how people select the best relation for a given compound (e.g. Costello & Keane, 2001, Wisniewski, 1996). However, there have been very few studies investigating how people learn and form these relations in the first place. In this paper we aim to fill this gap.

The paper describes an experiment investigating how people learn relations between two sets of novel concepts. In the experiment we designed four different relations that could hold between artificial, laboratory-generated 'beetle' and 'plant' concepts. Participants learned these relations from sets of examples, with each example showing one sort of relation holding between one type of plant and one type of beetle. After learning, participants were shown new pairs of plants and beetles, and asked to say which of the four learned relations could hold between those two items.

This experiment was designed to examine two different possible factors in people's learning of relations between concepts: the presence of *diagnostic* features for those relations, and the presence of *facilitating* features. By diagnostic features for a relation we mean features of a constituent concept that are strongly associated with a particular relation. Diagnostic features are most familiar in the case of single concepts: for example "has four legs" and "is made of wood" are diagnostic features for the single concept *chair*: most things that are chairs have those features, and most things that are not chairs do not. Similarly, the feature "has a flat surface raised off the ground" might be diagnostic for the relation *is-sat-on-by*: most instances of the *is-sat-on-by* relation have that feature; most instances of other relations do not. In the experiment we asked whether people would use the diagnostic features for relations when selecting likely relations for beetle-plant pairs.

By facilitating features we mean the features of a pair of concepts that are necessary for a given relation to be possible, and without which that relation cannot hold. For example, while the compound "steel chair" can easily be interpreted using the *made-of* relation, the compound "kitchen chair" cannot possibly be interpreted as "a chair made of kitchens" simply because kitchens are not a type of substance. Being a substance is a necessary facilitating feature for an item to take part in the *made-of* relation. Again, in the experiment we asked how people would use such facilitating features when selecting likely relations for beetle-plant pairs.

This paper is organised as follows. In the next section we discuss the representation of relations in terms of sets of examples, as used in our experiment. We then describe the experiment in detail. To foreshadow the results, we found that both diagnostic and facilitating features had a reliable influence on people's selection of likely relations for pairs of items. We then describe how an exemplar-based model of concept conjunction (Costello, 2000, 2001) can be applied to the results of this experiment, giving a close fit to people's judgements of relation likelihood. Finally, we conclude by discussing the implications of our findings for theories of conceptual combination.

Learning Relations from Exemplars

Our primary assumption is that the relations selected during conceptual combination are essentially categories, just as the concepts that they link are essentially categories. We use an exemplar representation to describe these relational categories. Exemplar theories of classification, which propose that a category is represented as the set of remembered instances of that category and that new items are classified on the basis of their similarity to those instances (e.g. Medin & Schaffer, 1978; Nosofsky, 1984), have successfully accounted for a number of patterns seen in people's learning of single categories. We extend the exemplar approach to allow both relations and the concepts that they link to be represented by sets of instances.

In a category representing a single concept, each exemplar consists simply of a single set of features. For a category representing a relation, however, each exemplar consists of two sets of features: the features of the two single-category exemplars that are being linked by that relation. For example, suppose we have two categories *A* and *B* consisting of the set of exemplars $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ and $\{b_1, b_2, b_3, b_4, b_5\}$ respectively. Each category represents a single concept, and each exemplar contains features describing one example of that concept. We can compute the membership of a given item in category *A*, for example, by comparing that item to the set of stored exemplars of category *A*. Then a relation *R* linking the concepts *A* and *B* might be represented as the set of exemplars of that relation, for example $\{(a_1, b_1), (a_1, b_2), (a_3, b_3), (a_4, b_4), (a_4, b_5)\}$. Regarding *R* as a category, we can compute the membership of any pair of items (x, y) in the relation *R* by comparing that pair to the set of exemplars of that relation. If more than one relation is defined, we can compute membership in each of the relations and make assertions about which relation the given pair of items is most likely to belong to.

This representation of relations in terms of a collection of pairs of category exemplars is motivated by how mathematical relations are defined in set theory. In set theory, binary relations are defined as sets of ordered pairs. To take a well-known example, the "is equal to" relation is a set denoted by =, and is defined on the integers to be the set {…, (-1, -1), (0, 0), (1, 1), (2, 2), …}. Thus the "is equal to" relation holds between two integers *x* and *y* if and only if the ordered pair (x, y) is a member of the set denoted =. We extend this set-theoretic idea of relations and propose that a relation between two concepts can be represented as a set of relation exemplars, where each of these relation exemplars is an ordered pair of category exemplars. Membership of a pair of items in a relational category is then computed by comparing that pair of items to the stored exemplars of that relational category, as is precisely the case for exemplar models of simple categories.

This approach assumes that when people are selecting the correct relation for a pair of items they may be performing a classification task in which they compare that pair of items to various sets of relation exemplars. What factors would we expect people to be influenced by in such a classification task? First, we would expect people pay attention to the features in items that are diagnostic for particular relations (that is, features that are present in most of the items that take part in the relation, and absent in most items that do not take part in the relation). If a particular feature is diagnostic for a relation, people should use that feature to identify new items likely to take part in that relation. Such a result would be consistent with other results in the classification literature, which reveal that people are attentive to diagnostic features when making determinations of category membership.

Second, we would expect people to pay attention to whether or not a given item has the facilitating features required for a given relation (as in our "kitchen chair" example). If a particular feature is present in every item that takes part in a certain relation, then we can assume that that feature may be necessary for that relation to take place: the feature may facilitate that relation. When confronted with a new pair of items which do not possess that facilitating feature, we would expect people not to select that relation for that pair of items. Note that a facilitating feature for a given relation may also be a diagnostic feature for that relation (if it occurs in every item that takes part in that relation *and* in no items that do not take part in that relation). However, a facilitating feature for a relation may also be non-diagnostic for that relation (if the feature occurs in every item that takes part in that relation, but also occurs in many items that do not take part in that relation). Next we describe an experiment examining the influence that facilitating and diagnostic features have on people's selection of relations for pairs of items.

A Categorisation Experiment

This experiment aims to test three hypotheses: that people can learn relations from sets of examples of those relations; that diagnostic features are important in people's selection of relations for pairs of items; and that facilitating features are also important in relation selection. The design of the experiment is essentially the same as other experiments in the category learning literature: a preliminary training phase where participants are exposed to exemplars of different artificial, laboratory controlled categories is followed by a test phase where participants are presented with new items and are asked to make judgements of category membership. The categories in this experiment are four different relations that can hold between pairs of objects. Each of the training items consists of two objects linked by one of these relations. The test items also consist of two objects; however these objects are not linked by any relation, and participants are asked to judge the likelihood of different relations holding between these items.

To examine the role of facilitating features in relation selection, the training items were designed so that two of the learned relations had facilitating features: we called these two relations the facilitated relations. Every time one of these relations occurred in the set of training items, a particular beetle or plant feature (the facilitating feature) was also present in that item. The other two relations did not have facilitating features: we called these two relations the independent relations. Similarly, the training items were designed so that some beetle and plant features were particularly diagnostic for some relations, and some features were not. In the training items, the diagnostic features for a particular relation occurred most frequently in beetle or plants taking part in that relation, and occurred rarely in other relations. The pairs of objects used in the test phase of the experiment consisted of various combinations of facilitating and diagnostic features for different relations. By examining participants' choice of relations to link the objects in these test items, we can then assess the influence of facilitating and diagnostic features in relation selection.

Method

Participants. 16 postgraduate students or recent college graduates volunteered to take part in the experiment. All were native speakers of English.

Materials. The materials for the training phase consisted of 18 visual stimuli on an A5-sized card depicting a cartoon

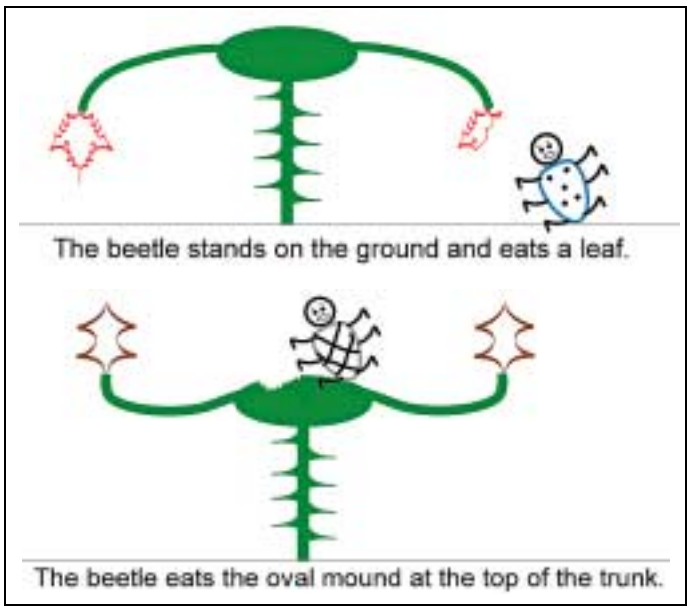


Figure 1: Two training phase stimuli.

beetle eating a plant. The beetles varied on three quaternary-valued feature dimensions: color of the shell, pattern on the shell, and facial expression. The plants varied on four feature dimensions: color of the leaves (quaternary-valued), shape of the leaves (quaternary-valued), droop of the branches (binary-valued), and whether there were buds or thorns on the trunk (binary-valued). There were four possible ways in which a beetle could eat a plant, corresponding to the four relational categories: the beetle could land on a leaf of a plant and eat the leaf (an independent relation), the beetle could eat from the top of the trunk of the plant (an independent relation), the beetle could eat from the trunk of the plant if there were buds rather than thorns on the trunk (a facilitated relation), or the beetle could stand on the ground and eat the leaf of a plant that had drooping branches (a facilitated relation). Underneath each picture was a sentence describing the eating behaviour that was talking place. Examples of the training phase's stimuli are presented in Figure 1.

These 18 items described a category structure for four relations, each relation being one of the different ways in which a beetle could eat a plant. The distribution of beetle and plant features was controlled so that some features would be facilitating for relations and so that some features would be diagnostic for relations. The distribution of beetle and plant features across the four different relation categories is shown in abstract form in Table 1. The numerical values in columns B1, B2 and B3 represent the different possible features that beetles could have; the values in columns P1, P2, P3, and P4 similarly represent the different possible features of plants. Each row in this table represents a different particular exemplar of one of the four relations R1, R2, R3 and R4.

In this experiment, we were interested in the influence which the distribution of features across items would have on people's selection of relations between items. We were not concerned with any effect which the physical properties of stimuli (e.g. the salience of different colours, the distinctiveness of different shapes) would have on people's

Table 1: The abstract relational category structures used in the training phase.

Item	Relation	Insect Features			Plant Features			
		B1	B2	B3	P1	P2	P3	P4
1	R1	1	1	1	1	1	1	2
2	R1	4	1	1	1	1	1	1
3	R1	2	2	2	4	1	1	2
4	R1	3	4	2	1	4	2	1
5	R1	3	3	3	1	1	1	1
6	R2	1	1	2	3	2	1	2
7	R2	2	2	2	3	2	1	1
8	R2	2	3	4	1	1	1	1
9	R2	3	4	3	2	3	2	2
10	R3	1	4	1	4	2	2	1
11	R3	1	1	4	2	2	2	1
12	R3	2	2	2	2	3	2	2
13	R3	3	2	4	2	3	2	1
14	R3	2	2	3	3	2	2	1
15	R4	1	3	1	4	3	2	2
16	R4	2	3	3	2	2	1	2
17	R4	3	2	3	3	3	1	2
18	R4	3	3	3	3	3	1	2

relation selection. Thus, while each participant's set of training items had the abstract structure shown in Table 1, each participant saw a unique set of physical stimuli. The abstract-to-physical mappings for the category dimensions and values for the two independent relations and for the two facilitated relations were balanced across participants. This was done so that the physical dimensions, values and relations would not be confounded with their abstract counterparts.

Facilitating and Diagnostic Features. The four relations in Table 1 were designed so that two relations had facilitating features (R3 and R4) and two did not (R1 and R2), and two relations had highly diagnostic features (R1 and R3) and two had less diagnostic features (R2 and R4). Relations R1 and R2 were the independent relations, while relations R3 and R4 were the facilitated ones. P3 is the facilitating dimension for relation R3: every exemplar of R3 involves an item with a value of 2 on dimension P3. Similarly P4 is the facilitating dimension for relation R4: every exemplar of R4 involves an item with a value of 2 on dimension P4. (In the experimental materials, the facilitating features were instantiated in a causally meaningful way. For example, the physical relation "the beetle stands on the ground and eats the leaf" had the facilitating feature "drooping branches on the plant").

Of the two independent relations, R1 has highly diagnostic features while R2 has less diagnostic features. Relation R1 has two highly diagnostic features: a 1 on P1 and a 1 on P2. Relation R2, however, has no particularly diagnostic features. (R2 is therefore a very vague category, not well distinguished by either diagnosticity or facilitating features). Of the two facilitated relations, R3 has highly diagnostic features while R4 has less diagnostic features. For relation R3, a 2 on

dimension P3 is a highly diagnostic feature for that relation, occurring five out of five times in examples of that relation and only three times outside it. (Note that this feature, a 2 on P3, is also the facilitating feature for relation R3). Relation R4, however, has no such highly diagnostic feature, although a 3 on B2 and a 3 on B3 are both moderately diagnostic for that relation. (R4's facilitating feature is not very diagnostic, occurring four times within the category and five times outside it.)

The materials for the test phase consisted of more visual stimuli depicting beetles and plants; however in these pictures the beetles and plants were shown separately, without any eating or any other interaction taking place. Underneath each test picture was the question "How likely are the different types of eating behavior?", followed by the four relation description sentences, each of which was accompanied by a 7-point scale ranging from -3 (labelled "not at all likely") to +3 ("extremely likely"). The order in which the four scales were presented was balanced across participants.

The test phase consisted of 29 beetle-plant pairs. Of these, nine pairs were selected from the 18 beetle-plant pairs that had been presented in the training phase, but now without the eating behaviour shown. These nine previously-seen pairs were used to assess how accurately participants learned the training items they had studied. The remaining 20 test items (beetle-plant pairs) had not been seen previously by participants. For these items the properties of interest are whether or not the facilitating feature of relation R3 or of relation R4 is present, and whether or not the item had features diagnostic for particular relations.

Procedure. The experiment consisted of two sections: a training phase where participants studied the training items (pairs of beetles and plants taking part in particular relations), and a test phase where they had to rate the likelihood of the different possible relations for a sequence of beetle-plant pairs. Participants were asked to pretend to be biologists interested in learning about imaginary plants and beetles and the relationship between them. The seven dimensions on which the beetles and plants could vary were explicitly pointed out to participants. It was pointed out to participants that they might find it useful to try to learn about the eating behaviour by looking for relationships between features and types of eating, or by learning the features of individual examples and remembering the type of eating occurring with them. Participants spent about five minutes reading the instructions, during which time the experimenter answered any questions they had. After reading the instructions participants were presented with the 18 training items at a large desk area. Participants were given 12 to 15 minutes to study the training items.

After the training phase, the 18 training items were removed and participants were given the 29 test items. Participants were first shown the nine test items corresponding to items they had studied in the first part of the experiment. Participants were told to mark an integer value on each of the four scales describing how likely they felt the four possible types of eating behaviour were. Following these nine items 20 new test items were presented to the participants. The order in which the items were presented was randomized for each participant, and participants were allowed to rate the items at their own pace.

Results

Participants' learning of the training items. For these nine items there was a "correct" relation (each item was a member of one category during the learning phase) and three "incorrect" relations (corresponding to the other three categories). The responses for each relation and each experimental item were classified as either positive (> 0) or non-positive (≤ 0), depending on how the participant responded on each scale. On average, participants gave a positive rating to correct relations 71% of the time and a positive rating to incorrect relations 33% of the time. Two participants gave a positive score to only four correct relations; these two participants were excluded from the analysis. The remaining 14 participants gave a positive rating to correct relations 75% of the time and a positive rating to incorrect relations only 25% of the time. These results indicate that participants learned to distinguish between the categories in the training phase.

Participants' sensitivity to facilitating features. One-tailed binomial tests with $\alpha = 0.05$ were used to identify whether the presence or absence of the facilitating features for a relation had an effect on how participants responded when grading the likelihood of that relation. The proportion of positive responses for each of the four relations was the statistic of interest.

First we considered the items in which the facilitating features for a given relation were absent. Of the 29 test items, 16 were items for which the facilitating feature for relation R3 was absent and 16 were items for which the facilitating feature for relation R4 was absent. For relation R3, the binomial test was significant for 13 of the 14 participants; in other words, 13 of the 14 participants were significantly more likely to produce a non-positive rather than a positive response to relation R3 when the facilitating feature for relation R3 was absent. (Indeed, 10 participants *never* produced a positive response). For relation R4, 10 of the 14 participants were significantly more likely to produce a non-positive rather than a positive response. (Here, five participants never produced a positive response).

A similar analysis was performed looking at the items where the facilitating feature was present. Of the 29 test items, 13 were items for which the facilitating feature for relation R3 was present and 13 were items for which the facilitating feature for relation R4 was present. For relation R3, 8 of the 14 participants were significantly more likely to produce a positive rather than a non-positive response. For relation R4, 7 of the 14 participants were significantly more likely to produce a positive rather than a non-positive response. These results are sensible considering that in many cases participants will rate a relation as having low likelihood for a given item, even when that relation's facilitating feature is present in the item: the facilitating feature doesn't mean that the relation must be selected for this item, only that it is a possibility. The difference in these results between items that had and items that had not the facilitating feature for a relation indicate that participants were highly sensitive to the presence and absence of those features.

Participants' sensitivity to diagnostic features. The diagnosticity of a feature for a category is a measure of how

good that feature is at identifying membership of that category. If a feature appears in many items in a category and few items outside a category then that feature will have high diagnosticity. More formally, we can define the diagnosticity of a feature *f* for a category *C* to be

$$D(f,C)=\frac{|C\cap E_f|}{|C\cup E_f|} \qquad \text{(Eq. 1)}$$

where E_f denotes the set of exemplars that have feature *f*. Using this formula we calculated the average diagnosticity of the features of each of the 29 test items for each of the four relational categories and compared this to the observed data. For two of the four relations, the amount of diagnosticity for items had a high correlation with the observed membership ratings for the items (for R1, $r = 0.83$, $p < 0.01$, *%var* = 69%; for R4, $r = 0.81$, $p < 0.01$, *%var* = 65%). For the other two relations, the correlation was less strong though still significant (for R2, $r = 0.66$, $p < 0.01$, *%var* = 43%; for R3, $r = 0.70$, $p < 0.01$, *%var* = 49%). These results indicate that participants were sensitive to diagnostic features when making their category judgements.

Diagnostic and facilitating features. Relations R1 and R2 were the independent relations: membership in these relations did not depend on facilitating features. Diagnosticity was very important for identifying members of R1 but was not very important for identifying members of R2. We would therefore expect the correlation of diagnosticity to the observed memberships to be higher for relation R1 than relation R2; this is the case in the above analysis of the effect of diagnostic features.

Relations R3 and R4 were the facilitated relations: membership in these relations depended on both diagnostic features and facilitating features. R3 was designed to have highly diagnostic features, while R4 was designed to have less diagnostic features. The diagnosticity analysis above, however, shows that the correlation of the observed memberships with diagnosticity was lower for R3 than for R4. The divergence between diagnosticity and membership ratings for these relations suggests an interaction between diagnosticity and facilitating features.

As a way of examining this interaction we looked at the total number of positive and non-positive responses across all participants for relations R3 and R4; they are presented in Table 2. For cases where the facilitating feature is present, there are less positive responses for R4 than for R3 and more non-positive responses for R4 than R3. Conversely, for cases where the facilitating feature is absent, there are less positive responses for R3 than for R4 and more non-positive responses for R3 than R4. These data suggest that participants are more sensitive to the presence or absence of the facilitating feature for R3 than they are for R4. This is consistent with the fact that the facilitating feature for R3 is more diagnostic for R3 than the facilitating feature for R4 is diagnostic for R4. In other words, people seem to be using both the facilitating nature and the diagnosticity of features together in deciding relation likelihood for these relations. In the next section we investigate this interaction between diagnosticity and facilitating features in more detail by applying a model of classification to our data.

Table 2: Total number of positive and non-positive responses across all participants

	Facilitating Feature Present		Facilitating Feature Absent	
	Positive	Non-positive	Positive	Non-positive
R3	130	52	14	210
R4	123	59	30	194

Modelling Relation Selection

We are interested in how people used facilitating and diagnostic features when making judgements of relation likelihood in our experiment. As we have seen, our results indicate an interaction between diagnosticity and facilitating features. We are also interested in whether or not our view of relations in terms of exemplar-represented categories can successfully account for the results of the above experiment. To explore these issues, we examined whether an exemplar model of categorization could be used to model participants' responses of how likely each relation is for each item in the experiment. We used as our starting point Costello's (2000, 2001) Diagnostic Evidence Model (DEM) which uses diagnostic features to model classification in concept combination. This model calculates an evidence score for an item *x* in a category *C* using the diagnosticity of each feature of the item for that category according to the formula

$$E(x,C)=1-\prod_{f\in F_x}\left(1-D(f,C)\right) \qquad \text{(Eq. 2)}$$

where F_x is the set of features of *x* and $D(f, C)$ is computed as in Equation 1. (Equation 2 essentially sees category membership as a disjunction of the feature diagnosticities and is not dissimilar to simply averaging the diagnosticities as we did in the previous section). As a preliminary step we applied this model to the experimental data without using any information about facilitating features: the model is only sensitive to features' diagnosticity. In this form the model still produces a reasonable fit to the data (for R1, $r = 0.88$, $p < 0.01$, *%var* = 78%; for R2, $r = 0.63$, $p < 0.01$, *%var* = 40%; for R3, $r = 0.73$, $p < 0.01$, *%var* = 53%; for R4, $r = 0.72$, $p < 0.01$, *%var* = 52%), with no free parameters.

The model in this form uses diagnostic information alone. However, the results of our experiment indicate that people make use of both diagnosticity and facilitating features in determining the relations. One possible account of how people use both these types of information is that people are applying diagnosticity information after they have been constrained by the presence or absence of the facilitating features. Perhaps people check if an item has the necessary facilitating features for a particular relation and then, if it does, use the diagnostic evidence of the features. In modifying the model we therefore assume that participants' do not use known exemplars which depict a relation that is impossible for the test item at hand: facilitating features restrict the universe of discourse so that membership of an item in a relational category is a calculation across the subset of the learned exemplars that belong to relational categories that are possible for the current item. Our formula for

diagnosticity then becomes

$$D(f,x,C) = \frac{\left|C \cap E_f \cap R_x\right|}{\left|\left(C \cup E_f\right) \cap R_x\right|} \qquad \text{(Eq. 3)}$$

where R_x is the set of known exemplars that belong to relations that are not impossible given the features of *x*. This modified model gives a much closer fit to the observed relation selection ratings (for R1, $r = 0.89$, $p < 0.01$, *%var* = 79%; for R2, $r = 0.78$, $p < 0.01$, *%var* = 61%; for R3, $r = 0.98$, $p < 0.01$, *%var* = 0.96; for R4, $r = 0.96$, $p < 0.01$, *%var* = 0.92), again with no free parameters. Clearly, both information about the diagnosticity of features and information about the presence or absence of facilitating features are required to accurately model the experimental data.

Though this modified DEM model may not be the best way of modelling the data, it does suggest that using information about both the facilitating features and diagnosticity of features of an item are important in selecting relations. The fact that an exemplar model of classification can predict how people rate the likelihood of different relations linking pairs of items is also evidence in support of the hypothesis that relations can be represented as categories. This suggests that relation selection can be thought of as a kind of classification task.

Conclusions and Discussion

Our study yields three findings. First, people can learn which relations are possible between concepts from sets of examples of those relations. Second, people pay attention to facilitating features for those relations and use those features when judging relation likelihood for new examples. Third, people also pay attention to, and use, diagnostic features for those relations. Such findings are consistent with our hypothesis that relations can be represented with an exemplar category structure, and that the selection of a relation between two constituents can be seen as a categorisation task. These findings have implications for current theories of how relational links are used in conceptual combination. In particular, these findings may be problematic for Gagné & Shoben's (1997) CARIN model, which proposes that in conceptual combination people select the correct relational link between two concepts from a fixed set of 16 relational links called *thematic relations*. First, the thematic relations used in the CARIN model have no internal structure: there is no way, in that model, in which facilitating or diagnostic features could be associated with those relations (for example, the MADE-OF relation in the CARIN model how has no way of requiring that a concept taking part in it is type of substance). Furthermore, the four different relations we used in our experiment do not occur in the CARIN model's fixed set of thematic relations: it would be hard for that model to explain how people used these relations in our experiment.

Our findings are consistent with other theories of conceptual combination (e.g. Costello and Keane, 2000; Murphy, 1988; Wisniewski, 1997) which do allow internal conceptual structure to influence relation selection. These theories use some variation of the idea that a concept representation can contain 'slots' such as MADE-OF or LOCATED and that conceptual combination involves one concept filling a slot in another concept (so that "kitchen chair", for example, would involve the "kitchen" concept filling the LOCATED slot in "chair"). The exemplar-based model of relation selection described in this paper provides an alternative to this slot-based representation of relations, showing that relations can be represented as sets of paired-item exemplars, rather than as slots in concepts. This exemplar-based model has the advantage of giving a simple account of how people learn which facilitating and diagnostic properties are associated with each relation.

As for future work: in our experiment, participants learned relational categories only, and did not learn conceptual categories (they did not learn different categories of beetle or plant, for example). A possible extension of this work would be to have participants learn, from sets of exemplars, both conceptual categories and the relational categories that link them. This experiment could reveal more both about how relations are learned and used, and about how exemplar-level and conceptual-level information interact in conceptual combination.

Acknowledgements

This research was supported by a grant from the Irish Research Council for Science, Engineering and Technology, funded by the National Development Plan.

References

Costello, F. J. (2000). An exemplar model of classification in simple and combined categories. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, (pp 95-100). Mahwah, N. J.: Erlbaum.

Costello, F. J. (2001) A computational model of categorisation and category combination: Identifying diseases and new disease combinations, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, (pp. 238-243). University of Edinburgh: Erlbaum.

Costello, F., & Keane, M.T. (2000). Efficient Creativity: Constraints on conceptual combination. *Cognitive Science, 24,* 299-349.

Costello, F. J., and Keane, M. T. (2001) Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts, *Journal of Experimental Psychology: Learning, Memory & Cognition, 27,* 255-271.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23* (1), 71-87.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review,* 85 (3), 207-238.

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science*, 12, 529-562.

Nosofsky, R. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* (1), 104 114.

Wisniewski, E. J. (1996). Construal and similarity in conceptual combination. *Journal of Memory and Language,* 35, 434-453.

Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review,* 4(2), 167-183.

Simple Ways to Construct Search Orders

Anja Dieckmann (dieckmann@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Peter M. Todd (ptodd@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Abstract

Simple decision heuristics that process cues in a particular order and stop considering cues as soon as a decision can be made have been shown to be both accurate and quick. But one criticism of heuristics such as Take The Best is that these owe much of their simplicity and success to the not inconsiderable computations necessary for setting up the cue search order before the heuristic can be used. The criticism, though, can be countered in two ways: First, there are typically many cue orders possible that will achieve good performance in a given problem domain. And second, as we will show here, there are simple learning rules that can quickly converge on one of these useful cue orders through exposure to just a small number of decisions. We conclude by arguing for the need to take into account the computation necessary for not only the application but also the setup of a heuristic when talking about its simplicity.

One-Reason Decision Making and Ordered Search

In the book *Simple heuristics that make us smart*, Gigerenzer and colleagues (1999) propose several decision making heuristics for predicting which of two objects or options, described by multiple binary cues, scores higher on some quantitative criterion. These heuristics have in common that information search is stopped once one cue is found that discriminates between the alternatives and thus allows an informed decision. No integration of information is involved, leading these heuristics to be termed "one-reason" decision mechanisms. These heuristics differ only in the search rule that determines the order in which information is searched. But where do these search orders come from?

"Take the Best" (TTB; Gigerenzer & Goldstein, 1996, 1999) is the heuristic that has received most attention to date, both theoretically and empirically. TTB consists of three building blocks:

1. Search rule: Search through cues in the order of their validity. Validity is the proportion of correct decisions made by a cue out of all the times that cue discriminates between pairs of options.

2. Stopping rule: Stop search as soon as one cue is found that discriminates between the two options.

3. Decision rule: Select the option to which the discriminating cue points, that is, the option that has the cue value associated with higher criterion values.

The performance of TTB has been tested on several real-world data sets, ranging from professors' salaries to fish fertility (Czerlinski, Gigerenzer & Goldstein, 1999). Cross-validation comparisons have been made against other more complex strategies, such as multiple linear regression, by training on half of the items in each data set to get estimates of the relevant parameters (e.g., cue order based on validities for TTB, beta-weights for multiple linear regression) and testing on the other half of the data. Despite only using on average a third of the information employed by multiple linear regression, TTB outperformed regression in accuracy when generalizing to the test set (71% vs. 68%).

The even simpler heuristic Minimalist was tested in the same way. It is another one-reason decision making heuristic that differs from TTB only in its search rule. Minimalist searches through cues randomly, and thus requires even less knowledge and precomputation than TTB – all it needs to know are the directions in which the cues point. Again it was surprising that this heuristic performed reasonably close to multiple regression (65%). But the fact that Minimalist lagged behind TTB by a noticeable margin of 6 percentage points indicates that part of the secret of TTB's success lies in its ordered search.

In this paper, we explore how such useful cue orders can be constructed in the first place, by testing a variety of simple order-learning rules in simulation. We find that simple mechanisms at the learning stage can enable simple mechanisms at the decision stage, such as one-reason decision heuristics, to perform well.

Experimental Evidence for Ordered Search

From an adaptive point of view, the combination of simplicity and accuracy makes one-reason decision making with ordered search, as in TTB, a plausible candidate for human decision processes. Consequently, TTB has been subjected to several empirical tests. Because TTB explicitly specifies information search as one aspect of decision making, it must be tested in situations in which cue information is not laid out all at once, but has to be searched for one cue at a time, either in the external environment or in memory (Gigerenzer & Todd, 1999).

In situations where information must be searched for sequentially in the external environment, particularly when there are direct search costs for accessing each successive cue, considerable use of TTB has been demonstrated (Bröder, 2000, experiments 3 & 4; Bröder, 2003). This also holds for indirect costs, such as from time pressure (Rieskamp & Hoffrage, 1999), as well as for internal search in memory (Bröder & Schiffer, 2003). The particular search order used has not always been tested separately, but when such an analysis at the level of building blocks has been done, search by cue validity order has received support (Newell & Shanks, 2003; Newell, Weston & Shanks, 2003).

However, none of these experiments tested search rules other than validity ordering. One other very important dimension on which cues can be ordered is discrimination rate, which refers to the proportion of all possible decision pairs in which a cue has different values for (i.e., discriminates between) the two alternatives[1]. A closer look into the experimental designs of the studies cited above reveals that they all used systematically constructed environments in which discrimination rates of the cues were held constant. Now, when the discrimination rates of cues are all the same, there are not many orders besides validity that make sense. To put it differently, identical discrimination rates make several alternative ordering criteria that combine discrimination rate and validity (e.g., Martignon & Hoffrage, 2002) all lead to the same (validity) order. Examples for such criteria are *success*, which is the proportion of correct discriminations that a cue makes plus the proportion of correct decisions expected from guessing in the non-discriminating trials (success = $v \cdot d + 0.5 \cdot (1-d)$, where v = validity and d = discrimination rate of the cue), and *usefulness,* the portion of correct decisions not including guessing (usefulness = $v \cdot d$).

Because these criteria collapse to a single order (validity) in the reported experiments, nothing can be said about how validity and discrimination rate may interact to determine the search orders that participants apply. It remains unclear what information participants would base their decisions on when both validity and discrimination rate vary. There are hints that when information is costly, making it sensible to consider both how often a cue will enable a decision (i.e., its discrimination rate) and the validity of those decisions, other criteria such as success that combine the two measures show a better fit to empirical data (e.g., Newell, Rakow, Weston & Shanks, in press; Läge, Hausmann, Christen & Daub, submitted). But these studies, too, remain silent about how these criteria, or an order based on these criteria, could possibly be derived by participants.

In sum, despite accumulating evidence for the use of one-reason decision making heuristics, the basic processes that underlie people's search through information when employing such heuristics remain a mystery. While some clues can be had by considering the size of the overlap or correlations between the search orders people use and various standard search orders (as reported by Newell et al., in press, and Läge et al., submitted), they do not come close to telling us how cue orders could possibly be learned.

Search Order Construction – the Hard Way

But how can the search order of TTB be constructed? Although TTB is a very simple heuristic to apply, the set-up of its search rule requires knowledge of the ecological validities of cues. This knowledge is probably not usually available in an explicit precomputed form in the environment, and so must be computed from stored or ongoing experience. Gigerenzer at al. (1999) have been relatively silent about the process by which people might derive validities and other search orders, a shortfall several peers have commented on (e.g., Lipshitz, 2000; Wallin & Gärdenfors, 2000). The criticism that TTB owes much of its strength to rather comprehensive computations necessary for deriving the search order cannot easily be dismissed. Juslin and Persson (2002) pay special attention to the question of how simple and informationally frugal TTB actually is, debating how to take into account the computation of cue validities for deriving the search order. They differentiate two main possibilities on the basis of when cue validities are computed: precomputation during experience, and calculation from memory when needed.

When potential decision criteria are already known at the time objects are encountered in the environment, then relevant validities can be continuously computed and updated with each new object seen. But if it is difficult to predict what decision tasks may arise in the future, this pre-computation of cue validities runs into problems. In that case, at the time of object exposure, all attributes should be treated the same, because any one could later be either a criterion or a cue depending on the decision being made. To use the well-known domain of German cities (Gigerenzer & Goldstein, 1996, 1999), the task that one encounters need not be the usual prediction of city populations based on cues such as train connections, but could just as well be which of two cities has an intercity train line based on cues that include city population. To keep track of all possible validities indicating how accurately one attribute can decide about another, the number of precomputed validities would have to be $C^2 - C$, with C denoting to the number of attributes available. In the German cities example, there are 10 attributes (9 cues plus the criterion population size), thus 90 validities would have to be pre-computed. This number rises rapidly with increasing number of attributes. Even ignoring computational complexity, this precomputation approach is not frugal in terms of information storage.

As a second possibility, Juslin and Persson (2002) consider storing all objects (exemplars) encountered along with their attribute values and postponing computation of validities to the point in time when an actual judgment is required. This, however, makes TTB considerably less frugal during its application. The number of pieces of information that would have to be accessed at the time of judgment is the number of attributes times the number of stored objects; in our city example, it is 10 times the number of known objects. With regard to computing validities for each of the $N \cdot (N-1)/2$ possible pairs that can be formed between the N known objects, each of the C cues has to be

[1] Other dimensions for ordering are possible, such as the temporal order of previous cue use, but we will not consider them here.

checked to see if it discriminated, and did so correctly. Thus the number of checks to be performed before a decision can be made is C·N·(N-1)/2, which grows with the square of the number of objects.

Although Juslin and Persson assume worst case scenarios in terms of computational complexity for the sake of their argument, they raise an important point, showing that one of the fundamental questions within the framework of the ABC research group (Gigerenzer et al., 1999) remains open: How can search orders be derived in relatively simple ways?

Many Roads Lead (Close) to Rome

From what we have said so far, the situation does not look too good for validity either in terms of empirical evidence or psychological feasibility. But what would be the consequence in terms of loss in accuracy if we drop the assumption that cue search follows the validity order? Simulation results can provide an answer. First of all, validity is usually not the best cue ordering that can be achieved. For the German city data set, Martignon and Hoffrage (2002) computed the performance of all possible orderings, assuming one-reason stopping and decision building blocks. The number of possible orders was 362,880 (9! orders of 9 cues). The mean accuracy of the resulting distribution corresponded to the performance expected from Minimalist, 70%, which was considerable above the worst ordering at 62%. Ordering cues by validity led to an accuracy of 74.2%, while the optimal ordering yielded 75.8% accuracy. More than half of all possible cue orders do better than the random order used by Minimalist, and 6,532 (1.8%) do better than the validity order. We can therefore conclude that many good orders exist. But how can one of these many reasonably good cue orders be constructed in a psychologically plausible way?

Search Order Construction – the Simple Way

A variety of simple approaches to deriving and continuously updating search orders can be proposed. Indeed, computer scientists have explored a number of self-organizing sequential search heuristics for the purpose of speeding retrieval of items from a sequential list when the relative importance of the items is not known a priori (Rivest, 1976; Bentley & McGeoch, 1985). The mechanisms they have focused on use transposition of nearby items and counting of instances of retrieval. Our problem of cue ordering is slightly different from that of the standard sequential list ordering, because cues can fail in ways that retrieved items cannot: a cue may not discriminate (necessitating the search for another cue before a decision can be made), or it may lead to a wrong decision. Still, the mechanisms of transposition and counting will be central to the heuristics we propose.

We focus on search order construction processes that are psychologically plausible by being frugal both in terms of information storage and in terms of computation. The decision situation we explore is different from the one assumed by Juslin and Persson (2002) who strongly differentiate between learning of (or about) objects and making decisions. Instead of assuming this unnecessary separation, we will explore a learning-while-doing situation. Certainly there are many occasions akin to Juslin and Persson's situation where individuals have to make decisions based on knowledge they have learned about objects encountered previously and in a different task context. But perhaps more common are tasks that have to be done repeatedly with feedback being obtained after each trial about the adequacy of one's decision. For instance, we can observe on multiple occasions which of two supermarket checkout lines, the one we have chosen or (more likely) another one, is faster, and associate this outcome with cues including the lines' lengths and the ages of their respective cashiers. In such situations, one can learn about the differential usefulness of cues for solving the task via the feedback received over time. It is this case – decisions made repeatedly with the same cues and criterion and the opportunity to learn from outcome feedback – which we will now look at more closely.

We consider several explicitly defined cue order learning rules that are designed to deal with probabilistic inference tasks. In particular, the task we use is forced choice paired comparison, in which a decision maker has to infer which of two objects, each described by a set of binary cues, is "bigger" on a criterion – the task for which TTB was formulated. Thus, in contrast to Juslin and Persson (2002), we assume individuals encounter decision situations instead of objects. After an inference has been made, feedback is given about whether a decision was right or wrong. Therefore, the learning algorithm has information about which cues were looked up, whether a cue discriminated, and whether a discriminating cue led to the right or wrong decision. There are different possibilities for taking these pieces of information into account. For example, correct decisions could be counted up for each cue (essentially keeping tallies). Or the information could be used to compute cue validities and discrimination rates based on the cases in which the cue has actually been looked up so far. These tallies, validity estimates, etc., would then be used for creating and adjusting the current cue order.

The rules we propose differ in the pieces of information they use and how they use them. We classify the learning rules based on their memory requirement – high versus low – and their computational requirements (see Table 1). The computational requirements include whether the entire set of cues is completely reordered after each decision or only adjusted locally via swapping of neighboring cue positions, and whether reordering is done on the basis of measures involving division, such as validity, or simple tallying.

The *validity rule* is the most demanding of the rules we consider in terms of both memory requirements and computational complexity. It keeps a count of all discriminations made by a cue so far (in all the times that the cue was looked up) and a separate count of all the correct discriminations. Therefore, memory load is comparatively high. The validity of each cue is determined by dividing its current correct discrimination count by its

Table 1: Learning rules classified according to memory and computational requirements

High memory load, complete reordering	High memory load, local reordering	Low memory load, local reordering
Validity: reorders cues based on their current validity Tally: reorders cues by number of correct minus incorrect decisions made so far	Tally swap: moves cue up (down) one position if it has made a correct (incorrect) decision if its tally of correct minus incorrect decisions is ≥ (≤) that of next higher (lower) cue	Simple swap: moves cue up one position if it has made a correct decision, and down if it has made an incorrect decision
Variants: - *reorder based on tally of discriminations so far* - *reorder based on tally of correct decisions only*	*Variants:* - *only upward swapping after correct decisions* - *tally of correct decisions only*	*Variants:* - *moving cues more than one position* - *only upward swapping after correct decisions*

total discrimination count. Based on these values computed after each decision, the rule reorders the whole set of cues from highest to lowest validity.

The *tally rule* only keeps one count per cue, storing the number of correct decisions made by that cue so far minus the number of incorrect decisions. So if a cue discriminates correctly on a given trial, one point is added to its tally. If it leads to an incorrect decision, one point is subtracted from its tally. The tally rule is less demanding both in terms of memory and computation: Only one count is kept, and no division is required.

While the validity and tally rules rely on a counting mechanism, the *simple swap rule* uses the principle of transposition (cf. Bentley & McGeoch, 1985). This rule has no memory of cue performance other than an ordered list of all cues, and just moves a cue up one position in this list whenever it leads to a correct decision, and down if it leads to an incorrect decision. In other words, a correctly deciding cue swaps positions with its nearest neighbor upwards in the cue order, and an incorrectly deciding cue swaps positions with its nearest neighbor downwards.

The *tally swap rule* is a hybrid of the simple swap rule and the tally rule. It keeps a tally of correct minus incorrect discriminations per cue so far (so memory load is high) but only locally swaps cues: When a cue makes a correct decision and its tally is greater than or equal to that of its upward neighbor, the two cues swap positions. When a cue makes an incorrect decision and its tally is smaller than or equal to that of its downward neighbor, the two cues also swap positions.

As indicated in table 1, many variants of these basic types of learning rules are possible. Here we will focus on these four rules spanning the space of possibilities, and look at how they perform in simulations. Elsewhere we consider evidence for their use in experimental decision settings, and use these simulation results to assess human performance.

Simulation Study of Simple Ordering Rules

To test the performance of these order learning rules, we use the German cities data set (Gigerenzer & Goldstein, 1996), consisting of the 83 highest-population German cities described on 9 cues. The question we want to address is, what would happen if a decision-maker does not search for cues in validity order from the beginning, but instead must construct a search order using feedback received about each decision made? We assume that cue directions are known. Furthermore, instead of allowing the decision maker to look up information about all 9 cues in each pair comparison, we assume that TTB's stopping and decision rule are used on all decisions. We do this because it is more natural to assume that learning happens in the ongoing context of decision making, which does not necessarily involve exhaustive information search. This runs counter the approach taken by Juslin and Persson (2002) who in their worst case scenarios assume exhaustive information search for validity computations. In our approach, only the limited information gathered until the first discriminating cue is found can be taken into account.

We simulated 10,000 learning trials for each rule, starting from random initial cue orders. Each trial consisted of 100 decisions between randomly selected decision pairs. Below we report average values across the 10,000 trials.

Results

We start by considering the cumulative accuracies (i.e., online or amortized performance – Bentley & McGeoch, 1985) of the rules, defined as the total percentage of correct decisions made so far at any point in the learning process. (The contrasting measure of offline accuracy – how well the current learned cue order would do if it were applied to the entire test set – is a less psychologically useful indication of a real decision maker's performance using some rule.) The mean cumulative accuracies of the different search order learning rules when used with one-reason decision making are shown in Figure 1. Cumulative accuracies soon rise above that of the Minimalist heuristic (proportion correct = 0.70) which looks up cues in random order and thus serves as a lower benchmark. However, at least throughout the first 100 decisions, cumulative accuracies stay well below the (offline) accuracy that would be achieved by using TTB for

all decisions (proportion correct = 0.74), looking up cues in the true order of their ecological validities.

The four learning rules all perform on a surprisingly similar level, with less than one percentage point difference in favor of the most demanding rule (i.e., validity) compared to the least (i.e., simple swap; mean proportion correct in 100 decisions: validity learning rule: 0.719; tally: 0.716; tally swap: 0.715; simple swap: 0.711). Importantly, though, the more demanding learning rules outperform Minimalist earlier. Whereas the tally swap and simple swap rule lead to accuracies that are significantly higher than Minimalist only after 48 and 61 decisions, respectively, the validity learning rule does significantly better already after 37 decisions, and the tally rule after 35 decisions ($z = 1.65$, $p = 0.05$).

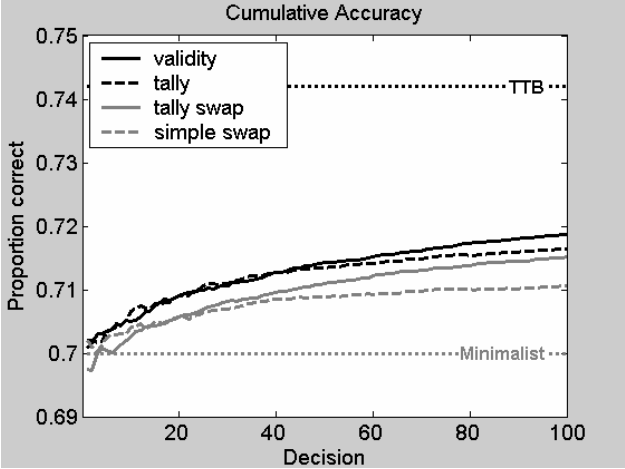


Figure 1: Mean cumulative accuracy of order learning rules

These four learning rules are, however, all more *frugal* than TTB, and even more frugal than Minimalist. On average, they look up fewer cues before reaching a decision (see Figure 2). Again, there is little difference between the rules (mean number of cues looked up in 100 decisions: validity learning rule: 3.17; tally: 3.07; tally swap: 3.13; simple swap: 3.18). The validity learning rule and the tally rule lead to cue orders that are significantly more frugal than Minimalist very early (after 16 and 14 decisions, respectively), whereas the two swapping rules take longer: The tally swap rule takes 27 decisions, and the simple swap rule 32 decisions.

Consistent with this finding, all of the learning rules lead to cue orders that show positive correlations with the discrimination rate cue order (reaching the following values after 100 decisions: validity learning rule: $r = 0.18$; tally: $r = 0.29$; tally swap: $r = 0.24$; simple swap: $r = 0.18$). This means that cues that often lead to discriminations are more likely to end up in the first positions of the order. In contrast, the cue orders resulting from all learning rules but the validity learning rule do not correlate with the validity cue order, and even the correlations of the cue orders resulting from the validity learning rule after 100 decisions only reach an average $r = 0.12$.

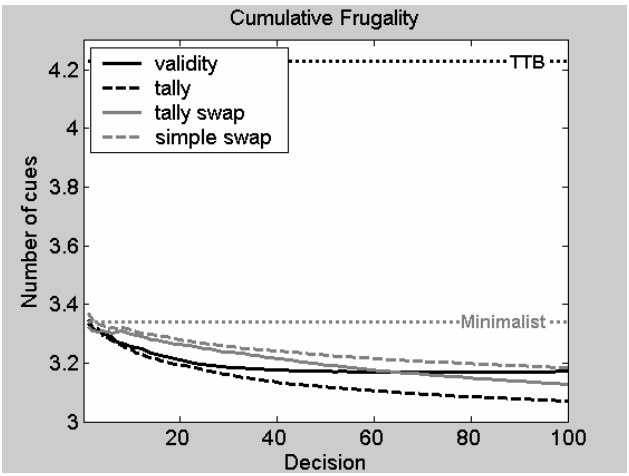


Figure 2: Mean cumulative frugality of order learning rules

But why would the discrimination rates of cues exert more of a pull on cue order than validity, even when the validity learning rule is applied? Part of the explanation comes from the fact that in the city data set we used for the simulations, validity and discrimination rate of cues are negatively correlated. Having a low discrimination rate means that a cue has little chance to be used and hence to demonstrate its high validity. Whatever learning rule is used, if such a cue is displaced downward to the lower end of the order by other cues, it may never be able to escape to the higher ranks where it belongs. The problem is that when a decision pair is finally encountered for which that cue would lead to a correct decision, it is unlikely to be checked because other, more discriminating although less valid, cues are looked up before and already bring about a decision. Thus, because one-reason decision making is intertwined with the learning mechanism and so influences which cues can be learned about, what mainly makes a cue come early in the order is producing a high *number* of correct decisions and not so much a high *ratio* of correct discriminations to total discriminations regardless of base rates.

In sum, all of the learning rules lead to accuracies between that of the heuristics TTB and Minimalist, but some rules reach orders that are better than Minimalist sooner. The rules are highly frugal, with a (slight) tendency to change the order in the direction of discrimination rate.

Discussion

The simpler cue order learning rules we have proposed do not fall far behind a validity learning rule in accuracy. This holds even for the simplest rule, which only requires memory of the last cue order used and moves a cue one position up in that order if it made a correct decision, and down if it made an incorrect decision. All of the rules considered here make one-reason decision heuristics perform above the level of Minimalist in the long run.

On the other hand, the four rules, even the validity learning rule, stay below TTB's accuracy across a relatively high number of decisions. But often it is necessary to make good decisions without much experience. Therefore, learning rules should be preferred that quickly lead to orders

with good performance. Both the validity and tally learning rules quickly beat Minimalist. At the same time, the tally rule leads to considerably more frugal cue orders.

Remember that the tally rule assumes full memory of all correct minus incorrect decisions made by a cue so far. But this does not make the rule implausible. There is considerable evidence that people are actually very good at remembering the frequencies of events. Hasher and Zacks (1984) conclude from a wide range of studies that frequencies are encoded in an automatic way, implying that people are sensitive to this information without intention or special effort. Estes (1976) pointed out the role frequencies play in decision making as a shortcut for probabilities. Further, the tally rule is comparatively simple, not having to keep track of base rates or perform divisions as does the validity rule. From the other side, the simple swap rule may not be much simpler, because storing a cue order may be about as demanding as storing a set of tallies. We therefore conclude that the tally rule should not be discounted on grounds of implausibility without further empirical evidence. Of course, a necessary next step (currently underway) will be to test how well these and other rules predict people's information search when they have to make cue-based inferences without knowing validities.

Our goal in this paper was to argue for the necessity of taking into account the set-up costs of a heuristic in addition to its application costs when considering the mechanism's overall simplicity. As we have seen from the example of the validity search order of TTB, what is easy to apply may not necessarily be so easy to set up. But simple rules can also be at work in the construction of a heuristic's building blocks. We have proposed such rules for the construction of one building block, the search order. We have seen that these simple learning rules enable a one-reason decision heuristic to perform only slightly worse than if it had full knowledge of cue validities from the very beginning. Giving up the assumption of full a priori knowledge for the slight decrease in accuracy seems like a reasonable bargain: Through the addition of learning rules, one-reason decision heuristics might lose some of their appeal to decision theorists who were surprised by the performance of such simple mechanisms compared to more complex algorithms, but they gain psychological plausibility and thus become more attractive as explanations for human decision behavior.

References

Bentley, J.L. & McGeoch, C.C. (1985). Amortized analyses of self-organizing sequential search heuristics. *Communications of the ACM, 28(4),* 404-411.

Bröder, A. (2000). Assessing the empirical validity of the "Take-The-Best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26 (5)*, 1332-1346.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 611-625.

Bröder, A. & Schiffer, S. (2003). "Take The Best" versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General, 132,* 277-293.

Czerlinski, J., Gigerenzer, G., & Goldstein, D.G. (1999). How good are simple heuristics? In G. Gigerenzer, P.M. Todd & The ABC Research Group, *Simple heuristics that make us smart*. New York: Oxford University Press.

Estes, W.K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37-64.

Gigerenzer, G., & Goldstein, D.G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103 (4)*, 650-669.

Gigerenzer, G., & Goldstein, D.G. (1999). Betting on one good reason: The Take The Best Heuristic. In G. Gigerenzer, P.M. Todd & The ABC Research Group, *Simple heuristics that make us smart*. New York: Oxford University Press.

Gigerenzer, G., & Todd, P.M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P.M. Todd & The ABC Research Group, *Simple heuristics that make us smart*. New York: Oxford University Press.

Gigerenzer, G., Todd, P.M., & The ABC Research Group (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Hasher , L., & Zacks, R.T. (1984). Automatic Processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*, 1372-1388.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): a "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563-607.

Läge, D., Hausmann, D., Christen, S. & Daub, S. (submitted). Take The Best: How much do people pay for validity?

Lipshitz, R. (2000). Two cheers for bounded rationality. *Behavioral and Brain Sciences, 23*, 756-757.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision, 52*, 29-71.

Newell, B.R., Rakow, T., Weston, N.J., & Shanks, D.R. (in press). Search strategies in decision-making: the success of 'success'. *Journal of Behavioral Decision Making.*

Newell, B.R., & Shanks, D.R. (2003). Take the best or look at the rest? Factors influencing 'one-reason' decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 53-65.

Newell, B.R., Weston, N.J., & Shanks, D.R. (2003). Empirical tests of a fast and frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes, 91*, 82-96.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer, P.M. Todd & The ABC Research Group, *Simple heuristics that make us smart*. New York: Oxford University Press.

Rivest, R. (1976). On self-organizing sequential search heuristics. *Communications of the ACM, 19(2),* 63-67.

Wallin, A. & Gärdenfors, P. (2000). Smart people who make simple heuristics work. *Behavioral and Brain Sciences, 23*, 765.

Influencing nonmonotonic reasoning by modifier strength manipulation

Kristien Dieussaert (kristien.dieussaert@psy.kuleuven.ac.be)
Department of Psychology, University of Leuven,
102 Tiensestraat, B-3000 Leuven, Belgium

Marilyn Ford (m.ford@griffith.edu.au)
School of Computing and Information Technology, Griffith University,
Nathan, Queensland, Australia, 4111

Leon Horsten (leon.horsten@hiw.kuleuven.ac.be)
Department of Philosophy, University of Leuven,
2 Kardinaal Mercierplein, B-3000 Leuven, Belgium

Abstract

Despite the current belief that much common sense reasoning is nonmonotonic in nature, research indicates that only a limited percentage of people are good at nonmonotonic reasoning. Good nonmonotonic reasoners recognize the logical strengths and weaknesses of some arguments. In the present study, we focus on differences in the probabilistic interpretation of the modifiers *typically* and *usually* and on the resulting differences in the strengths and weaknesses of arguments. We show that these implicit probabilistic strengths influence the reasoning process of good nonmonotonic reasoners.

Introduction

Most AI logicians and logic programmers, as well as philosophical logicians, ground their interest in nonmonotonic reasoning on the observation that common sense reasoning is largely nonmonotonic in nature. Ginsberg (1994, p.2), for example, states that: "*... flexibility is intimately connected with the defeasible nature of commonsense inference ... we are all capable of drawing conclusions, acting on them, and then retracting them if necessary in the face of new evidence. If our computer programs are to act intelligently, they will need to be similarly flexible*".

Pelletier and Elio (1997) argue that researchers like Ginsberg are right in grounding their research on human reasoning, but they also argue that AI researchers should bear the consequences of it.

As a first consequence, Pelletier and Elio (1997) argue for a more systematic study of human inferences. They plead against the use of the intuitions of a small group of AI researchers to decide on the acceptable answers to some 'benchmark' problems (e.g. Lifschitz, 1988). Within the AI community, at least some researchers are also supportive of this point of view (e.g. Schurz, 2001; Benferhat, Bonnefon, & Da Silva Neves, 2002).

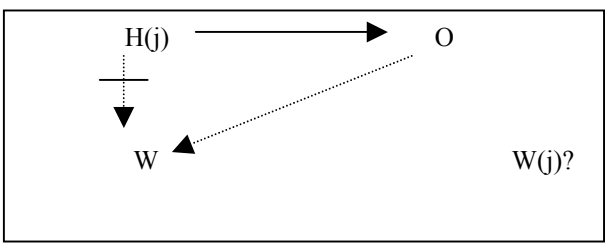
The second consequence of grounding default reasoning research on human common sense reasoning is more severe. Contrary to deductive reasoning, where classical logic is considered to be the norm, there is no standard norm for default reasoning. Since the long-term goal of most AI researchers is to simulate human reasoning, and since we do not have an objective theory of what rational default reasoning is, Pelletier and Elio (1997) argue that a psychologistic view of default reasoning should be adopted. Thus, the data that any default system should cover should be determined by the practices of ordinary people.

We agree with the importance of investigating human reasoning by controlled experimental research to gain more insight into the process of human reasoning. However, we also argue that it is important for AI not to develop a formal nonmonotonic logic or automated reasoning system on the basis of flawed reasoning.

Ford and Billington (2000) took the need for AI researchers to study human reasoning seriously and systematically investigated human nonmonotonic reasoning with abstract material. This way, participants could not rely on background knowledge, as they can do in daily life.

To clarify the discussion, we present an example of one problem, with a well-known Tweety-bird like structure:

Hittas are usually not waffs.	Penguins do not fly
All of the hittas are oxers.	All penguins are birds
Oxers are usually waffs.	Birds fly
Jukk is a hitta.	Tweety is a Penguin
Is Jukk a waff?	Does Tweety fly?



Ford and Billington (2000) summed up their main finding by presenting five negative (N) and three positive (P) factors

that influenced people's reasoning about nonmonotonic problems:
(N1) Most participants were not willing to draw a tentative conclusion when faced with conflict and with non-strict rules. (N2) Some participants weighed up the perceived number of relevant positive and negative paths, though the perceived paths were not paths. (N3) Some participants considered path length regardless of the ordering of rule types. Most participants preferred the shorter path to the longer path. (N4) Some participants gave weight to the presence of the universal quantifier, even when this was inappropriate. (N5) Some participants interpreted 'usually not' as evidence that 'some are' and thus gave preference to a positive conclusion. (See Hewson & Vogel, 1994, and Vogel & Tonhauser, 1996, for more evidence that many people have difficulty with nonmonotonic reasoning problems).
Besides these negative factors, Ford and Billington (2000) also extracted some positive factors from their experiment. (P1) Some participants recognized the relevance of the fact that if *all of the Xs are Ys* there might be *Ys that are not Xs*. (P2) Some participants recognized the relevance of the fact that if *all of the Ys are Zs* then *any Xs that are Ys are also Zs*. (P3) Some participants recognized that given a sentence *Xs are usually Ys* there are potentially many Ys that are not Xs.
People who recognize these positive points are able to see differences in the logical strength of arguments. Consider the following:

X → Y ⇢ Z (a)
(all of the Xs are Ys, Ys are usually Zs)

X ⇢ Y → Z (b)
(Xs are usually Ys, all of the Ys are Zs)

X ⇢ Y ⇢ Z (c)
(Xs are usually Ys, Ys are usually Zs)

An appreciation of P2 allows people to recognize the strength of (b): for (b), it must be the case that Xs are usually Zs because the Xs that are Ys must also be Zs. In contrast, an appreciation of P1 and P3 allows people to see the weakness of (a) and (c), respectively: for (a) and (c), it might be that none of the Xs are Zs because it could be that the Ys that are not Xs are not Zs.
Ford (2004, In Press) argues that people who see the differences in the logical strength of arguments are more likely to give logically justifiable answers on nonmonotonic reasoning problems, since they rely on logically valid principles to form their answers. For example, with problems such as (1), they answer 'unlikely' more frequently and more strongly than people who do not see differences in the logical strength of arguments. They give this answer because of their recognition of the weakness of (a).
Note that these reasoners are not relying on a notion of 'specificity', where information stemming from a subclass overrides information from a superclass; Ford and Billington's (2000) subjects did not articulate this notion of specificity as it is used in AI and the three P factors they identified make no mention of such specificity. The subjects instead rely on the logical strength of conflicting paths in an argument.
In this manuscript, we investigate further the nature of the logical strengths and weaknesses that reasoners who appreciate P2 and P3 use. We will argue for differences in the probabilistic interpretation of the modifiers *usually* and *typically* and consequent differences in the logical strengths and weaknesses of arguments. Given the results of Ford (2004, In Press), we would thus expect variations in conclusions given by reasoners who appreciate P2 and P3, with these reasoners giving more weight to the stronger side of an argument.
In a pilot experiment, we will extend a former study in which it was shown that researchers should be careful how to phrase their 'default relations' (Dieussaert, 2003). Researchers do not seem to make a distinction between sentences such as 'birds fly', 'birds normally fly', 'birds usually fly', 'birds typically fly' and so on. However, Dieussaert showed that the interpretation of these sentences, and the inferences yielded from them, differ greatly.
For the present study, we focus on the difference between *usually* and *typically*. We confirm the finding that *typically* is interpreted as indicating more instantiations of a type than *usually*. This implies that 'birds typically fly' represents a stronger relation than 'birds usually fly' since more instances of 'bird' are supposed to fly in the former case.
In a second experiment, we use this finding to show how the strengths and weaknesses of arguments can influence nonmonotonic reasoning. Reasoners who are shown to appreciate P2 and P3 are given problems with relations phrased with *typically* and *usually*. The data show clearly that the strength of arguments influences the nonmonotonic reasoning process for these subjects.

Pilot Experiment

In an earlier experiment (Dieussaert, 2003), the influence of phrasing on the interpretation of default sentences was shown. In a within subjects design, participants estimated the positive outcome of a sequence such as 'Hilo are typically waff. Jukk is a hilo. Is Jukk a waff?' significantly higher than for sequences like 'Brant are usually glent. Kerdo is a brant. Is Kerdo a glent?'
To obtain confirmatory evidence, we extended the earlier experiment.

Method

Participants

Ninety-nine first year students in Psychology at the University of Leuven, who had not taken a logic course, participated as a partial fulfillment of a course requirement.

Design
The design was completely within subjects. The dependent variable was the percentage entered per item.

Material and Procedure
Each participant received a booklet with written instructions and 18 items in randomized order. Each participant solved 18 problems: 9 positive items and 9 negative items. They solved this paper-and-pencil task individually and in a self-paced manner.
Here, we focus only on the difference between *typically* and *usually*, since these terms form the core of the main experiment.
The relevant material was:

- Nagdals are usually pirasos.
- Hittas are typically waffs
- Nilo are usually not riza
- Koki are typically not liri

The instructions for the positive items were as follows:
On each of the following pages a sentence will be presented. We ask you to mark how you interpret the underlined word within this sentence. To clarify the task, we give you an example.
The sentence: JY members are *normally* singers.
The question: Does the word '*normally*' mean that:
Mark one or more answers:
0 A certain percentage of JY members have features that characterise singers. If so, given the sentence, what would you assume would be the approximate % (0-100) of JY members that have features that characterize singers. .....% [further referred to as: **Feature**]
0 A certain percentage of time JY members are singers. If so, given the sentence, what would you assume would be the approximate % (0-100) of time JY members are singers. .....% [further referred to as: **Time**]
0 A certain percentage of JY members are singers. If so, given the sentence, what would you assume would be the approximate % (0-100) of JY members that are singers. .....% [further referred to as: **Number**]
For the negative items, the task were rephrased in a negative form e.g.: ...% (0-100) of JY members that are <u>not</u> singers.

Results

Table 1: Mean percentages given for Feature, Time, and Number (see Material).[a]

Problem	Feature	Time	Number	Mean
Typically	*91.8*	65.0	92.3	91.8
	(N=85)	(N=01)	(N=24)	[SD=11.7]
Usually	77.9	74.3	*79.6*	78.4
	(N=31)	(N=17)	*(N=70)*	[SD=12.5]
Typically not	*87.5*	88.3	89.1	87.3
	(N=73)	(N=04)	(N=36)	[SD=17.9]
Usually not	77.2	76.8	*76.4*	77.4
	(N=27)	(N=38)	*(N=52)*	[SD=14.3]

[a]The number of responses (N) do not add up to 99 because participants were allowed to mark 1-3 answers. Some participants marked only the answers and did not enter a percentage.

Overall percentages entered for *typically* are higher than percentages entered for *usually* (91.8 vs. 78.4; t (92) =7.82, $p < .00001$). Percentages entered for *typically not* are higher than percentages entered for *usually not* (87.3 vs. 77.4; t (46) =4.8, $p < .00001$).
If we take into account only the single choices of participants (and remove items for which more than one answer was marked): Feature is the preferred category for *typically*, while Number is the preferred category for *usually*. A Sign test shows a higher number of Feature choices for *typically* (73) than for *usually* (17; Sign test, non-ties = 63, $Z = 6.3$, $p < .00001$). The same pattern is found for *typically not* (59) versus *usually not* (15; Sign test, non-ties = 70, $Z = 8.3$, $p < .00001$). A Sign test shows a higher number of Number choices for *usually* (55) than for *typically* (13; Sign test, non-ties = 56, $Z = 5.5$, $p < .00001$). A similar pattern is found for *usually not* (39) versus *typically not* (23; Sign test, non-ties = 33, $Z = 2.4$, $p < .00005$).

Discussion
This experiment confirms the results of Dieussaert (2003): *typically* and *usually* are interpreted somewhat differently. Most importantly for our purposes, having *typically* in a sentence is associated with higher percentages than having *usually*, with the former term thus being considered stronger.

Main Experiment
The pilot experiment provides confirmatory evidence for the stronger relation between two propositions A and B when they are connected in a default sentence with *typically* than with *usually*.
Having established this firmly, we can now investigate how the strength of this relation influences the nonmonotonic reasoning process. In this experiment, reasoners who appreciate Ford and Billington's (2000) P2 and P3 are tested.

Method

Participants
Twenty-seven first year students in Psychology from the University of Leuven, who had not taken a logic course, participated as a partial fulfillment of a course requirement.

Design
The design was completely within subjects. The dependent variable was the score on a seven-point scale.

Material and Procedure
Each participant first received two critical questions (Ford, 2004) to see if they recognized Ford and Billington's (2000) P2 and P3.
They were told that there were no time limits. The questions were:

1) Given the following two statements:
Mary's friends are usually Ann's friends.
All of Ann's friends are Sue's friends.
Could it be the case that none of Mary's friends are Sue's friends? (Yes/No)

Given the following two statements:
Jim's friends are usually Tom's friends.
Tom's friends are usually Fred's friends.
Could it be the case that none of Jim's friends are Fred's friends? (Yes/No)

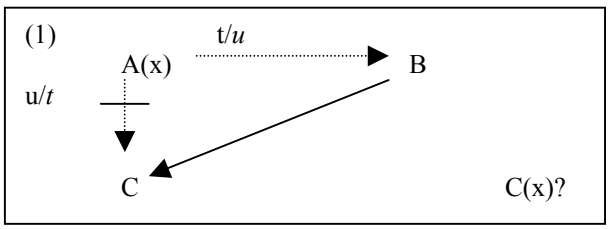
The first question contains the strong argument, with P2 relevant:
X ⋯⋯▶ Y ——▶ Z
while the second question contains the weak argument, with P3 relevant:
X ⋯⋯▶ Y ⋯⋯▶ Z

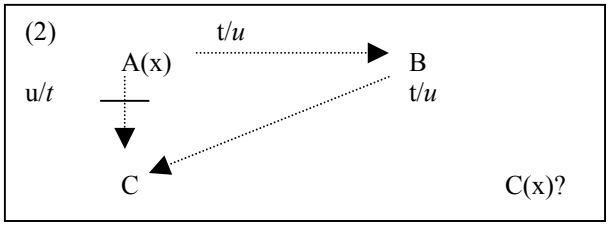
Only participants who answered the critical questions correctly (No on the first, Yes on the second) proceeded to the second part of the experiment, leaving 11 subjects.
These participants received a booklet with written instructions and 18 problems (1 per page). Each participant gave their answer to each problem verbally and then indicated a likelihood estimation on a seven point scale.

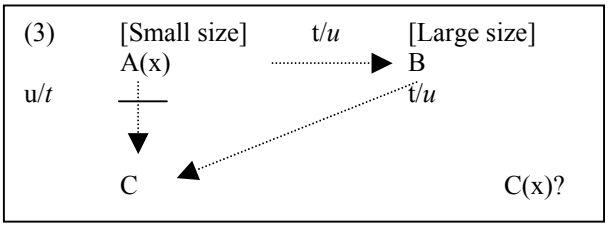
On a scale of 1 to 7, estimate the likelihood of Z(x).

extremely unlikely						extremely likely
1	2	3	4	5	6	7

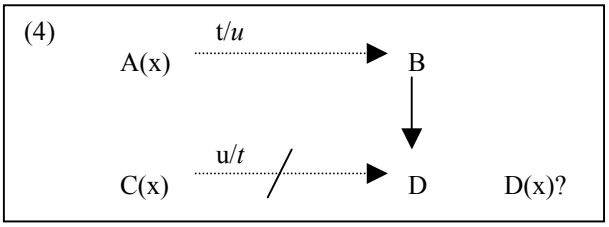
Problems in which the modifier *typically* (t) was used for the positive arguments and in which the modifier *usually* (u) was used for the negative arguments will be referred to as TU problems. Problems in which the modifier *usually* (u) was used for the positive arguments and in which the modifier *typically* (t) was used for the negative arguments will be referred to as UT problems.
Six differently structured problems can be distinguished. Participants received two examples of each of two versions of the following 3-argument structures, thus making 12 problems:

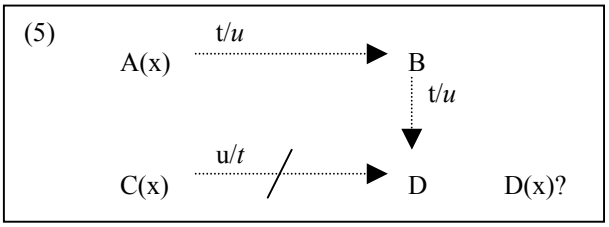


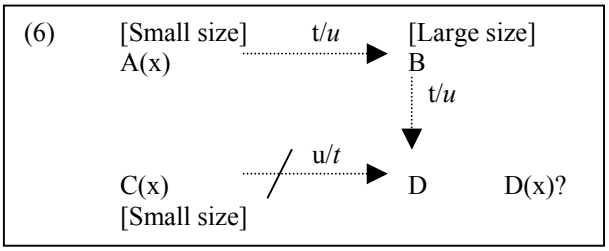




Participants also received one example of each of the two versions of the following 4-argument structures, making a further 6 problems:







It should perhaps be noted here that of the six structures studied here, the notion of specificity, if it were used, could only be applied to Problems 2 and 3, where information from a subclass (A) conflicts with information from its superclass (B). The good reasoners we are using, however, would be expected to use P2 and P3 to compare the strength of the conflicting arguments in all the problems.
Problems usually requiring 'Can't tell'
Problem 1 represents a strong positive versus a strong negative argument. The expected response when the modifier phrase is the same for both sides of the argument is thus around 4, meaning 'can't tell'. However, an additional manipulation was added.

In 1a, the positive non-strict relation was phrased with *typically* (t; As are typically Bs), while the negative non-strict relation was phrased with *usually* (u; As are usually not Cs). For problem 1b, the phrasing was vice versa (*usually* for the positive relation and *typically* for the negative one).
Problem 4 represents a similar problem to (1), but with four propositions involved. The expected response here is also 4, meaning 'can't tell', when the modifiers are the same, but again TU and UT versions were given.
Given that *typically* is stronger than *usually*, the TU versions would be expected to result in a higher rating (more positive) than the UT versions.

Problems usually indicating 'unlikely'
Problem 2 represents a weak positive versus a strong negative argument. The expected response when the modifier phrase is the same for both sides of the argument is thus lower than 4, meaning 'unlikely'. However, the phrasing manipulation could be expected to have an additional influence on the final rating.
Problem 5 is similar to Problem 2, with a weak positive versus a strong negative argument, and with the phrasing manipulation expected to influence the final rating.
Problems 3 and 6 differ from 2 and 5, respectively, in that information is given on the relative subset/superset sizes to which the respective items belong: a small subset for A and a large one for B. It seems (Ford, In Press) that relative size information can sometimes help good reasoners in their reasoning.
For problems 2, 3, 5 and 6, the ratings for the TU versions would be expected to move higher, becoming more positive than would otherwise be expected. With the UT versions, the ratings would be expected to move lower, becoming even less positive than would otherwise be expected.

Results

Table 2: The mean likelihood ratings as a function of modifiers used in the positive and negative arguments. Standard deviations are given in square brackets.

(N = 11)		Problem	TU	UT
3-arg		(1)	4.3 [1.2]	3.7 [0.8]
	w/o relative size	(2)	3.9 [1.1]	2.7 [0.8]
	with relative size	(3)	4.2 [1.1]	2.5 [0.9]
4-arg		(4)	5.0 [1.7]	3.4 [1.2]
	w/o relative size	(5)	4.2 [1.7]	2.5 [0.8]
	with relative size	(6)	4.2 [1.3]	3.4 [1.4]

3-argument problems.
The mean likelihood rating was higher for TU problems than for UT problems (4.1 vs. 3.0; $F(1,10) = 17.5$, $MSE = 1.2$, $p < .005$). Planned comparisons showed that TU ratings are higher than UT ratings for Problem 1 (4.3 vs. 3.7; $F(1,10) = 7.7$, $MSE = .2$, $p < .05$), for Problem 2 (3.9 vs. 2.7; $F(1,10) = 10.1$, $MSE = .7$, $p < .01$), and for Problem 3 (4.2 vs. 2.5; $F(1,10) = 14.4$, $MSE = 1.1$, $p < .005$).
No difference between the problems was observed ($p = .08$). However, an interaction between problem and modifier was observed ($F(2,20) = 4.4$, $MSE = .4$, $p < .05$). A planned comparison shows only a significant difference for UT between Problem 1 and 2 (3.7 vs. 2.7; $F(1,10) = 6.9$, $MSE = .8$, $p < .05$) and between Problem 1 and 3 (3.7 vs. 2.5; $F(1,10) = 31.0$, $MSE = .3$, $p < .0005$). This difference is due to the particularly low ratings of UT Problem 2 and 3.
Notice, too, that relative size information did not influence the ratings.

4-argument problems.
A similar pattern is found for 4-argument problems. The mean likelihood rating was higher for TU problems than for UT problems (4.5 vs. 3.1; $F(1,10) = 7.0$, $MSE = 4.5$, $p < .05$). Planned comparisons showed that TU ratings are higher than UT ratings for Problem 4 (5.0 vs. 3.4; $F(1,10) = 5.2$, $MSE = 2.8$, $p < .05$), for Problem 5 (4.2 vs. 2.5; $F(1,10) = 6.8$, $MSE = 2.4$, $p < .05$), but not for Problem 6 (4.2 vs. 3.4; $p = .2$).
No difference between the problems was observed ($p = 0.5$). No interaction between problem and modifier was observed.

Discussion

This study was set up to gain more insight into the role that modifiers of non-strict relations play in the nonmonotonic reasoning process. Generally, researchers do not pay much attention to the specific wording of default expressions. We showed in a pilot experiment that this neglect is undeserved: default expressions vary in interpretation. However, only if this interpretation also affects the nonmonotonic reasoning process does the topic become particularly noteworthy for AI researchers and philosophical logicians doing research on nonmonotonic reasoning.
In the main experiment we showed that the use of different modifiers in non-strict relations does indeed lead to a variation in nonmonotonic reasoning, more precisely in likelihood ratings on nonmonotonic reasoning problems. We presented reasoners who appreciate P2 and P3, with problems of two kinds: problems with structures where we would normally expect them to give a 'can't tell' answer and problems where we normally expect them to give an 'unlikely' answer. So, if participants bore only the structure of the problem in mind, we would expect a 'can't tell' answer for Problems 1 and 4, and an 'unlikely' answer for Problem 2-3 and 5-6, despite the modifier manipulation. However, if the reasoning process was influenced by the modifier used, we would see a shift in answers, depending on the specific modifier used to express the positive and negative non-strict relations.
The data show clearly that reasoners who appreciate P2 and P3 are influenced by the modifier used. Ratings on TU problems differ significantly from ratings on UT problems. With the 'can't tell' Problem 1, we observed ratings staying close to the 'can't tell' rating, although a positive shift was noted for TU problems, while a slightly negative shift was noted for UT problems, resulting in an overall difference.

With the 'can't tell' Problem 4, the pattern was more extremely pronounced, with a large positive shift for TU problems and a large negative shift for UT problems.

The TU versions of problems 2-3 and 5-6 are lifted up to a 'can't tell' level, while UT versions receive an 'unlikely' rating. While adding relative size information has been shown to sometimes help good nonmonotonic reasoners (Ford, In Press), it did not affect the reasoning process in our experiment, possibly because these subjects did not need this help.

It is clear that although the matching TU and UT versions of problems have the same structure, they are not considered as being equivalent. The modifier *typically* or *typically not* makes a non-strict relation stronger compared with its counterpart *usually* or *usually not.*

It is clear that people who show an appreciation of P2 and P3 and who solve nonmonotonic problems by comparing the logical strength of conflicting arguments, rather than by using a notion of specificity, also use the strength of modifiers to guide their reasoning. Just as it is rational to take the logical strength of conflicting arguments into account, rather than using a notion of specificity, so too it is rational to take into account modifier strength in conflicting arguments.

Conclusion

Ford (2004, In Press) has shown that good reasoners use the logical strength of different sides of an argument to guide their reasoning. The present study adds credence to this effect of weighing up the strengths of the different sides of an argument. The study shows that different modifiers can differentially weaken or strengthen an argument and that they thereby influence reasoning.

Acknowledgments

This research was made possible by the financial support of the Fund for Scientific Research Flanders (project G.0239.02: K. Dieussaert and L. Horsten).

References

Benferhat, S., Bonnefon, J. F., & Da Silva Neves, R. M. (2002, July). *An overview of possibilistic handling of default reasoning: Applications and empirical studies.* Paper presented at the 1st Salzburg Workshop on Paradigms of Cognition (SWPC 1/2002), 'Nonmonotonic and uncertain reasoning in the focus of competing paradigms of cognition', Salzburg, Austria.

Dieussaert, K. (2003). Do typical birds usually fly normally?. In A. Markman & L. Barsalou (Eds.). *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society, Inc.

Ford, M. (2004). System LS: A three-tiered nonmonotonic reasoning system. *Computational Intelligence, 20 (1),* 89-108.

Ford, M. (In Press). Human nonmonotonic rasoning: The importance of seeing the logical strength of arguments. *Synthese.*

Ford, M., & Billington, D. (2000). Strategies in human nonmonotonic reasoning. *Computational Intelligence, 16 (3),* 446-468.

Ginsberg, M. L. (1994). AI and nonmonotonic reasoning. In D. M. Gabbay, C. J. Hogger, & J. A. Robinson (Eds.), *Handbook of logic in artificial intelligence and logic programming. Vol. 3: Nonmonotonic reasoning and uncertain reasoning*. Oxford: Clarendon Press.

Hewson, C. and Vogel, C. (1994). Psychological evidence for assumptions of path-based inheritance reasoning. *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Atlanta, Georgia.

Lifschitz, V. (1988). Benchmark problems for formal nonmonotonic reasoning, version 2.00. In J. Siekmann (Series ed.) & M. Reinfrank, J. de Kleer, M.L. Ginsberg, & E. Sandewall (Vol. Eds.), *Lecture notes in Artificial Intelligence. Nonmonotonic reasoning*. Berlin : Springer-Verlag.

Pelletier, F. J., & Elio, R. (1997). What should default reasoning be, by default? *Computational Intelligence, 13,* 165-187.

Schurz, G. (2001). What is 'normal'? An evolution-theoretic foundation for normic laws and their relation to statistical normality, *Philosophy of Science, 68* (4), 476-497.

Vogel, C. and Tonhauser, J. (1996). Psychological constraints on plausible default inheritance reasoning. In Aiello and Shapiro (Eds.), *Proceedings of the 5th International Conference on Principles and Practice of Knowledge Representation, KR`96*. Cambridge, Mass.: Morgan Kaufmann. 608-19.

The Role of Explanation Coherence of Two Premises on Property Induction

Kyung Soo Do (ksdo@skku.edu)
Ju Hwa Park (jhpark@skku.edu)
Department of Psychology, Sungkyunkwan University
Seoul 110-745, KOREA

Abstract

Three experiments on property induction were conducted to explore whether an incoherent premise discounted the believability of the conclusion when there were two premises. In all three experiments, a single premise increased or decreased the likelihood of the conclusion depending on the nature of explanatory coherence of a premise and a conclusion. However, when there were two premises, one that shared the reason with the conclusion (coherent premise), and one that does not (incoherent premise), the believability of the conclusion was affected differently in three experiments. When two premises and the conclusion were presented simultaneously in Experiment 1, the believability of the conclusion was increased. That is, an incoherent premise did not seem to affect the believability of the conclusion as much as the coherent premise. The incoherent premise seemed to decrease the believability of the conclusion a little bit when the two premises and the conclusion were presented sequentially so that each premise was not ignored in Experiment 2. The incoherent premise decreased the believability of the conclusion below the baseline condition in Experiment 3, where participants were asked to write down reasons for each premise being true. Results of three experiments suggested that only the confirming evidences were processed under natural conditions. A few possible theoretical implications were considered.

Introduction

When someone asks a question whether a target object has a certain property (target property), such as "Does an ostrich lay eggs?", and you do not know the answer, you might induce the answer by checking whether some object (source object), usually objects that are similar to the target object, has the target property. In the ostrich example, you would answer "yes" if you think ostriches are similar to geese and know that geese lay eggs. As this example shows, what conclusion you make depends on what objects are used as source objects.

What object is an effective source object in property induction depends on a number of factors: The nature of the target property, the level of knowledge of the person, and the cultural background of the person, to name a few. People used different source objects depending on their knowledge and occupation (Proffitt, Coley, & Medin, 2000). Also there seems to be a culture difference in property induction (Choi, Nisbett, & Smith, 1997). Even though the level of knowledge and the cultural background of the person affect the property induction, the effect of the nature of the target property on property induction has been the focus of most research. More specifically, what objects are effective as source objects in inducing two types of properties, and how the information of the source objects are used for property induction have been more widely investigated.

There are two types of target properties, blank properties and nonblank properties, and the effectiveness of source objects seems to differ between the two types. The effectiveness of a source object seems to depend on the similarity between the target object and the source object for blank properties for which we do not have any other information to rely on (e.g., 'have BCC in blood') (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975; Sloman, 1993). However, the similarity between the target object and the source object does not seem to work for nonblank properties for which we have other information to infer about the target object having the target property (e.g., 'can cut the wire') (Smith, Shafir, & Osherson, 1993). As the relationship between objects and the target property are diverse (Murphy & Medin, 1985), there are many ways of inducing nonblank properties: People seem to use other relevant information, such as body size or strength, in inducing nonblank properties (Smith *et al*., 1993). People rated the believability of the conclusion differently when the target property is about shape from when the target property is about behavior (Heit & Rubinstein, 1994).

One way of inducing nonblank properties is comparing the reason for the target object having the target property with the reason for the source object having the target property. Sloman (1994, 1997) proposed that the explanation coherence between the premise and the conclusion affect the plausibility of the conclusion. If the target object and the source object share the same explanation, informing the participants that the source object has the target property increases the believability of the conclusion that the target object has the target property. For instance, computer programmers and secretaries have bad backs because they sit all day long. Therefore, informing the participants that "Computer programmers have bad backs" would make the conclusion "Secretaries have bad backs" more plausible than when the participants are not informed about the computer programmers having bad backs. However, if the target object and the source object do not share the same reason, informing the participants that the source object has the target property decreases the believability of the conclusion that the target object has the target property. In the bad back example, for instance, furniture movers had bad backs because they lift heavy things. Therefore, informing that "Furniture movers had bad backs" would make the conclusion "Secretaries

have bad backs" less plausible. That is, premises that have different explanations seem to discount the plausibility of the conclusion. Sloman called this the explanation discounting principle.

As has been described, the explanation discounting principle explains empirical results quite well when there is just one piece of relevant information. However, it is not specific about how it works when there are multiple pieces of relevant information, especially when there is conflicting information. There can be a few modified versions of the explanation discounting principle. The most extreme form of the explanation discounting principle would assume that the participants use all the information in inducing properties in equal degree (equivalence hypothesis, but hereafter I use equivalence hypothesis and the explanation discounting principle interchangeably). However, the equivalence hypothesis needs to be tested, because when there is more than one piece of relevant information, people do not use all the information they have. People are known to be cognitive misers (Nisbett & Ross, 1980). They seem to use information that confirms their predictions or hypotheses, but ignore information that disconfirms. They produced cases that confirmed their hypotheses (Wason, 1960) or they were more willing to search information that confirms than information that is likely to disconfirm (Shaklee & Fischhoff, 1982). Therefore if the confirmation strategy is the default way of using information, then it is quite likely that the explanation discounting principle may not apply when there is more than one piece of relevant information.

In this paper, we intend to explore whether the explanation discounting principle works when there are two premises in property induction tasks. There are two kinds of premises: "Same" premises, in which the source object has the target property for the same reason as the target object, and "Different" premises, in which the source object has the target property for reasons different from that of the target. In three experiments, we are mainly interested in the believability of the conclusion of the mixed conditions where one Same premise and one Different premise are presented. If the implicit equivalence hypothesis of the explanation discounting principle was correct, the conclusion in the mixed conditions should be rated not higher than that of the baseline condition, where the conclusion is presented without any premises. On the other hand, if the confirmation strategy is the default mode of using multiple pieces of information in property induction, the conclusion in the mixed conditions should be rated not lower than that of the baseline condition.

The premises and the conclusion were presented simultaneously in Experiment 1 to explore whether the explanation discounting principle applies when there are two premises. Experiment 1 is regarded as a natural condition because we did not try any manipulation to make the premises being processed. Experiments 2 and 3 were intended to find the boundary condition where the explanation discounting principle applies. Each premise and the conclusion were presented successively in Experiment 2 to make each premise salient and not be ignored. In Experiment 3, participants were asked to write down the reason why the object has the target property for each premise.

Experiment 1

There were two goals for Experiment 1. First, we wanted to replicate Sloman's (1994, 1997) finding that one Same premise increased the plausibility of the conclusion, and one Different premise decreased the plausibility of the conclusion. Second, we wanted to compare the explanation discounting principle against the confirmation strategy by presenting two premises, one Same premise and one Different premise.

Method

Design There were five experimental conditions in Experiment 1. In two single-premise conditions, one premise was presented on top of the conclusion. There was a horizontal line between premises and the conclusion. In the Same condition, one Same premise was presented, and in the Different condition, one Different premise was presented. Two premises were presented with the conclusion in the remaining three two-premises conditions: In the S+S condition, two Same premises were presented on top of the conclusion. In the S+D condition, the Same premise was presented on the top line and the Different premise was presented on the next line. In the D+S condition, the Different premise was presented on the top line and the Same premise was presented on the next line. Each participant was randomly assigned to one of the five experimental conditions. Therefore, the premise condition was a between subjects variable.

Participants Ninety-five Sungkyunkwan University students who attended an "Introduction to Psychology" course participated as a requirement for the course. Nineteen participants were randomly assigned to each experimental condition. None of them had participated in property induction experiments prior to the current experiment.

Material Twelve properties were used in the experiment as the target property. The target properties and the corresponding occupations were selected based on the results of an item selection experiment. In the item selection experiment, two hundred Sungkyunkwan University students were asked to write down at least two occupations that have the target property and the reasons they have the target properties over 24 properties. The 24 properties were selected from 32 items used in Sloman (1994, 1997) and judged appropriate in Korea. Of the 24, 12 properties were selected as experimental material. All premises and conclusions took the form of an occupation or class of people having the target property, such as "Veterans have problems getting jobs."

Procedure There were three stages in the experiment. At Stage 1, participants were presented only the conclusion of

twelve induction problems, and were asked to rate the probability of each conclusion. Each conclusion was presented on a computer monitor screen one at a time. The presentation order of the twelve conclusions was randomized within a subject. The rating at Stage 1 was used as a baseline rating of the participant. After they finished baseline estimation, they did an intervening task for more than five minutes (Stage 2). The intervening task was not related with property induction, or any of the properties or the occupations used in the experiment. After the participants completed the intervening task, they were presented twelve experimental property induction problems and asked to rate the probability of each conclusion considering the premises. After they finished rating twelve induction problems, they were given the induction problems and their ratings for each problem and were asked to type in the reason for their response. The presentation order of the twelve induction problems was randomized within a subject. Presentation of the items and recording of responses in Stages 1 and 3 were manipulated by a program written in Visual Basic 6.0. Pentium-class PCs and computer monitors were used in Stages 1 and 3.

Results and discussion

Rating Average ratings of the baseline (Stage 1) and the experiment phase (Stage 3) for each premise condition are presented in Fig. 1. As the difference between the rating in the baseline phase and that of the experiment phase was the main interest, ratings in the baseline phase and that of the experiment phase were regarded as a within-subjects variable, and one factor within-subjects ANOVA was conducted for each premise condition**.**

In single-premise conditions, presenting a Same premise increased the rating in the Same condition, $F(1, 18) = 4.87$, $p < .05$, $MSE = 92.96$, and presenting a Different premise decreased the rating in the Different condition, $F(1, 18) = 51.93$, $p < .001$, $MSE = 21.92$. Results in the single-premise conditions replicated Sloman's (1994) results and corroborated the explanation discounting principle.

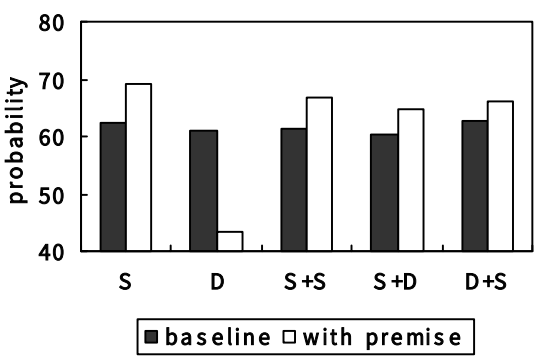


Figure 1. Average ratings of the conclusion: Experiment 1. (S: Same; D: Different)

However, the explanation discounting principle did not seem to apply in the two-premises conditions. Presenting two premises increased the rating of the conclusion in the S+S condition, $F(1, 18) = 21.92$, $p < .001$, $MSE = 13.25$, and in the S+D condition, $F(1, 18) = 11.09$, $p < .001$, $MSE = 17.54$, and did not affect the rating of the conclusion in the D+S condition, $F(1, 18) = 2.14$, ns. According to the explanation discounting principle, the conclusion in the S+D and D+S conditions, in which there was a premise that has the target property for reasons different from that of the target object, was expected to yield ratings at least not higher than that of the baseline phase. However, even in the D+S condition, where the difference from the baseline is smaller than the S+D condition, the average rating for the experiment phase was a little larger than that of the baseline, though not statistically significant. Thus, the results in S+S, S+D, D+S conditions seemed to fair better with the confirmation strategy. That is, participants might have processed only the information that can confirm or strengthen the plausibility of the conclusion when there are two pieces of conflicting information. The possibility of adopting the confirmation strategy got further support from participants' subjective reasons for their responses.

Subjective report The reasons participants wrote down for their conclusions in the experiment were classified into 11 possible categories in the single-premise condition and 20 possible categories in the two-premises conditions.

In the single-premise conditions, participants seemed to use the information in the premise. More specifically, in the Same condition, about 65% of the reasons matched that of the experimenter. In the Different condition, about 35% reported that the reasons for the premise and the conclusion did not agree. In general, results in the single-premise conditions suggested the explanation discounting principle seemed to apply when there is just one piece of relevant information.

However, participants seemed to mainly use confirming information and ignore disconfirming information when there were two premises. More specifically, in the S+S condition, 58% of the responses mentioned the premises and the conclusion had the same reason. In the S+D and D+S conditions, 29% of the responses mentioned only the Same premise, and 21% mentioned reasons they spontaneously made to make both the premise and the conclusion shared the same reason. In other words, in about half of the responses, participants searched for reasons that are the same as the conclusion. Of the remaining 50%, 22% of the response mentioned only the conclusion. As a whole, participants' subjective reports in two-premises conditions strongly suggested that they adopted the confirmation strategy.

Experiment 2

The results of Experiment 1 strongly suggested that participants processed only a part of the information given to them. That is, they seemed to use information that gave

support for the conclusion, and ignore information that was incoherent with the conclusion. Experiments 2 and 3 were intended to test this possibility of non-use of disconfirming information by making premises salient so that disconfirming information was not to be ignored. In Experiment 2, premises were made salient by presenting the premise(s) and the conclusion one after the other.

Method

Participants Ninety-five Sungkyunkwan University students participated in Experiment 2. They were recruited in the same way as that of Experiment 1. Nineteen participants were randomly assigned to one of the five premise conditions.

Material The materials of Experiment 2 were identical to that of Experiment 1.

Procedures The procedures of Experiment 2 were identical to that of Experiment 1, except for the following changes in the temporal order of presenting premises and conclusion at Stage 3. At Stage 3 of Experiment 2, the premise on the top line of the screen appeared and remained visible until participants made responses indicating their rating of the conclusion. The second premise, if there was one, appeared on the screen 3 seconds after the start of the first premise and remained visible until participants made responses. Finally, the conclusion appeared on the screen 3 seconds after the onset of the last premise, and remained visible until the response. The sequential presentation of the premises and the conclusion was intended to make sure that the premises not be ignored.

Results and Discussion

Rating Average ratings of the baseline and the experiment phase for each premise condition are presented in Fig. 2. In single-premise conditions, presenting a Same premise increased the rating in the Same condition, $F(1, 18) = 6.00$, $p < .001$, $MSE = 21.92$, and presenting a Different premise decreased the rating in the Different condition, $F(1, 18) = 16.29$, $p < .001$, $MSE = 133.96$.

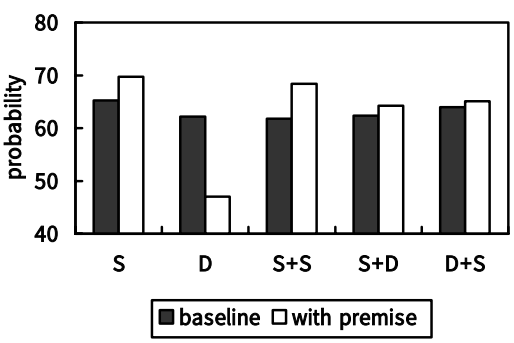


Figure 2. Average ratings of the conclusion: Experiment 2. (S: Same; D: Different)

Presenting two premises increased the rating of the conclusion in the S+S condition $F(1, 18) = 14.15$, $p < .001$, $MSE = 29.26$, but did not affect the rating of the conclusion in the S+D condition, $F(1, 18) = 1.50$, ns, and in the D+S condition, $F(1, 18) = .79$, ns. Different from Experiment 1, the ratings of the two mixed conditions, the S+D and the D+S conditions, were not different from that of a baseline, which suggested that making premises not ignored by presenting one after the other makes all the information attended and as a consequence can exert both facilitating and discounting effect on property induction, even though the discounting effect seems not as strong as the facilitating effect.

Subjective report As in Experiment 1, participants seemed to use the information in the premise in the single-premise conditions. More specifically, in the Same condition, about 73% of the reasons matched that of the experimenter. In the Different condition, about 44% reported that the reasons for the premise and the conclusion did not agree.

The pattern of responses in the two-premises conditions of Experiment 2 was similar to that of Experiment 1. 52% of the responses mentioned that the premises and the conclusion had the same reason in the S+S condition. In the S+D and D+S conditions, 28% of the responses mentioned only the Same premise, and 22% mentioned reasons they spontaneously made to make both the premise and the conclusion share the same reason. Furthermore, 26% of the responses mentioned only the conclusion.

In general, results of Experiment 2 were quite similar to that of Experiment 1, but sequentially presenting premises did at least partially succeed to make information that disconfirms the conclusion affect the believability of the conclusion.

Experiment 3

Presenting the premises and the conclusion successively changed the pattern of results a little in Experiment 2. In Experiment 3, the disconfirming premise was forced to be processed by asking participants to write down reasons why each premise could be true.

Method

Participants Ninety-five Sungkyunkwan University students participated in Experiment 3. They were recruited in the same way as that of Experiment 1. Nineteen participants were randomly assigned to one of the five premise conditions**.**

Material The materials of Experiment 3 were identical to that of Experiment 1.

Procedures The procedures of Experiment 3 were identical to that of Experiment 1, except for the following three changes. First, in Experiment 3, participants were tested in groups. Nineteen participants in each premise condition were seated in a large class room. They were seated in a way such that there was at least one seat unoccupied in all directions. Second, participants were given a small booklet. Third, participants were asked to write down the reasons for the premises being true and the believability rating of the conclusion. Separate booklets were given at each stage, so that participants could not look at their baseline ratings when they did induction problems. In the booklet for Stage 3, the conclusion was printed in a page following the page where premises and their responses for the premises were written, so that participants could not read their reasons for the premises**.**

Results and Discussion

Rating Average ratings of the baseline and the experiment phase for each premise condition are presented in Fig. 3. In single-premise conditions, presenting a Same premise increased the rating in the Same condition, $F(1, 18) = 8.03$, $p < .05$, $MSE = 61.53$, and presenting a Different premise decreased the rating in the Different condition, $F(1, 18) = 38.02$, $p < .001$, $MSE = 78.76$.

Presenting two premises increased the rating of the conclusion in the S+S condition $F(1, 18) = 9.55$, $p < .01$, $MSE = 74.76$, but decreased the rating of the conclusion in the S+D condition, $F(1, 18) = 8.70$, $p < .01$, $MSE = 38.00$, and D+S condition, $F(1, 18) = 9.71$, $p < .01$, $MSE = 65.67$.

In general, making premises salient by writing down reasons why they can be true did not affect the effects of confirming information, probably because the confirming information had already exerted its influence due to the confirmation strategy people spontaneously use in most situations. However, presenting a disconfirming premise decreased the rating of the conclusion in Experiment 3 in a much larger degree, and succeeded to give very strong support for the explanation discounting principle when there were two conflicting premises.

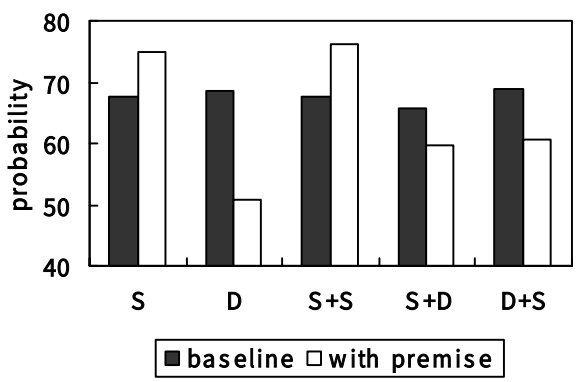


Figure 3. Average ratings of the conclusion: Experiment 3. (S: Same; D: Different)

As a whole, the results of the three experiments seemed to suggest that the explanation discounting principle seemed to work only when the disconfirming information became salient by either being presented one by one or by forcing respondents to think about the reasons.

General Discussion

Three experiments were conducted to explore whether the explanation discounting principle works when there are two conflicting premises. The results of the three experiments can be summarized as follows: (1) Both the confirmation strategy and the explanation discounting principle seemed to work when there was just one premise. In three experiments, it has been consistently observed that the Same premise increased the rating of the conclusion, supporting the confirmation strategy, and that the Different premise decreased the rating of the conclusion, supporting the explanation discounting principle. Results of the single-premise conditions suggested that people seemed to search for relevant information and use it when they had just one piece of relevant information. (2) However, only the Same premise(s) seemed to affect the plausibility of the conclusion when there were two premises under natural conditions. Ratings in the two mixed conditions were higher than or equal to the baseline in Experiment 1, but got equal or lower than the baseline when the premises were forced to be processed in Experiments 2 & 3. Our interpretation that only the confirming information seemed to influence the judgments and decisions is in good agreement with the information processing strategies generally accepted in cognitive psychology, such as Johnson-Laird & Byrne (1991) and Nisbett & Ross (1980).

However, the explanation discounting principle can explain the results of the three experiments if the implicit assumption that all information is processed in the same degree was modified. First, as I mentioned in the Introduction, the explanation discounting principle did not make any explicit assumption concerning the fate of conflicting information. Therefore, our interpretation of the discounting principle might be an unfair test of the discounting principle. Second, the relevance of the confirming premise and the disconfirming premise might be different. For instance, if we adopt the coherence of Thagard (1992), confirming premises share more attributes with the conclusion than the disconfirming premises. More specifically, the confirming premise shares the reason and the consequences of the reason with the conclusion (e.g., in the bad back example, both programmers and secretaries share two attributes, sit all day long and have bad backs), whereas the disconfirming premise shares only the consequences (e.g., furniture movers and secretaries have one attributes in common, they have bad backs). Therefore, the explanation discounting principle can explain the results of current experiments if their relevance were used as

relative weights of each premise. However, the explanation discounting principle still has problems explaining why making premises salient decreased the believability of the conclusion below the baseline in Experiment 3.

One aspect that has to be solved in the preceding argument is who, what, or when determines the processing order of the information. That is, deciding whether certain information is confirming or disconfirming to the conclusion can be solved only after we figure out the conclusion in the property induction tasks. Therefore the order of processing information might be different from the order the information is given. If this is the case, then there have to be multiple stages of processing. For instance, a primitive assessment of the relevance/confirmation of premises to the conclusion precedes the detailed processing of the relevant or confirming information.

Acknowledgements

This work was supported by the Korea Research Foundation Grant (KRF-2002-074-HS1002).

References

Choi, I., Nisbett, R. E., & Smith, E. E. (1997). Culture, categorization, and inductive reasoning. *Cognition, 65*, 15-32.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 411-422.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Erlbaum.

Murphy, L., M., & Medin, D. M. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Nisbett, R. E., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgments.* Englewood Cliffs, NJ: Prentice Hall.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category based induction. *Psychological Review, 97*, 185-200.

Proffitt, J. B., Coley, D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 811-828.

Rips, L. J. (1975). Inductive judgments about natural categories *Journal of Verbal Learning and Verbal Behavior, 14,* 665-681.

Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition, 10*, 520-530.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25, 231-280.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition, 52*, 1-21.

Sloman, S. A. (1997). Explanatory coherence and the induction of property. *Thinking and Reasoning, 3*, 81-110.

Smith, E.E., Shafir, E., & Osherson, D.N. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67-96.

Thagard, P. (1992). *Conceptual revolution*. Princeton; Princeton University Press.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

A Fundamental Limitation of Symbol-Argument-Argument Notation As a Model of Human Relational Representations

Leonidas A. A. Doumas (adoumas@psych.ucla.edu)
John E. Hummel (jhummel@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
405 Hilgard Ave.
Los Angeles, CA 90095-1563

Abstract

Human mental representations are both structure-sensitive (i.e., symbolic) and semantically rich. Connectionist models have well-known limitations in capturing the structure-sensitivity of mental representations, while traditional symbolic models based on varieties of symbol-argument-argument notation (SAA) have difficulty capturing their semantic richness. We argue that this limitation of SAA is fundamental and cannot be alleviated in the notational format of SAA itself. Finally, we review an approach to human mental representation that captures both its structure-sensitivity and semantic richness.

Relational reasoning—reasoning constrained by the relational roles that objects play, rather than just the features of the objects themselves—is ubiquitous in human mental life, and includes analogical inference, schema induction, and the application of explicit rules (Gentner, 1983; Holyoak & Thagard, 1995). In order to support human-like relational thinking, a representational system must meet two general requirements (Hummel & Holyoak, 1997): First, it must represent relations independently of their fillers and simultaneously specify how fillers are bound to relational roles (i.e., it must be a *symbol system*; Newell, 1990). Second, it must explicitly specify the semantic content of relational roles and their fillers. In this paper, we consider in detail the implications of the latter requirement, with an emphasis on symbol-argument-argument notation (SAA), which includes propositional notation, high-rank tensors, and many varieties of labeled graphs.

Properties of Relational Representations

Human Relational Representations are Symbolic

Symbolic representations have the property that symbols are invariant with their role in an expression[1], and the meaning of the expression as a whole is a function of both the symbols and their arrangement (i.e., role-filler bindings). For example, the expressions *chase* (Pat, Don) and *chase* (Don, Pat) mean different things, but Pat, Don, and *chase* mean the same things in both expressions. Formal symbol systems have this property by assumption: It is given in the definition of the system that symbols retain their meanings across different expressions, and that the meaning of an expression is a function of both its constituent symbols and their arrangement. Physical symbol systems (Newell, 1990)—such as digital computers and human brains—cannot simply "assume" these properties, but instead must actively work to ensure that both are satisfied. The claim that human mental representations are symbolic in this sense is controversial (e.g., Elman et al., 1996), however, relational generalization has proved unattainable for non-symbolic models of cognition, and there is reason to believe it is fundamentally unattainable for such models (see Doumas & Hummel, in press; Halford, Wilson, & Phillips, 1998; Hummel & Holyoak, 2003a; Marcus, 1998).

Human Relational Representations Specify the Semantic Content of Objects and Relational Roles

A second important property of human relational representations is that they explicitly specify the semantic content of objects and relational roles (e.g., the *lover* and *beloved* roles of *love* (x, y) or the *killer* and *killed* roles of *murder* (x, y)): We know what it means to be a lover or a killer, and that knowledge is part of our representation of the relation itself (as opposed to being specified in a lookup table, a set of inference rules, or some other external structure). As a result, it is easy to appreciate that the patient (i.e., *killed*) role of *murder* (x, y) is like the patient role of *manslaughter* (x, y), even though the agent roles differ (i.e., the act is intentional in the former case but not the latter); and the agent (i.e., *killer*) role of *murder* (x, y) is similar to the agent role of *attempted-murder* (x, y), even though the patient roles differ.

More evidence that we represent the semantics of relational roles explicitly is that we can easily solve mappings that violate the "n-ary" restriction (Hummel & Holyoak, 1997). That is, we can map n-place predicates onto m-place predicates where $n \neq m$. For instance, given statements such as *taller-than* (Abe, Bill), *tall* (Chad) and *short* (Dave), it is easy to map Abe to Chad and Bill to

[1] This is not to say that symbols may not have shades of meaning that vary from one context to another. For example, the symbol "*loves*" suggests different relations in *loves* (Mary, John) vs. *loves* (Mary, Chocolate). However, as noted by Hummel and Holyoak (2003a), this kind of contextual shading must be a function of the system's knowledge, rather than an inevitable consequence of the way in which it binds relational roles to their fillers.

Dave. Given *shorter-than* (Eric, Fred), it is also easy to map Eric to Bill (and Dave) and Fred to Abe (and Chad). These mappings are based on the semantics of the individual relational roles rather than the formal syntax of propositional notation, or, say, the fact that *taller-than* and *shorter-than* are logical opposites: *love* (*x*, *y*) is in some sense the opposite of *hate* (*x*, *y*), but in contrast to *taller-than* and *shorter-than* (in which the first role of one relation maps to the second role of the other) the first role of *love* (*x*, *y*) more naturally maps to the first role of *hate* (*x*, *y*). In short, the similarity and/or mappings of various relational roles are idiosyncratic, based on the semantic content of the individual roles in question. The semantics of relational roles matter, and are an explicit part of the mental representation of relations.

The semantic properties of relational roles are also evidenced in numerous other ways in human cognition, including memory retrieval (e.g., Gentner, Ratterman & Forbus, 1993; Ross, 1989), and analogical mapping and inference (Bassok, Wu & Olseth, 1995; Kubose, Holyoak & Hummel, 2002; Krawczyk, Holyoak & Hummel, in press; Ross, 1987). Indeed, the meanings of relational roles influence relational thinking even when they are irrelevant or misleading (e.g., Bassok et al., 1995; Ross, 1989), suggesting that access to and use of role-based semantic information is quite automatic. This information appears to be an integral part of the mental representation of relations. Given its centrality in human cognition, an important criterion for a general account of human mental representation is that it must represent relations in a way that captures the semantics of their roles.

SAA Accounts of Relational Representations

Numerous models of human cognition account for the symbolic nature of relational representations by postulating representations based on varieties of symbol-argument-argument notation (SAA). These include propositional notation, varieties of labeled graphs, and high-rank tensor products (e.g., Anderson & Lebiere, 1998; Eliasmith & Thagard, 2001; Falkenhainer, Forbus, & Gentner, 1989; Forbus, Gentner, & Law, 1995; Halford et al., 1998; Holyoak & Thagard, 1989; Keane, Ledgeway, & Duff, 1994; Ramscar & Yartlett, 2003; Salvucci & Anderson, 2001). In SAA notation relations and their fillers are represented as independent symbolic units, which are bound together to form larger relational structures. For example, in propositional notation the statement "John loves Mary" is represented by binding the symbol for the predicate *love* to a list of its arguments, here John and Mary, forming the proposition *love* (John, Mary). By virtue of their positions in the list of arguments, John is taken to be the filler of the first role of the *loves* relation (i.e., the *lover*), and Mary is the filler of the second role (i.e., the *beloved*). Labeled graphs use location in the graph to indicate relational bindings in much the same way: Relations are represented independently of their arguments (fillers), and relation-filler bindings are represented in terms of the fillers' locations in the graph. As such, SAA is meaningfully symbolic in the sense described above.

Because SAA systems are meaningfully symbolic they naturally support operations that require relational representations, such as structure mapping (a.k.a., analogical mapping; see e.g., Falkenhainer et al., 1989; Forbus et al., 1995) and matching symbolic rules (see Anderson & Lebiere, 1998). This is no small accomplishment. Numerous representational schemes, including traditional distributed connectionist representations (e.g., Elman, et al., 1996) and the kinds of representations formed by latent semantic analysis (e.g., Landauer & Dumais, 1997) have not succeeded in modeling relational perception or cognition, and, as noted above, there are compelling reasons to believe they are fundamentally ill-suited for doing so.

The successes of SAA notation as a model of human mental representation are thus decidedly non-trivial. At the same time, however, SAA notation has greater difficulty capturing the semantic content of human mental representations. Indeed, this limitation was a central part of the criticisms leveled against symbolic modeling by the connectionists in the mid-1980s (e.g., Rumelhart, McClelland, & the PDP Research Group, 1986; for a more recent treatment, see Hummel & Holyoak, 1997). One of the strengths of distributed connectionist representations is their natural ability to capture the semantic content of the objects they represent. By contrast, propositional notation and labeled graphs have difficulty making this semantic content explicit. As a result, symbolic models based on SAA notation often resort to a patchwork of external fixes such as look-up tables, inference rules and arbitrary, hand-coded similarity matrices (e.g., Salvucci & Anderson, 2001; high-rank tensors, e.g., Halford et al., 1998, are a notable exception, but see Doumas & Hummel, in press, and Holyoak and Hummel, 2000, for detailed treatments of the limitations of this approach).

Importantly, these fixes are external to the relational representations themselves: They are not instantiated in the notation of the representations that encode relations, but rather only at the level of the system that *processes* these representations. This separation is fine for accounts of cognition at Marr's (1982) level of computational theory, but for accounts at the level of representation and algorithm such fixes entail specifying multiple sets of representations (at minimum one for the notational system capturing the relations and a second for the system that captures the relations between different relations and/or semantics of those relations), and an interface for moving between them.

Moreover, relations in SAA are represented as indivisible wholes, leaving the roles implicit as semantically empty place-holders or arcs in a graph. For example, the *love* (*x*, *y*) relation is represented by a single symbol (in propositional notation), a single vector (in high-rank tensor products), or a single node (in a labeled graph)[2], and its roles, *lover* (*x*) and

[2] Not all models that employ labeled-graphs are based on SAA (e.g., Larkey & Love, 2003), although most are.

beloved (*y*), are not represented explicitly at all. As such, even if a relational symbol (as a whole) is assumed to have semantic content, SAA notation itself does not specify how this content is differentially applicable to its arguments. It is this limitation that gives rise to the *n*-ary restriction: Even if the symbol *taller-than* is assumed to have meaning, the expression *taller-than* (*x*, *y*) does not explicitly specify how the role filled by *x* is semantically similar to *tall* (*z*), so it provides no basis for mapping *taller-than* (*x*, *y*) onto *tall* (*z*).

Recall the semantic relations between the roles of *murder* (*x*, *y*), *manslaughter* (*x*, *y*), and *attempted-murder* (*x*, *y*). SAA notation can, at best, treat the entire relations as simply "similar" or "different". High-rank tensor systems, for instance, represent entire relations (but not their roles) as vectors, and compute the similarities between relations based on the similarities of their vectors. Models using propositional notation and labeled graphs can "recast" predicates into more abstract forms (e.g., coding both *murder* (*x*, *y*) and *manslaughter* (*x*, *y*) as *commit-violence-against* (*x*, *y*)). Neither of these approaches expresses the similarity relations between the individual roles of relations, however, because they fail to make individual relational roles (and thus their semantic content) explicit. As a result, they do not make clear how the roles of *manslaughter* (*x*, *y*) and *attempted-murder* (*x*, *y*) are related to the roles of *murder* (*x*, *y*), nor how they differ.

It is easy to imagine a look-up table or set of inference rules that would supply this information, but the number of rules required to specify the similarity relations between all relational roles would scale minimally with $(n^2 - n)/2$ (or $n^2 - n$ if equivalent bidirectional similarity is not assumed), where n is the number of relational roles in the system. Worse, these rules would be external to the system's SAA-based representational system, so an account of how the system operates at the level of representation and algorithm requires a description of the rule set, and an additional control structure to read the SAA and access the rules as necessary. All of this is obviated in a system that simply codes the semantic content of individual relational roles explicitly in the notation of the relational representations.

The convenience of "postulate them as you need them" inference rules makes it is tempting to assume that the lack of role-specific semantic information in SAA is merely a thorny inconvenience: Surely the problem of role-based semantics in SAA is solvable, and will be solved as soon as it becomes important enough for someone to give it attention. In the meantime, it is certainly no reason to abandon SAA as a model of mental representation, especially if the only alternatives are non-symbolic representational schemes (to anticipate, they are not).

But it turns out that the problem is more than just a thorny inconvenience. As elaborated in the next section, role-specific semantic information cannot be made internal to (i.e., explicit in) the notation of SAA. Instead, the knowledge of the semantics of individual roles can only be specified at the level of the whole system, instantiated in external routines or representational structures. As a result, systems employing SAA (i.e., traditional symbolic models) are at best incomplete as algorithmic accounts of human cognition.

Representing Relational Roles in SAA

In a symbol system, representing something explicitly means representing it as a structure (e.g., a proposition). Thus, to represent relational roles and their semantic content explicitly in SAA, it is necessary to represent them structurally (i.e., to *predicate* them). For example, to represent the *murder* (*x*, *y*) relation in terms of its *killer* (*x*) and *victim* (*y*) roles, SAA must represent *killer* (*x*) and *victim* (*y*) as propositions (or tensors, or nodes), and simultaneously represent the fact that together they compose the *murder* (*x*, *y*) relation.

A relation specifies that its arguments are engaged in certain specific functions or states. For instance, the statement *drive* (Brutus, the Honda) specifies that Brutus is playing the role of the individual driving, and that the Honda is playing the role of the thing being driven. This information is entailed by the relational statement itself. As a separate matter, the relation may also imply or suggest other relations (e.g., that Brutus knows how to drive, that the Honda is in running condition, etc.), but it directly entails only the *driver* and *driven-object* roles, and the binding of Brutus and the Honda to these roles.

A relation does more than simply imply its roles (just as the roles do more than simply imply their parent relation), it *consists of* them: Collectively, the roles, along with their linkage to one another, *are* the relation. Imagine a relation *R* (*p*, *q*), with roles $R_1(p)$ and $R_2(q)$, that implies relation *S* (*p*, *q*), with roles $S_1(p)$ and $S_2(q)$. Although *R* (*p*, *q*) entails or implies a total of four roles (i.e., those of *R* and those of *S*), only two (R_1 and R_2) compose *R* itself. Therefore, representing relational roles explicitly requires explicitly specifying which relations consist of which roles.

As summarized above, this information seems to come gratis (i.e., without computational cost) as an integral component of human relational representations. An adequate model of human relational representations should, therefore, capture this information in its relational representations, rather than relegating it to the processes that operate on these representations or to external representational systems. This distinction is subtle, but important. It is not enough that the system as a whole capture this information; instead the same representations that encode the relations should capture it. That is, the information should be captured at the level of the notation, or language, of the system rather than in an external system of inference rules, look-up tables, etc.

However, it is not possible to capture relations, their roles and the relation between them in the notation of SAA. In SAA representing a relation explicitly requires instantiating a proposition. If representing a relation implies representing its roles, and representing a relation's roles requires representing the relation between the relation and its roles, then representing any given relation implies a second

proposition representing the relation between the first relation and its roles. This second proposition contains a new relation, though, which implies roles of its own that must be related back to it in a third proposition, which also contains a new relation with roles that must be related back to it, and so on.

For example, imagine the propositions *love* (John, Mary) and *love* (Bill, Sally) and their role bindings, *lover* (John), *beloved* (Mary), *lover* (Bill) and *beloved* (Sally). The representations must specify which role bindings go together to form which relations (e.g., that *lover* (John) goes with *beloved* (Mary) – as opposed to *beloved* (Sally) or *lover* (Bill) – to form the complete relation *love* (John, Mary)). Specifying this composition relation in SAA notation requires a proposition of the form *consists-of* (*love* (John, Mary), *lover* (John), *beloved* (Mary))[3]. The predicate *consists-of*, however, is itself a relation, and so specifies a set of role bindings of its own, which must be explicitly represented and linked back to the original *consists-of* relation with a second *consists-of* relation. This second *consists-of* is also a relation and so specifies a set of role bindings of its own that must be linked back to it via a third *consists-of* relation, and so on.[4] The consequence is an infinite regress that can render the resulting representational system ill-typed (Manzano, 1996).

In other words, it is not possible to represent the relation between relational roles and complete relations explicitly in the notation of SAA. For the same reason, it is not possible to specify the semantic content of relational roles explicitly in SAA. As noted previously, this information can be specified in a complete *system* based on SAA (after all, many such systems are Turing complete), but it is not possible to capture this information in the SAA notation itself. Instead, the knowledge is necessarily contained in the processes that operate on the SAA representations, or in external representational systems that must be interfaced with the SAA notation. This property renders SAA notation fundamentally inadequate as a model of human mental representations. It does not imply that representing the semantic content of relational roles explicitly is impossible in an SAA based system; it simply indicates that such systems cannot capture this content in the same way that people seem to—i.e., as an integral part of the relational representation itself.

It is important to emphasize that the problem of infinite regress is by no means an argument against the general utility of SAA. On the contrary, propositional notation, for example, is an extremely useful tool for the purpose for which it was created, namely theorem proving. For this purpose, semantic emptiness is a virtue. It just happens that the design requirements for a representational system for theorem proving differ markedly from the design requirements for a representational system for supporting perception and cognition in living, behaving systems: While the former requires pristine semantic emptiness, the latter must deal with the semantically-rich, sometimes ugly realities of the world as it is. Given this, it is not surprising that a representational system designed for theorem proving should prove inadequate as a model of human mental representation.

Possible Objections and Responses

We are arguing that SAA is ill-suited for modeling human mental representations, including relational representations, for which it appears at first to be ideally suited. Proponents of traditional symbolic models of cognition are likely to object strenuously to our arguments. We shall try to anticipate some of these objections and respond to them.

One potential objection is that the infinite regress is irrelevant because it is a trivial matter to simply terminate the regress at any arbitrary point and declare it done. The problem with this objection is that the system does not finish the process it began. Once the progress halts, the relation between roles and relations remains unspecified.

Another potential objection is that although role-specific information cannot be specified in the SAA notation itself, it might be adequate to have the information specified at the level of the system as a whole. The problem with this objection is that it is an appeal to the status quo: Using lookup tables, etc., is precisely what current SAA-based models are forced to do. As we have argued, the data on the role of semantics in human cognition suggest that this approach is not adequate. Among other problems, it is too deliberate: The data on the role of semantics in memory retrieval, analogical mapping, etc., suggest that people use role-based semantic information automatically and without computational cost. Specifying information at multiple levels of representations and postulating routines for moving between the two is computationally costly. Although it is impossible, given the current state of our knowledge, to rule such an account out definitively, it certainly seems more plausible to assume that role-based information is simply a part of relational representations than to assume that the mind is equipped with a complex system of look-up tables. Indeed, the use of lookup tables is perhaps even more awkward than it appears at first blush. Without an explicit representation of role-specific semantics in its representations, an SAA system must use an external meta-system to interpret its SAA representations, decide when the look-up table (another representational system external to SAA) should be accessed, retrieve the relevant role information, translate that information into SAA, and insert it into the original SAA format. Importantly, the external control structure could not, itself, utilize SAA: if it

[3] Naming the relation that specifies the relation of a relation to its roles *consists-of* is, of course, completely arbitrary.

[4] One might argue that this problem could be solved by using binary rather than unary representations of roles. For example, representing *lover* (John) as *lover* (John, *loves* (John, Mary)), thus representing the role, its filler, and the relation to which it belongs. The problem with this approach is that it simply transfers the same problem to a new representation. In SAA binary predicates dictate relations. As a result, the system would still have to specify the roles of this binary predicate and how they relate back to it.

did, then it too would require an external meta-system to keep it's own relation and role information straight. This solution is thus both inelegant and impractical, and in the limit results in a regress, not of nested "consists-of" relations, but of external control structures.

A third objection to our argument concerns the *n*-ary restriction. This problem might be solved simply by postulating that all *n*-place predicates (where *n* is free to vary) be replaced by *m*-place predicates (where *m* is a constant, say, 3). For example, *tall* (*x*) would be recast as *tall* (*x*, -, -), and *taller-than* (*x*, *y*) as *taller-than* (*x*, *y*, -), where "-" denotes a permanently empty "dummy" slot. In this way, *tall* (*x*, -, -) could be mapped to *taller-than* (*x*, *y*, -) by virtue of their both taking three "arguments". One problem with this proposal is that it assumes some procedure for deciding how to recast the predicates. For example, should *short* (*y*) be recast as *shorter-than* (*y*, -, -) (the intuitive recasting) or as *taller-than* (-, *y,* -)? Note that this question must be answered *prior* to discovering the very mapping(s) the recasting is supposed to permit. A second problem with this proposal is that it still leaves the problem of mapping non-identical predicates: How should the system map the arguments of *taller-than* (*x*, *y*, -) to those of *shorter-than* (*x*, *y*, -) without knowing what the roles of each relation "mean"?

Semantically Rich Representations of Relational Roles

We have argued that SAA is at best incomplete as a model of human mental representation. However, this limitation does not by any means imply that we must abandon symbolic accounts of mental representation. Rather, what is needed is a representational system that makes relational roles, their semantics, their bindings to their fillers and their composition into complete relations all explicit. This approach is commonly known as *role-filler binding* (e.g., Halford, et al. 1998).

In a role-filler binding system roles are represented explicitly and bound to their fillers. Relations are composed of linked role-filler bindings. One example is the representational system employed by Hummel and Holyoak's (1997, 2003a) LISA model. LISA uses a hierarchy of distributed and localist codes to represent relational structures. At the bottom, "semantic" units represent objects and roles in a distributed fashion (e.g., for the relation *love* (John, Mary) "John" might be represented as *human*, *adult*, *male*, etc., and Mary by *human*, *adult*, *female*, etc.; similarly, the *lover* and *beloved* roles of the *love* relation would be represented by units capturing their semantic content). At the next level, these distributed representations are connected to localist units representing individual objects and relational roles. Above the object and role units, localist role-binding units (SPs) link object and role units into specific role-filler bindings. At the top of the hierarchy, localist P units link SPs into entire propositions.

In this representation, the long-term binding of roles to their fillers is captured by the conjunctive SP units, thus violating the role-filler independence required for symbolic representation. However, when a proposition becomes active, its role-filler bindings are also represented dynamically by synchrony of firing: SP units in the same proposition fire out of synchrony with one another, causing object and predicate units, along with their corresponding semantics, to fire in synchrony with each other if they are bound together, and out of synchrony with other bound roles and objects. On the semantic units, the result is a collection of mutually desynchronized patterns of activation, one for each role-filler binding. Thus, when a proposition is active, role-filler independence is maintained on its object, predicate and semantic units, with the role-filler bindings carried by the synchrony relations among these units.

This is only one way to represent relational knowledge in a role-filler binding scheme (see also Halford et al, 1998, Smolensky, 1990). As demonstrated by Hummel, Holyoak and their colleagues (see Hummel & Holyoak, 2003b for a review), however, it is a very useful one. LISA has been shown to simulate aspects of memory retrieval, analogical mapping, analogical inference, schema induction, and the relations between them. It also provides a natural account of the capacity limits of human working memory, the effects of brain damage and normal aging on reasoning, the relation between effortless ("reflexive") and more effortful ("reflective") reasoning, and aspects of the perceptual-cognitive interface. It inherits these abilities from its ability to capture both the relational structure and the semantic richness of human mental representations.

Conclusion

Few would claim that people literally think in the predicate calculus. However, many researchers have argued that SAA-based representations serve *at least* as a plausible shorthand for human mental representations. We have argued that SAA notation is *at most* a shorthand for human relational representations—a shorthand that must necessarily leave the messy business of the semantics of relational roles to fundamentally external representations and processes. Inasmuch as the semantics of roles are an important internal component of human mental representation, as we, and others, have argued they are, this fact leaves an important facet of human mental representation necessarily beyond the reach of SAA-based models. The solution to this problem is not to abandon symbolic representations altogether, as proposed by some, but rather to replace SAA with a role-filler binding approach to the representation of relational knowledge in models of cognition. Doing so provides a natural basis for capturing both the symbolic structure of human mental representations and their semantic richness.

Acknowledgements

The authors would like to thank Keith Holyoak, Keith Stenning, Ed Stabler, Tom Wickens, and Charles Travis for

helpful discussion and comments. This work was supported by a grant from the UCLA Academic Senate.

References

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: LEA.

Bassok, M., Wu, L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition*, 23, 354-367.

Doumas, L. A. A., & Hummel, J. E. (in press). Modeling human mental representations: What works, what doesn't, and why. In K. J. Holyoak & R. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press.

Eliasmith, C. & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, 25, 245-286.

Elman, J. L., Bates, E., Johnson, M. H., Karmaloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Inateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Faulkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141-205.

Gentner, D., Ratterman, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25, 524-575.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Brain and Behavioral Sciences*, *21*, 803-864.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich and A. Markman (Eds.). *Cognitive Dynamics: Conceptual Change in Humans and Machines* (pp. 229 - 264). Hillsdale, NJ: Erlbaum.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295-355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427-466.

Hummel, J. E., & Holyoak, K. J. (2003a). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220-263.

Hummel, J. E., & Holyoak, K. J. (2003b). Relational reasoning in a neurally-plausible cognitive architecture: An overview of the LISA project. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 10*, 58-75.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science, 18*, 387-438.

Krawczyk, D. C., Holyoak, K. J., & Hummel, J. E. (in press). Structural constraints and object similarity in analogical mapping and inference. *Thinking and Reasoning*.

Kubose, T. T., Holyoak, K. J., & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of Memory and Language, 47*, 407-435.

Larkey, L. & Love, B. (2003). CAB: Connectionist Analogy Builder. *Cognitive Science* , 27, 781-794.

Landauer, T.K. and Dumais, S. T. (1997) A solution to Plato's problem: TheLatent Semantic Analysis theory of acquisition, induction andrepresentation of knowledge. *Psychological Review*, 104, 211-240.

Manzano, M. (1996). *Extensions of first order logic*. Cambridge: Cambridge University Press.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3), 243-282.

Marr, D. (1982). *Vision*. Freeman: San Francisco.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Ramscar, M. & Yarlett, D. (2003) Semantic Grounding in Models of Analogy: An Environmental Approach. *Cognitive Science, 27*, 41-71.

Ross, B. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *JEP: Learning, Memory, and Cognition*, *13*, 629-639.

Ross, B. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *JEP: Learning, Memory, and Cognition*, *15*, 456-468.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. (Vol. 1). Cambridge, MA: MIT Press.

Salvucci, D. D., & Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: The path mapping theory. *Cognitive Science*, 25, 67-110.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159-216.

Structure Mapping and the Predication of Novel Higher-Order Relations

Leonidas A. A. Doumas (adoumas@psych.ucla.edu)
John E. Hummel (jhummel@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
405 Hilgard Ave.
Los Angeles, CA 90095-1563

Abstract

Relations play a central role in human perception and cognition, but little is known about how relational concepts are acquired and predicated. For example, how do we come to understand that physical force is a higher-order multiplicative relation between mass and acceleration? We report an experiment demonstrating that structure mapping (a.k.a., analogical mapping) plays a key role in the predication of novel higher-order relations. This finding suggests that structure mapping—i.e., the appreciation of analogies—may play a pivotal role in the acquisition and predication of novel relational concepts.

Relational Reasoning

The processing of relations plays a central role in human perception and thought. It permits us to perceive and understand the spatial relations among an object's parts (Hummel, 2000; Hummel & Biederman, 1992; Hummel & Stankewicz, 1996), comprehend arrangements of objects in scenes (see Green & Hummel, 2004, for a review), and comprehend abstract analogies between otherwise very different situations or systems of knowledge (e.g., between the structure of the solar system and the structure of the atom; Gentner, 1983; Gick & Holyoak, 1980, 1983; Holyoak & Thagard, 1995). The power of relational thinking resides in its ability to generate inferences and generalizations that are constrained by the *roles* that elements play, rather than strictly the properties of the elements themselves: The sun is similar to the nucleus of an atom, not because of its literal features, but because of their shared relations to planets and electrons, respectively.

Experience can cause profound changes in the way we process relations. The difference between an expert chess player and a novice, for example, lies in the ability to quickly perceive and reason about the meaningful relations among the pieces on the board (and relations among those relations). Relational learning is central to both the most abstract and uniquely human cognitive abilities (including mathematical and scientific reasoning), and the most "everyday" reasoning using analogies, schemas and rules (Gentner, 1983; Holland, et al., 1986; Hummel & Holyoak, 1997, 2003).

In order to reason explicitly about a relation it is necessary to *predicate* that relation, that is, to represent it as an explicit predicate that takes arguments. Consider an example. In a match-to-sample task, an animal is shown a sample stimulus (e.g., a red square), and two alternatives, one that matches the sample (another red square) and one that does not (e.g., a green square). The animal's task is to indicate which alternative matches the sample. Many animals, including honeybees (Giurfa et al., 2001), can learn to perform this task with simple stimuli such as colors and shapes (see Holyoak & Thagard, 1995, Thompson & Oden, 2000). The computational requirements for performing this task include the ability to explicitly represent values of the relevant feature dimension (e.g., "red" for the dimension "color"), and the ability to remember the value of that dimension in the sample for the purposes of choosing the correct alternative. Despite initial appearances, the task does *not* require the animal to explicitly appreciate that the correct choice item is in any way the "same" as the sample. For example, if color is the relevant dimension, then after the presentation of a red sample, the animal need only maintain a representation of "red" until the choice items appear. The animal need never reflect explicitly on the fact that the sample and the correct choice are the same color (Thompson & Oden, 2000).

However, the task can be generalized to require an explicit appreciation of "sameness." Consider a *relational* match-to-sample task, in which the sample depicts two triangles, alternative A depicts of circle and a diamond, and alternative B depicts two squares. Choosing B as the correct match to the sample requires the reasoner to represent B and the sample in terms of their shared relation (i.e., *same-shape* (x, y)). College students find this comparison trivial, yet only humans and symbol-trained chimpanzees are known to be able to perform this task reliably (Thompson & Oden, 2000). (Fagot, Wasserman & Young, 2001, claim to show relational matching to sample in the baboon, *Papio papio*. However, their data—in particular, the baboons' failure to learn the task when the sample and choice options each contained only two objects—are more consistent with the baboons' responding to stimulus entropy as a holistic perceptual feature, akin to color, rather than *same* as an explicit relation; Hummel & Holyoak, 2003.)

The assumption that people represent the relation *same-shape* in the same way for the squares as for the triangles provides an intuitive account of our ability to perform the relational match to sample, but it begs the question of *why* we see the relation "same shape" in the squares, whereas most other animals only see squares. What are the mental operations that allow us to discover and predicate *same-shape* as an explicit relation that retains its properties over

the sameness of squares to squares and the sameness of triangles to triangles?

The question of how we discover and predicate new relations is central to cognitive science because the kinds of problems a person (or cognitive model) can solve, and the characteristics of its solutions, depend critically on the relations the person (or model) does and does not represent explicitly. Models of human perception and cognition that represent relations explicitly (i.e., as predicates that take arguments) can solve problems far beyond the scope of models that do not represent relations explicitly (e.g., traditional connectionist models, which represent all concepts as simple lists of features; for reviews see Doumas & Hummel, in press; Hummel & Holyoak, 1997, 2003; Marcus, 1998). But to date, all the models that do represent relations are simply given, by the modeler, a vocabulary of relational concepts with which to reason (examples include ACT-R [Anderson & Lebiere, 1998], LISA [Hummel & Holyoak, 1997, 2003], SME [Falhenhainer, Forbus & Gentner, 1989] and, SOAR [Rosenbloom, Newell, & Laird, 1991], among many others). The question of where these concepts come from, and the related question of how we know which relations to predicate in which contexts, is rarely if ever addressed, and the answer to this question is far from well understood. Understanding how the mind comes to represent relations as explicit predicates would contribute substantially to our understanding of the origins of human perception and thinking, and to the development of symbolic thought (Smith, 1989).

Relational Predication

The question of relational predication subsumes at least two related questions: First, how do we recognize and predicate familiar relations for use in novel situations? It is one thing to understand abstract relational notions such as *same-as*, *threatens* or *covaries-with*; it is another to recognize that a relation applies in a given situation and to explicitly predicate it in the service of understanding that situation. Second, how do we discover new relations? For example, what happens in the mind of a child between the time when she does not understand the relation *same-shape* (x, y), and the time when she does? Inasmuch as new relations are learned as combinations of familiar relations, or as familiar relations applied to novel dimensions, the question of relational discovery is clearly related to the question of predication: Especially for adults, discovering new relations may often be a process of discovering which familiar relations apply in a novel situation, and discovering how they are linked together by higher-order relations. Consider, for example, the physics student who is first learning to reason about force as a relation between mass, a basic property of an object, and acceleration, itself a relation between velocity and time. It is this version of the relation discovery question—how do we discover novel higher-order relations among familiar relations—that is the focus of the present paper.

Our ability to appreciate that the relation between the squares in the relational match to sample task is the same as the relation between the triangles—and to choose a pair of squares over a circle and a diamond as the correct match to a pair of triangles on the basis of that relation—illustrates that relations are *invariant* with their arguments (Hummel & Holyoak, 2003): *same-shape* (x, y) is the same relation, regardless of the particular shapes that happen to be bound to x and y at the time. It is precisely this invariance that allows us to appreciate what *same-shape* (triangle1, triangle2) has in common with *same-shape* (square1, square2). As a result of this invariance, *same-shape* ranges over *all possible* shapes, so it is not learnable in terms of the perceptual features of any particular pair of shapes (see Kellman, Burke, & Hummel, 1999). The ability to perform tasks based on such relations—and to discover and predicate them—is therefore fundamentally beyond the reach of any learning algorithm based strictly on the statistical regularities among the elements of the stimuli in its training set—i.e., the vast majority of all theories of learning (see Hummel & Holyoak, 2003).

The problem of relational learning and predication is further complicated by the sheer number of potentially relevant relations present in any given situation. The number of first-order relations among n items increases minimally with $(n^2 - n)/2$ (and this assuming that all relations are commutative, which is not the case for most relations). Worse yet, the number of higher-order relations over these first-order relations is literally unbounded. Any task (e.g., category learning, problem solving, etc.) that calls for the discovery of new higher-order relations is therefore functionally impossible without additional constraints on the selection of which relations to predicate.

Given this, how do people discover and predicate new relations? An important theme that has emerged in the literature on relational reasoning is that *structure mapping* (a.k.a. *analogical mapping)*—the process of finding relational correspondences between the elements of two systems—plays a central role in all forms of relational reasoning (see Hofstadter, 2001; Holyoak & Thagard, 1995). A primary hypothesis motivating the present research is that structure mapping may also play a central role in discovery and predication of new relations. The reason is that structure mapping is driven more by the relational roles that objects play than by the features of the objects themselves. By revealing relational similarities between otherwise different-seeming systems, structure-mapping may bootstrap the discovery of any higher-order relations the two systems have in common. Consistent with this hypothesis, several previous studies have demonstrated that structure mapping bootstraps the induction of abstract relational schemas (e.g., Gick & Holyoak, 1983; Ratterman & Gentner, 1998; Sandhofer & Smith, 2001; Yamauchi & Markman, 2000), and that comparison helps people appreciate what lower-order relations might be relevant to a specific task (Gentner & Namy, 1999; Namy & Gentner, 2002; Yamauchi & Markman, 1998).

The current experiment was designed to investigate whether analogical mapping may also help people to discover the higher-order relations that analogous systems have in common—that is, whether analogical mapping may bootstrap the discovery of novel higher-order relations. If it does, then analogical mapping may not only be a process that depends on the relations we can predicate, but may also be a process that aids us in the predication of new relations.

Experiment

The experiment used a category-learning paradigm to measure relational predication. Categories were defined by an unfamiliar higher-order relation between the elements of exemplars. Each exemplar consisted of drawings of three simple "cells" inside a circular frame (see Figure 1). Within an exemplar, the cells varied in their location in the frame, their shape, the thickness of their membrane, the roundness of their nucleus, and the number of organelles. Categories were defined by a higher-order relation between the cells' membrane thickness and the roundness of their nuclei: In Category A, the thicker a cell's membrane, the rounder its nucleus; in Category B, the thicker the membrane the more elliptical its nucleus. The cells' locations in the frame, shape, and number of organelles varied randomly and were uncorrelated with category membership.

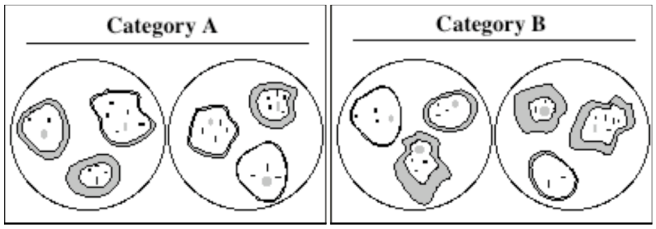


Figure 1. Two stimuli from Category A and two from Category B.

The exemplars were designed to make category learning impossible without discovering the higher-order relation between relative membrane thickness and nucleus roundness. Absolute thickness and roundness were non-predictive of category membership because the thinnest membrane (or least round nucleus) in one exemplar of a category was potentially the thickest (or roundest) in another exemplar of the same category. For the same reason, conjunctions of specific roundnesses and thicknesses were also non-predictive. Every exemplar, regardless of category, had three cells, one of which had a thickest membrane and another of which had a thinnest (with the third in between), so the categories were not learnable in terms of relative (or absolute) membrane thickness. Likewise, every exemplar, regardless of category, had one cell with a more round nucleus than the others and one with a more elliptical nucleus (with the third in between). In other words, the categories were not definable, or learnable, in terms of any basic features or even first-order relations.

For this reason the category structure is unlearnable by any model that codes exemplars in terms of their features (e.g., location, color, width, orientation, etc.) or conjunctions of their features, but cannot explicitly represent relations among those features and relations among relations. Such models constitute the vast majority of mathematical and computational models of category learning (e.g., Krushke, 1992, 2001; Nosofski, 1988; Nosofski & Palmeri 1998), including all connectionist models (see Doumas & Hummel, in press; Hummel & Holyoak, 2003; Marcus, 1998).

By contrast, the categories are learnable in the space of *conjunctions* of *relative* membrane thickness and *relative* nucleus roundness—that is, in terms of a higher-order relation between the first-order relations of relative thickness and relative roundness (which is simply a restatement of the category-defining higher-order relation).

Two groups of subjects were trained to categorize exemplars into the two categories. One group (the *Map* group) performed a mapping task halfway through the category-learning task; the other group (the *No Map* group) did not. Subjects in the No Map condition simply studied a pair of exemplars from the same category (either A or B, counterbalanced); subjects in the Map condition viewed a pair of exemplars from the same category and were asked to indicate which cell in one exemplar corresponded to which in the other and why.

Our predictions were as follows: (1) To the extent that the category-relevant higher-order relation between cells' relative membrane thickness and relative nucleus roundness is unfamiliar to our subjects, categorization performance on the pre-mapping trials ought to be near chance. (2) To the extent that mapping helps subjects to predicate this relation, post-mapping categorization performance of subjects who map correctly in the Map condition should jump abruptly to ceiling (as a result of predicating the category-defining relation), but performance in the No Map condition, and the performance of those who map incorrectly in the Map condition, should remain near chance.

Methods

Participants: 20 UCLA undergraduates participated for course credit.

Materials: Each exemplar consisted of three drawings of simple cells in a circular frame. The cells differed in their shapes, location, membrane thickness, nucleus roundness, and number of organelles (see Figure 1).

Seven membrane thicknesses and seven nucleus roundnesses were used to construct the stimuli, making it possible for the thickest membrane (or roundest nucleus) in one exemplar of a category to be the thinnest (or most elliptical) in another exemplar of the same category, thus making it impossible for subjects to categorize correctly based on absolute thickness or roundness (i.e., it is necessary to respond on the basis of *relative* thickness and roundness between cells in an exemplar).

The locations, shapes, and number of organelles of the cells in an exemplar varied randomly, subject to the constraint that no cells ever overlapped in the frame. Each cell was one of 6 different shapes and contained between 1 and 6 organelles.

We created exemplars used in the pre-mapping, post-mapping, and mapping phase of the experiment (described more fully below) withholding membrane thicknesses 3 and 7 (the thickest), and nucleus roundnesses 1 (least round) and 5 for construction of transfer exemplars. The exemplars used in the transfer phase were created under the constraints described above, with the additional constraints that at least one novel thicknesses and one novel roundness appeared in each exemplar, and each novel thickness and roundness appear in at least three of the six transfer exemplars (see below). The withheld thicknesses and roundnesses consisted of values both within the bounds of the values seen by subjects during the training and test phases of the experiment and values outside those bounds. Thus, transfer trials required subjects to both interpolate and extrapolate learning to new values.

The exemplars used during the mapping phase consisted of two exemplars from the same category placed side by side.

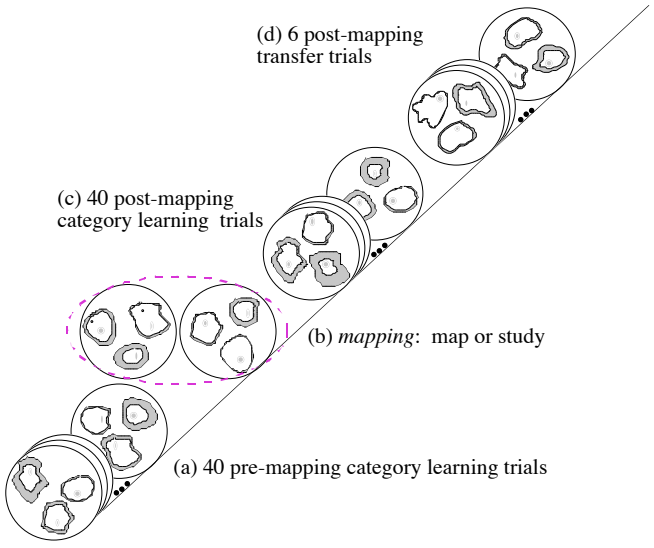


Figure 2. Structure of the experimental procedure. (a) Training phase, 40 trials; (b) mapping phase; (c) test phase, 40 trials; (d) transfer phase, 6 trials.

Procedure: Ten subjects were randomly assigned to each experimental condition. All stimuli were presented on a computer screen. All subjects received 40 pre-mapping training exemplars (20 A's and 20 B's) in a random order (Figure 2a). Their task was to indicate (with a key press) whether each exemplar belonged to Category A or B. Each response was followed by accuracy feedback. Following the initial training phase, subjects were presented with one of the two mapping sets (either two As or two Bs; Figure 2b). Subjects were informed that both exemplars belonged to the same category but they were not told which category they belonged to. Subjects in the No Map condition were instructed to study the mapping set for one minute. Subjects in the Map condition were asked to indicate which cell in the exemplar on the left corresponded to each cell in the exemplar on the right, and to state the reason or reasons for each correspondence. Conditions and mapping sets were fully counter-balanced. All subjects then received 40 post-mapping training trials (20 A's and 20 B's) in a random order (Figure 2c). Responses were followed by accuracy feedback as before. In the final transfer stage of the experiment subjects were presented with the six transfer exemplars (3 A's and 3 B's) in random order and their task was to categorize each (Figure 2d). They received no accuracy feedback during this part of the experiment.

Scoring: All participants were scored for number of correct responses in the pre- and post-mapping trials (maximum 40 correct for each) and the transfer trials (maximum 6 correct). We also recorded the mappings made by participants in the Map condition. At the end of the experiment all subjects were also asked to state the rule(s) they had used to categorize the exemplars.

Results

The results of the experiment were exactly as predicted. An independent-samples t-test showed no main effects for mapping, $t(18) = 1.13$, $p > .25$, on the pre-mapping training trials (mean-proportion-correct$_{MAP}$ = .52, mean-propotion-correct$_{NO\text{-}MAP}$ = .47), which is expected, as the groups received exactly the same treatment prior to mapping. Performance in neither group differed significantly from chance (50% correct).

A second independent-samples t-test was run for performance on the post-mapping trials (Figure 3a). Post-mapping, categorization performance in the Map condition was significantly more accurate (mean-proportion-correct$_{MAP}$ = .77) than in No Map (mean-propotion-correct$_{NO\text{-}MAP}$ = .48), $t(18) = 3.84$, $p < .01$. Accuracy in the No Map group did not differ from chance.

A similar pattern of results obtained on the transfer trials (Figure 3b). Subjects in the map condition performed significantly more accurately (mean-proportion-correct$_{MAP}$ = .83) than those in the no map condition (mean-propotion-correct$_{NO\text{-}MAP}$ = .42), $t(18) = 4.16$, $p < .01$. Performance of the No Map group on the transfer trials did not differ from chance.

The participants' reports of their mappings and rule use also revealed interesting patterns. First, none of the 10 subjects in the No Map group, and 7 of the 10 subjects in the Map group correctly stated the rule defining category membership at the end of the experiment. Second, there was a perfect 1:1 correspondence between subjects who correctly mapped the cells during the mapping phase and those who correctly stated the rule: All and only those subjects who identified the correct mappings were able to

state the category-defining rule at the end of the experiment. All other subjects either missed the relevant dimensions completely or mapped based on absolute membrane width and absolute nucleus roundness, which, as stated previously, were not sufficient for correct categorization.

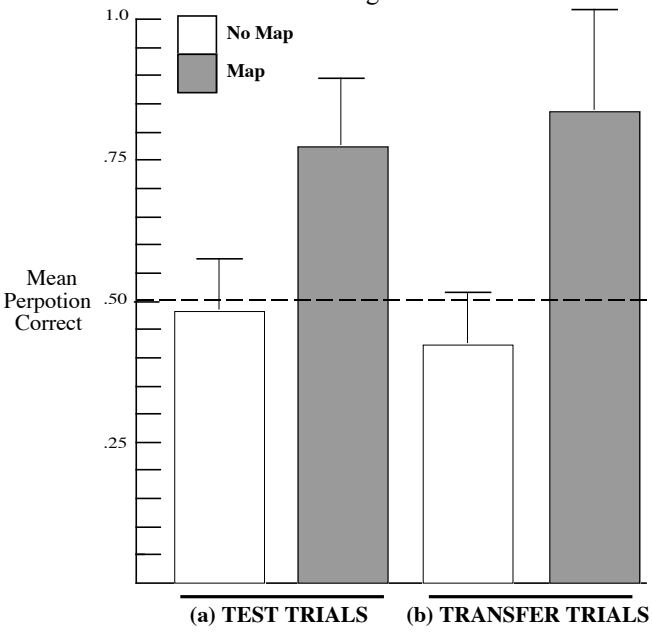


Figure 3. (a) Mean number of correct responses on post-mapping trials as a function of condition. (b) Mean number of correct responses on transfer trials as a function of condition. The dashed line indicates chance.

Discussion

Relations play a central role in human perception and thinking, yet little is known about how relational concepts are acquired and predicated. The problem of relational predication is especially difficult because it is underconstrained. We hypothesized that structure mapping might aid in the discovery and predication of novel higher-order relations.

The results of a category learning experiment support this hypothesis. Subjects who mapped exemplars from the same category onto one another were much better able to learn the novel, category-defining higher-order relation than subjects who did not map. Indeed, performance of the latter group never got above chance.

Additionally, subjects who mapped were able to both interpolate and extrapolate learning to new exemplars with novel stimulus values (i.e., novel membrane thicknesses and nucleus roundnesses) and to verbally state the relational rule that defined category membership. Subjects who did not map were unable to either transfer to new stimuli or to state the category-defining rule. These findings suggest that mapping aids in the predication of novel relations, and that the resulting relations are explicit, in the sense of being available to bind to novel inputs (recall the transfer trials; also, see Hummel & Holyoak, 1997, 2003).

More broadly, the findings reported here suggest that the same cognitive mechanisms that underlie our ability to make analogies—namely, those underlying structure mapping—may also underlie our ability to discover and predicate new relational concepts. If this suggestion is correct, then the evolution of the capacity for generalized structure mapping may well be the "great leap forward" (Newell, 1990) that ultimately gave rise to our capacity for generalized symbolic thought.

Acknowledgements

The authors thank Keith Holyoak and Collin Green for helpful comments and discussions, and Ben Stein and Andrew Pak for tireless work in the lab and for running subjects. This work was supported by a grant from the UCLA Academic Senate.

References

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: LEA.

Beiderman, I, (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, 94, 115-147.

Doumas, L. A. A., & Hummel, J. E. (in press). Approaches to modeling human mental representations: What works, what doesn't, and why. In K. J. Holyoak & R. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. Cambridge: Cambridge University Press.

Fagot, J., Wasserman, E. A, Young, M. E. (2001). Discriminating the relation between relations: The role of entropy in abstract conceptualization by baboons (Papio papio) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes, 27(4)*, 316-328.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, 14, 487-513.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1-38.

Giurfa, M., Zhang, S., Jennett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, *410*, 930-933.

Green, C., & Hummel, J. E. (2004). Relational perception and cognition: Implications for cognitive architecture and the perceptual-cognitive interface. In B. H. Ross (Ed.) *The psychology of learning and motivation*. Vol 44. (pp. 201-223). San Diego: Academic Press.

Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K.J. Holyoak, & B.N. Kokinov

(Eds.), *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Holyoak, K. J., & Hummel, J. E. (2001). Toward an understanding of analogy within a biological symbol system. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT Press.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman (Eds.), *Cognitive Dynamics: Conceptual Change in Humans and Machines*. Hillsdale, NJ: Erlbaum.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99, 480-517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*.

Hummel, J. E., & Stankiewicz, B. J. (1996). Categorical relations in shape perception. *Spatial Vision*, *10*, 201-236.

Kellman, P. J., Burke, T. & Hummel, J. E. (1999). Modeling perceptual learning of abstract invariants. *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 264-269). Mahwah, NJ: Erlbaum.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67, 2797-2822.

Kruschke, J. K. (1992). ALCOVE: An examplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Krushke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45*, 812-863.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*, 243-282.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Children's use of comparison in category learning. *Journal of Experimental Psychology: General*, 131, 5-15.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition and typicality. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *14*, 700-708.

Nosofsky, R. M., & Palmeri, T. J. (1998). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Ratterman, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, 13, 453-478.

Rosenbloom, P. S, Newell , A., & Laird , J. E. (1991). Toward the knowledge level in Soar: The role of the architecture in the use of knowledge. In K. VanLehn (Ed.), *Architectures for intelligence: The Twenty-second Carnegie Mellon Symposium on Cognition*. Hillside, NJ: LEA.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology*, 22, 260-492.

Sandhofer, C. M., & Smith, L. B. (2001). Why children learn color and size words so differently: Evidence from adult's learning of artificial terms. *Journal of Experimental Psychology: General*, 130, 600-617.

Smith, L. B. (1989). From global similarity to kinds of similarity: The construction of dimensions in development. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning*, 146-178, Cambridge: Cambridge University Press.

Spalding, T. L., & Ross, B. H. (1994). Comparison-based learning: Effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1251-1263.

Thompson, R. K. R. & Oden, D. L. (2000). Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. *Cognitive Science. Special Issue: Primate cognition*, 24(3), 363-396.

Yamauchi, T. & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124-148.

Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference and structural alignment. *Memory & Cognition*, 28, 64-78.

A Day in the Life of a Spoken Word

Nicolas Dumay and M. Gareth Gaskell (n.dumay/g.gaskell@psych.york.ac.uk)
Department of Psychology, University of York
York, YO10 5DD, UK

Xiaojia Feng (fengxj@yorku.ca)
Department of Psychology, York University
4700 Keele Street, Toronto, M3J 1P3, Canada

Abstract

Two experiments tracked the emergence of lexical competition effects for newly learnt spoken words (e.g., "cathedruke"). Experiment 1 compared form-only learning with learning in semantically rich sentence contexts. In both cases, although immediate explicit recognition of the novel words was good, lexical competition effects (e.g., "cathedruke-cathedral") emerged only after a delay of at least 24 hours. Experiment 2 evaluated the timecourse of learning in more detail and used embedding (rather than cohort) new competitors (e.g., "shadowks"). Again results showed no evidence of lexicalization immediately after exposure, but clear lexical competition effects after 24 hours. Furthermore, recognition and free recall improved over time. These results are interpreted in terms of a consolidation process that integrates words into the mental lexicon over a relatively protracted period of time.

1. Introduction

Our knowledge about what information is relevant for language acquisition has increased greatly in the last decade. Factors such as statistical properties of the input (Saffran, Aslin, & Newport, 1996) and current lexical knowledge (Dahan & Brent, 1999) have been shown to influence lexical development. However, less is known about exactly how new vocabulary items are integrated into one's mental lexicon, a process called "lexicalization". The main reason for this state of affairs is that studies on word acquisition have typically used only direct measures of learning, such as the performance in familiarity judgment or recollection tasks. Yet, such measures only tell us about the strength of the traces left by exposure, not whether a new lexical entry *per se* has been created.

A critical methodology for addressing the lexicalization issue looks at whether newly learnt words influence how the learner recognizes preexisting words. For models of spoken word recognition, a key feature of a lexical entry is its ability to be evoked when compatible with the input, and to compete with similar-sounding entities for identification (e.g., McClelland & Elman, 1986). Therefore, a strong test of whether a speech sequence has been lexicalized is whether it engages in lexical competition, and thereby affects the activity within the mental lexicon.

In a recent study (Gaskell & Dumay, 2003), we began to explore how and when newly learnt spoken words become involved in the lexical competition process, or in our terms, produce a "lexical footprint". Adults were familiarized with made-up words that overlapped strongly with existing words (such as "cathedruke" for "cathedral"), through repeated presentation in a phoneme-monitoring task. In one experiment, good explicit knowledge of the novel words was obtained after only one training session (i.e., 12 presentations of each item), whereas the inhibitory influence of these new competitors on the identification of existing words in a lexical decision task (LD) required three (successive) days of exposure to emerge.

In another experiment, we disentangled the roles of time and level of exposure in the lexicalization process, using a single training session at a high exposure rate (i.e., 36 presentations). We also swapped LD with a more implicit test of lexical activity, the pause detection task. Here, participants made speeded decisions as to whether a short silence was present towards the offset of the existing words (e.g., "cathedr_al"). As Mattys and Clark (2002) showed, pause detection latencies are positively correlated with the amount of lexical activity elicited by the portion of speech preceding the pause. They hypothesized that the activation of lexical candidates involves the use of processing resources that would otherwise be allocated to pause detection. Our experiment showed good explicit recognition of the novel items right after exposure. In contrast, an increase in lexical activity as indexed by longer pause detection latencies when a new competitor had been learnt was not immediately observed, but had emerged when re-tested a week later.

So, in contrast to phonological storage, lexicalization is apparently not instantaneous and in fact may require a substantial amount of time, possibly to allow the consolidation of episodic traces (cf. O'Reilly & Norman, 2002). Nonetheless, on the basis of the above findings, it is not possible to tell how long it takes after a sufficient level of exposure has been reached for a newly learnt word to be lexically operational.

Furthermore, in Gaskell and Dumay (2003) participants had to learn only the sound-form of the novel words in quite an artificial situation, i.e., phoneme monitoring. Therefore, whether these data give us a good picture of what happens in more normal circumstances when semantic and thematic information are usually available must be addressed. In particular, the delay observed in the emergence of lexical

footprint could well result from the relatively impoverished conditions in which the novel words were acquired. Word learning, as measured by recognition and recall, is often improved when a meaning is available to attach to the novel phonological form (e.g., Rueckl & Olds, 1993; Whittlesea & Cantwell, 1987). On these grounds, linking the form of the novel words to some semantic representations during encoding may give rise to a faster lexicalization and, potentially, to a "deeper" lexical footprint.

Finally, so far the onset-matched competitors that have been used to test for the emergence of lexical competition were cohort competitors, i.e., novel and existing items that mismatch towards their offset. Therefore, we do not know whether these effects can be extended to a more general view of lexical competition encompassing all words that overlap to any degree (cf. McQueen, Norris, & Cutler, 1994). The following experiments address these issues.

2. Experiment 1

Experiment 1 examined whether providing some semantic information along with the form of the novel words during exposure would result in a deeper lexical footprint and/or the faster emergence of this effect. On two successive days of learning, novel words (e.g., "cathedruke" for "cathedral") were heard 12 times either in isolation, as carriers in a phoneme monitoring task, or in a sentential context during a semantic verification task. Here, they were associated with the name of a conceptual category (e.g., "vegetable"). The effect of exposure to these new cohort competitors on identification of the base words was evaluated immediately, 24 hours later (before the second exposure) and after a week, using a LD task. In addition, whether the novel words learnt under semantic exposure had acquired a meaning was tested in two ways. First, during the LD task, we also presented each novel word followed directly by their category name, and measured the extent to which the former could speed up responses to the latter (cf. Dagenbach, Horst, & Carr, 1990). Second, we looked at how much the novel words would elicit production of a word related to the meaning of the category name in a free association task.

2.1. Method

2.1.1. Participants. Thirty native British English speakers with no known auditory or language impairment were tested. They were students at the University of York (UK) or lived in the surrounding area, and were all paid for their participation.

2.1.2. Materials and Design. The key materials contained 12 bisyllabic and 24 trisyllabic item triplets (based on Gaskell & Dumay, 2003, Experiment 2). Each triplet included a base word, such as "cathedral", and two nonwords, such as "cathedruke" and "cathedruce". The nonwords diverged from the base word at the final vowel and from each other at the final consonant or consonant cluster. One nonword (e.g., "cathedruke") was presented as novel word, whereas the other one was used as alternative choice in a recognition test. Base words were monomorphemic nouns that ranged in frequency between 2 and 19 occurrences per million and had their uniqueness point (UP) located at or before the final vowel. Hence, if exposure led to lexicalization of the novel word, the latter was expected to become the main competitor of the base word, shifting its UP towards its offset.

For the semantic exposure phase, each novel word was assigned a meaning, based on a conceptual category unrelated the base word (cf. Battig & Montague, 1969). For example, "cathedruke" was associated with "vegetable". Two sentences in which each novel word appeared were then constructed. One explicitly conveyed information about the category membership of the novel word, such as "*A cathedruke is a variety of vegetable*"; the other provided a more general semantic context, such as "*The cook served the boiled cathedruke with a steak and baked potatoes*".

The test items were divided into three groups, as were the participants. During exposure, a given group of participants heard 12 novel words in a phoneme monitoring task and 12 others in a semantic verification task, the items being assigned to a different exposure condition (phonological, semantic or unexposed) across the three alternative versions of the experiment. Participants were presented with all base words during the LD lexicalization test. Thus, for any participant, new competitors were potentially acquired for 2/3 of the existing words, and overall each item was equally represented at the three levels of the factor "exposure".

Base words, novel words, alternative nonwords, category names and sentences were produced in a soundproof booth by a male native speaker of British English, recorded onto CD, and stored as separate files using CoolEdit.

2.1.3. Procedure. On day 1, participants were exposed to the novel words through the phoneme monitoring and semantic verification tasks, with task order counterbalanced across participants. Next, they were tested for lexicalization effects in a LD task, followed by a two-alternative forced choice (2-AFC) recognition test which assessed explicit knowledge of the novel words, and, finally, a free association task. On day 2, participants performed the LD task, the 2-AFC recognition test and the free association task before a second exposure phase took place. On day 8, the procedure was the same as on day 2 except that there was no further exposure.

The phoneme monitoring component of the *exposure phase* involved 12 novel words and consisted of 12 blocks in which each novel word occurred once. A target phoneme was specified for each block, and in all 6 phonemes were used (/n/, /d/, /t/, /s/, /p/ and /m/). Participants had to make speeded decisions as to whether the target was present or absent in the word, by pressing one of two buttons. The semantic verification component used 12 other novel words, each presented 6 times by way of their "category membership" sentence, and 6 times by way of their "semantic context" sentence. On each trial of a given block, participants had to make a yes/no judgment about the

meaning of the novel word. In all 6 questions were used, asking whether the novel word referred to something that was (1) man-made, (2) alive, (3) edible, (4) audible, (5) touchable, or (6) liked by the participant.

The *lexicalization test* required making timed LD to all the base words, the novel words and their associated category names intermixed with a large set of fillers (i.e., 64 words, 102 nonwords). The order of stimulus presentation was the same for each participant but varied every day. It was pseudorandomized such that each base word (e.g., "cathedral") occurred at least 20 trials before its related novel word (e.g., "cathedruke"), which was then immediately followed by the associated category name (e.g., "vegetable"). The proportion of semantically related pairs, i.e., a novel word followed by its category name, was 4.4%. Participants were instructed to press "yes" only to the existing real words, and had 3 s. from stimulus onset to respond. The inter-trial interval was 1 s. The LD latencies to the base words allowed us to estimate the amount of lexical competition induced by prior exposure to the novel words during the training phase. The LD latencies to the category names allowed an estimate of the extent to which these words were semantically associated with the immediately preceding novel word (which could act as a prime).

In the *2-AFC recognition test*, novel words and alternative nonwords were presented in pairs (e.g., "cathedruke-cathedruce"), and participants had to press a button to indicate the item they had to learn. The acoustic exemplar of the newly learnt words was the one presented at exposure.

Finally, in the *free association task*, only the novel words presented during the semantic exposure phase were played. After each item, participants had to write down the first word that came to mind. This gave us a second measure of how strongly the novel word was linked to the category name or its meaning.

2.2. Results and Discussion

Table 1. Top: Correct response rate in 2-AFC recognition. Bottom: Response probability in the free association task.

	Day		
	1	2	8
2-AFC recognition			
Phonological exposure	93.6	91.9	97.5
Semantic exposure	86.4	91.7	95.8
Free association			
Novel word meaning	.30	.31	.44
Base word	.38	.47	.38
Base word meaning	.08	.06	.06
Other	.24	.16	.11

Performance in the *2-AFC recognition test* was good, with a rate of correct responses of at least 90% on each session (see Table 1). Analyses of variance (ANOVAs) showed an effect of day ($F1(2,58) = 14.7$, $p < .01$; $F2(2,66) = 12.9$, $p < .01$), an effect of exposure, though marginally significant by participant ($F1(1,29) = 3.9$, $p < .06$; $F2(1,33) = 6.6$, $p < .05$), and a day x exposure interaction ($F1(2,58) = 5.1$, $p < .01$; $F2(2,66) = 5.6$, $p < .01$). As planned comparisons revealed, the performance was better on day 8 than on days 1 and 2 ($ps < .01$), which did not differ from each other. More interestingly, only on day 1 was the performance better for phonological than semantic exposure ($ps < .01$).

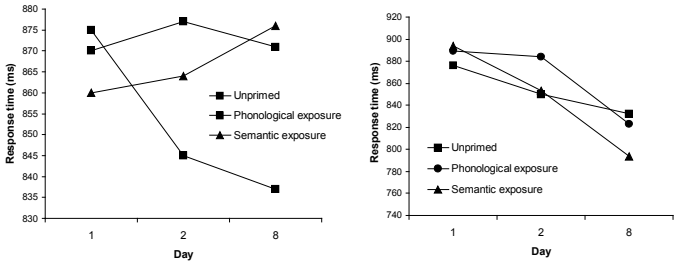


Figure 1. (Exp. 1) Left: Mean lexical decision latency to the base word. Right: Mean lexical decision latency to the category name.

In the *LD lexicalization test* (see Figure 1), latencies to the base words revealed an interaction between day and exposure ($F1(4,108) = 2.6$, $p < .05$; $F2(4,132) = 3.0$, $p < .05$). Here, the important thing was to assess the occurrence of reliable priming effects. Planned comparisons examined the difference between the unprimed condition and both the phonological and semantic conditions. No sign of delayed recognition caused by competition with the novel words was found right after exposure. However, 24 hours later, a clear inhibitory effect had emerged for the novel words trained phonologically ($F1(1,28) = 5.0$, $p < .05$; $F2(1,33) = 7.9$, $p < .01$), but still no significant effect was found in the semantic encoding condition. Finally, on day 8, both phonologically and semantically trained novel words induced inhibition of the base word recognition ($F1(1,28) = 4.1$, $p = .052$; $F2(1,33) = 9.3$, $p < .01$; $F1(1,28) = 6.8$, $p < .05$; $F2(1,33) = 7.4$, $p < .05$). Analyses of errors (2.8%) revealed no significant effect or interaction.

LD latencies to the *category names* also showed an interaction between day and exposure to the preceding novel word (i.e., untrained, phonologically trained vs. semantically trained), although marginally significant by participant ($F1(4,108) = 2.1$, $p = .081$; $F2(4,132) = 3.7$, $p < .01$). Planned comparisons revealed no effect of exposure on day 1. On day 2, an inhibitory effect was found unexpectedly for the phonological condition ($F1(1,28) = 4.3$, $p < .05$; $F2(1,33) = 8.3$, $p < .01$), whereas there no effect for the semantic condition. On day 8, the inhibitory effect in the phonological condition had disappeared, and a facilitatory effect only significant by item had emerged for the semantically trained novel words ($F1(1,28) = 2.3$, $p < .15$; $F2(1,33) = 6.8$, $p < .05$). Analyses of errors (2.2%) revealed no significant effect or interaction.

Responses in the *free association task* were classified using the taxonomy presented in Table 1. Base words and words related to the meaning of the novel words represented the majority of the responses (overall 76%). More interestingly, response probability showed an interaction

between day and response type ($F1(8,232) = 5.3$, $p < .01$; $F2(8,264) = 14.7$, $p < .01$). From day 1 to day 2, the probability of producing the base word increased with no parallel reduction in that of producing the meaning of the novel word. By contrast, from day 2 to day 8, there was an increase in the probability of producing the meaning of the novel word, clearly to the detriment of the probability of producing the base word.

Taken together, these results suggest that exposure to a novel word in a meaningful semantic context does not result in faster or deeper lexicalization compared to simple exposure to its phonological form. On day 2, only the competitors learnt on the basis of just their sound-form were able to delay recognition of the existing words, and on day 8, the two conditions of exposure did not differ in terms their lexical footprint effects. Interestingly, the emergence of competition effects on day 8 for the novel words learnt through semantic exposure coincided with a significant change in the ability of these words to prime their related category name, both in the semantically primed LD and in the free association task.

3. Experiment 2

The finding of a lexical footprint effect in the form-only condition after only 24 hours and 12 exposures in Experiment 1 stands in contrast with the late emergence of this effect after three days of exposure under similar conditions in Gaskell and Dumay's (2003) Experiment 2. This new result suggests that lexicalization may take place during the first 24 hours following exposure. To gain further evidence that this really is the case, the present experiment examined more closely the timecourse of lexicalization for phonologically trained novel words using another paradigm than LD as lexicalization test: pause detection. Contrary to LD, this paradigm provides a measure of lexical activity without requiring participants to make any judgment about the lexical properties of the input. Examination of the lexical footprint effect induced by a single massed exposure phase was performed at three time points: immediately after exposure, 24 hours later and a week later. To test whether the lexical footprint would generalize to the level of competition for segmentation, embedding rather than cohort competitors (e.g., "shadowks") were used.

3.1. Method

3.1.1. Participants. Thirty-two native British English speakers with no known auditory or language impairment were tested. They were all students at the University of York (UK), and none had taken part in Experiment 1. They received course credits or were paid for their participation.

3.1.2. Materials and Design. The key materials consisted of 72 bisyllabic item triplets. Each included a base word ending in an unreduced vowel, such as "shadow", and two nonwords such as "shadowks" and "shadowkt", derived from the base word by adding a consonant cluster and which differed from each other in one of the final consonants. As in Experiment 1, one nonword (e.g., "shadowks") was presented as novel word, whereas the other one was used as alternative choice in a recognition test.

Base words were stress-initial morphologically simple nouns that ranged in frequency between 0 and 403 occurrences per million and had their UP located before or at the final vowel. Here, in contrast to Experiment 1 which used cohort competitors, the novel word, if lexicalized, would be the only longer (embedding) competitor of the base word. To have a better chance to index lexicalization of this longer word, base words were therefore not presented in isolation during the pause detection paradigm but in longer carriers, such as "shadowk" or "shadow_k", derived from the new competitor itself (e.g., "shadowks").

In addition, 12 other novel words along with their alternative nonwords were devised, such as "trogist" and "trogisk". They were used as fillers to increase the amount of materials to be learnt, and potentially enhance the sensitivity of the 2-AFC recognition test.

All speech materials were produced by the same speaker and acquired using the same procedure as for Experiment 1.

The test items were divided into four groups, as were the participants. In the lexicalization test (i.e., pause detection) half of the carriers contained a short silence (e.g., "shadow_k"), whereas the other half did not, and within each of these groups, half of the items had potentially a longer competitor as a result of exposure, and the other did not. Four versions of the experiment allowed each item to be equally represented in the four (exposure x pause occurrence) subcells of the design.

3.1.3. Procedure. On day 1 participants were exposed to the novel words through a phoneme monitoring task. Then, the immediate effect of exposure on lexical activity was assessed using the pause detection paradigm. Finally, explicit knowledge of the novel words was examined in a free recall task and a 2-AFC recognition test. The effect of exposure on lexical activity and explicit knowledge of the novel words were re-tested on two subsequent occasions: 24 hours after exposure, and one week later. On each occasion, the pause detection task was administered first, followed by the free recall task and the 2-AFC recognition test.

The *exposure phase* was similar to the phoneme monitoring component of Experiment 1. Here, the 36 test novel words and the 12 lexically unrelated fillers were involved. Each of them was presented 36 times over 12 blocks of trials. The 6 target phonemes were /n/, /d/, /k/, /l/, /t/ and /s/.

The *lexicalization test* used the pause detection task. On each trial, participants had to decide by pressing one of two buttons whether a short silence (of 200 ms) was present in any location within a bisyllabic spoken item. On the pause-present trials, base word carriers had the silence inserted just before the final consonant (e.g., "shadow_k"). Fillers were 144 bisyllabic words ending in a consonant or consonant cluster, half of which contained a pause. The pause was inserted just before or after the first or second vowel.

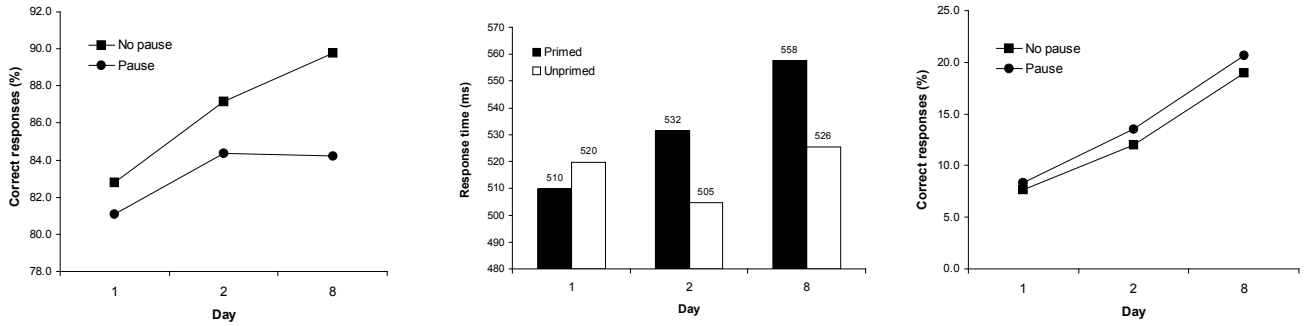


Figure 2. (Exp. 2) Left: Correct response rates in 2-AFC recognition. Middle: Mean pause detection latency (across pause-present and pause absent trials) as a function of day and exposure. Right: Correct response rates in free recall.

In the *free recall task*, participants had 3 min. to recall orally as many novel words as they could from the exposure phase. Finally, the *2-AFC recognition test*, similar to that of Experiment 1, involved all the 48 novel words presented during exposure, along with their choice (e.g., "shadowks-shadowkt").

3.2. Results and Discussion

Performance in the *2-AFC recognition test* was good, with a rate of correct responses of at least 80% on each session (see Figure 2). ANOVAs taking into account day and whether the item had been disrupted by a pause during the lexicalization test revealed that the main effects were significant (day: $F1(2,56) = 6.1$, $p < .01$; $F2(2,136) = 6.2$, $p < .01$; pause: $F1(1,28) = 6.9$, $p < .05$; $F2(1,68) = 7.9$, $p < .01$), but did not interact with one another (*Fs* close to 1). As planned comparisons showed, performance increased from day 1 to day 2 (*p*s < .05), but not from day 2 to day 8.

In the (pause detection) *lexicalization test*, latencies revealed a clear-cut interaction between day and exposure ($F1(2,56) = 6.4$, $p < .01$; $F2(2,114) = 6.9$, $p < .01$). On day 1, the immediate effect of exposure to a novel competitor was to speed-up the pause detection performance, although this effect was only marginally significant by participant ($F1(1,28) = 4.0$, $p < .06$; $F2(1,57) = 1.9$, $p < .2$). In contrast, 24 hours after exposure as well as one week later, the performance on the carriers for which a longer competitor had been learnt was clearly inhibited ($F1(1,28) = 5.7$, $p < .05$; $F2(1,57) = 7.0$, $p = .01$; $F1(1,28) = 9.0$, $p < .01$; $F2(1,57) = 14.0$, $p < .01$). There was no effect of exposure or interaction involving exposure and day on errors (7.4%).

In the *free recall task*, an ANOVA with day and presence or absence of a pause during the lexicalization test only revealed a significant effect of day ($F1(2,56) = 27.8$, $p < .01$; $F2(2,136) = 64.5$, $p < .01$), with better performance on day 2 than on day 1 ($F1(1,28) = 18.7$, $p < .01$; $F2(1,68) = 21.7$, $p < .01$), and better performance on day 8 than on day 2 ($F1(1,28) = 21.3$, $p < .01$; $F2(1,68) = 51.1$, $p < .01$).

On the basis of these results, it thus seems that following a sufficient amount of exposure, lexicalization of the novel word occurs within the next 24 hours, but not immediately. Whereas, on day 1, pause detection was facilitated by prior exposure to a new longer competitor of the base word, on day 2 (as on day 8), there was clear evidence that the new competitor was now contributing to lexical activity. Interestingly, the performance in direct recognition and free recall gradually increased over time.

4. General Discussion

The two experiments reported above allow us to make substantial progress in understanding the full range of factors involved in lexicalization of novel words. Gaskell and Dumay (2003) showed that when words are learned on the basis of only their phonological form, there is a delay associated with their engagement in lexical competition. Experiment 1 looked at whether this delay was eliminated when a richer linguistic context was available during learning. We found no evidence of any earlier or deeper lexicalization using a richer learning environment; if anything, the meaning and sentential context available at encoding led to an increased delay in lexicalization. This result suggests that exposure to a phonological form is both necessary and sufficient for normal engagement in lexical competition, supporting models of language acquisition that have a similar focus on phonological form (e.g., Saffran, Aslin, & Newport, 1996).

Experiment 1 also examined another hallmark of lexical processing: semantic/associative priming. The results suggest that this aspect of lexicalization emerges hand-in-hand with engagement in lexical competition. As for the lexical footprint in the semantic condition, a significant priming effect was observed on day 8, but not at the two preceding test points. We should be careful in interpreting this effect, since the associate of the novel item was repeatedly presented during the exposure session. It is possible that this exposure induced a repetition priming effect instead of, or in addition to, the associative facilitation caused by the pairing of novel items and their associates in LD (e.g., "cathedruke-vegetable). Yet, this account would predict that priming should be just as apparent on days 1 and 2 (cf. Tenpenny, 1995), whereas no such effects were found. Thus, the data do seem to be best

explained in terms of the emergence of a lexical link between the novel items and their associated superordinates. This link appears to rely on the establishment of a lexical entry capable of engaging in competition rather than simply a phonological trace.

Experiment 2 widened the domain of reference for our lexical footprint test. Previously we had employed standard "cohort" competitors, in which the novel and existing items mismatch towards the end of the word. In Experiment 2 the novel items had no segmental mismatch with the existing items, but instead they were embedding competitors (e.g., "shadowks"). This experiment marks the beginning of an extension of our research to lexical competition at the level of lexical segmentation. These items appeared to behave in a very similar way to standard cohort competitors, strengthening the general conceptualization of lexical competition as involving lexical items with any degree of overlap (cf. McQueen et al., 1994).

Experiment 2 had the further advantage of involving a larger set of stimuli with more sensitive measures of explicit recall and recognition performance. The explicit measures demonstrate that even in the absence of further exposure to the novel sequences, recall and recognition performance improves. One potential explanation of this finding is that the processes that operate to engage the novel representations in lexical competition also refine or focus the phonological representations. This interpretation has some support from developmental studies suggesting that well-established lexical representations are more clearly specified in terms of phonological form than newly learnt ones (Stager & Werker, 1997; Swingley & Aslin, 2000).

Perhaps the most conspicuous finding relates to the timecourse of lexicalization. In the phonological condition of Experiment 1, and more crucially in Experiment 2, we found a clear profile of lexical competition effects across the three testing occasions. Immediately after learning, there was no evidence that lexicalization had emerged, as defined by engagement in lexical competition. However, without further exposure, this lexical competition effect was observed 24 hours later, and was essentially unchanged by day 8. We can therefore narrow down the critical time period for emergence of lexical competition to somewhere between 1 and 24 hours after exposure. This suggests that under normal circumstances, lexicalization will not be hurried. This profile of learning fits in with the idea that engagement in lexical competition requires the new information to be interleaved with existing representation as is the case for distributed connectionist networks (O'Reilly & Norman, 2002). Our current research effort is focused on whether lexicalization is reliant on the kind of memory consolidation thought to occur during sleep (Walker, in press).

Acknowledgments

This research was supported by a grant from the UK Medical Research Council (G0000071) to Gareth Gaskell, and Experiment 1 was part of Xiaojia Feng's MSc. thesis. We thank Graham Hitch for valuable comments and discussions.

References

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, **80,** 1-46.

Dagenbach, D., Horst, S., & Carr, T. H. (1990). Adding new information to semantic memory: How much learning is enough to produce automatic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16,** 581-591.

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, **128,** 165-185.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, **89,** 105-132.

Mattys, S. L., & Clark, J. H. (2002). Lexical activity in speech processing: evidence from pause detection. *Journal of Memory and Language*, **47,** 343-359.

McClelland, J. L., & Elman, J. L. (1986). The Trace model of speech perception. *Cognitive Psychology*, **18,** 1-86.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: spotting words in other words. *Journal of Experimental Psychology: Learning Memory and Cognition*, **20**, 621-638.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, **6,** 505-510.

Rueckl, J. G., & Olds, E. M. (1993). When pseudowords acquire meaning: The effect of semantic associations on pseudoword repetition priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19,** 515-527.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, **274**, 1926-1928.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, **388**, 381-382.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, **76**, 147-166.

Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review,* **2**, 339-363.

Walker, M. P. (in press). A refined model of sleep and the time course of memory formation. *Behavioral and Brain Sciences*.

Whittlesea, B., & Cantwell, A. (1987). Enduring influence of the purpose of experiences: Encoding-retrieval interactions in word and pseudoword identification. *Memory and Cognition*, **15,** 465-472.

Bridging computational, formal and psycholinguistic approaches to language

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
se37@cornell.edu

Zach Solan, David Horn, Eytan Ruppin
Faculty of Exact Sciences
Tel Aviv University
Tel Aviv, Israel 69978
{*zsolan,horn,ruppin*}*@post.tau.ac.il*

Abstract

We compare our model of unsupervised learning of linguistic structures, ADIOS [1, 2, 3], to some recent work in computational linguistics and in grammar theory. Our approach resembles the Construction Grammar in its general philosophy (e.g., in its reliance on structural generalizations rather than on syntax projected by the lexicon, as in the current generative theories), and the Tree Adjoining Grammar in its computational characteristics (e.g., in its apparent affinity with Mildly Context Sensitive Languages). The representations learned by our algorithm are truly emergent from the (unannotated) corpus data, whereas those found in published works on cognitive and construction grammars and on TAGs are hand-tailored. Thus, our results complement and extend both the computational and the more linguistically oriented research into language acquisition. We conclude by suggesting how empirical and formal study of language can be best integrated.

The empirical problem of language acquisition

The acquisition of language by children — a largely unsupervised, amazingly fast and almost invariably successful learning stint — has long been the envy of natural language engineers [4, 5, 6] and a daunting enigma for cognitive scientists [7, 8]. Computational models of language acquisition or "grammar induction" are usually divided into two categories, depending on whether they subscribe to the classical generative theory of syntax, or invoke "general-purpose" statistical learning mechanisms. We believe that polarization between classical and statistical approaches to syntax hampers the integration of the stronger aspects of each method into a common powerful framework. On the one hand, the statistical approach is geared to take advantage of the considerable progress made to date in the areas of distributed representation, probabilistic learning, and "connectionist" modeling, yet generic connectionist architectures are ill-suited to the abstraction and processing of symbolic information. On the other hand, classical rule-based systems excel in just those tasks, yet are brittle and difficult to train.

We are developing an approach to the acquisition of distributional information from raw input (e.g., transcribed speech corpora) that also supports the distillation of structural regularities comparable to those captured by Context Sensitive Grammars out of the accrued statistical knowledge. In thinking about such regularities, we adopt Langacker's notion of grammar as "simply an inventory of linguistic units" ([9], p.63). To detect potentially useful units, we identify and process partially redundant sentences that share the same word sequences. We note that the detection of paradigmatic variation within a slot in a set of otherwise identical aligned sequences (syntagms) is the basis for the classical distributional theory of language [10], as well as for some modern works [11]. Likewise, the *pattern* — the syntagm and the *equivalence class* of complementary-distribution symbols that may appear in its open slot — is the main representational building block of our system, ADIOS (for Automatic DIstillation Of Structure).

Our goal in the present paper is to help bridge statistical and formal approaches to language [12] by placing our work on the unsupervised learning of structure in the context of current research in grammar acquisition in computational linguistics, and at the same time to link it to certain formal theories of grammar. Consequently, the following sections outline the main computational principles behind the ADIOS model, and compare these to select approaches from computational and formal linguistics. The algorithmic details of our approach and accounts of its learning from CHILDES corpora and performance in various tests appear elsewhere [1, 2, 3]. In this paper, we chose to exert a tight control over the target language by using a context-free grammar (Figure 1) to generate the learning and testing corpora.

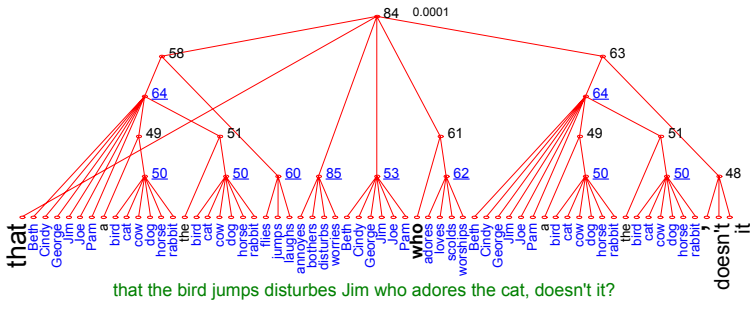
```
S:    P100 | P101 | P102 ;
P100: P1  P2  P3  P4;
P101: P18 P6  P7;
P102: P8  P2  P9  P6  P10  P2;
P2:   P11 P12 | P13;
P22:  P11 P12;
P11:  the | a;
P12:  cat | dog | cow | bird | rabbit | horse;
P13:  P14 P32 | P14;
P14:  Joe | Beth | Jim | Cindy | Pam | George;
P15:  P14 and P14 P36 | P14 P14 and P14 P36;
P16:  Beth | Pam | Cindy;
P3:   P18 and P19;
P32:  who P17 P22 | who P17 P14;
P4:   , don't they ?;
P35: believes | thinks;
P36: believe | think;
P19: P2  P20;
P18: meows | barks;
P20: laughs | jumps | flies;
P9:  is easy | is tough | is eager;
P7:  is easy | is though;
P6:  to please | to read;
P8:  that;
P10: annoys | worries | disturbs | bothers;
P17: scolds | loves | adores | worships;
P18: P14 P35 that P18 | P14 P35 that;
P5:  P16 P35 that P18;
P1:  P15 that P18;
```

Figure 1: the context free grammar used to generate the corpora for the acquisition tests described here.

The principles behind the ADIOS algorithm

The representational power of ADIOS and its capacity for unsupervised learning rest on three principles: (1) probabilistic inference of pattern significance, (2) context-sensitive generalization, and (3) recursive construction of complex patterns. Each of these is described briefly below.

Probabilistic inference of pattern significance. ADIOS represents a corpus of sentences as an initially highly redundant directed graph, in which the vertices are the lexicon entries and the paths correspond, prior to running the algorithm, to corpus sentences. The graph can be informally visualized as a tangle of strands that are partially segregated into *bundles*.



P84 → "that" P58 P63
E63 → E64 P48
E64 → "Beth" | "Cindy" | "George" | "Jim" | "Joe" | "Pam" | P49 | P51
P48 → "," "doesn't" "it"
P51 → "the" E50
P49 → "a" E50
E50 → "bird" | "cat" | "cow" | "dog" | "horse" | "rabbit"
P61 → "who" E62
E62 → "adores" | "loves" | "scolds" | "worships"
E53 → "Beth" | "Cindy" | "George" | "Jim" | "Joe" | "Pam"
E85 → "annoyes" | "bothers" | "disturbes" | "worries"
P58 → E60 E64
E60 → "flies" | "jumps" | "laughs"

Figure 2: *Left:* a pattern (presented in a tree form), capturing a long range dependency (equivalence class labels are underscored). This and other examples here were distilled from a 400-sentence corpus generated by the grammar of Figure 1. *Right:* the same pattern recast as a set of rewriting rules that can be seen as a Context Free Grammar fragment.

Each of these consists of some strands clumped together; a bundle is formed when two or more strands join together and run in parallel, and is dissolved when more strands leave the bundle than stay in. In a given corpus, there will be many bundles, with each strand (sentence) possibly participating in several. Our algorithm, described in detail elsewhere [3],[1] identifies significant bundles iteratively, using a context-sensitive probabilistic criterion defined in terms of local flow quantities in the graph. The outcome is a set of patterns, each of which is an abstraction of a bundle of sentences that are identical up to variation in one place, where one of several symbols (the members of the equivalence class associated with the pattern) may appear (Figure 2). This representation balances high compression (small size of the pattern lexicon) against good generalization (the ability to generate new grammatical sentences from the acquired patterns).

Context sensitivity of patterns. Because an equivalence class is only defined in the context specified by its parent pattern, the generalization afforded by a set of patterns is inherently safer than in approaches that posit globally valid categories ("parts of speech") and rules ("grammar"). The reliance of ADIOS on many context-sensitive patterns rather than on traditional rules can be compared to the Construction Grammar idea (discussed later), and is in line with Langacker's conception of grammar as a collection of "patterns of all intermediate degrees of generality" ([9], p.46).

Hierarchical structure of patterns. The ADIOS graph is rewired every time a new pattern is detected, so that a bundle of strings subsumed by it is represented by a single new edge. Following the rewiring, which is context-specific, potentially far-apart symbols that used to straddle the newly abstracted pattern become close neighbors. Patterns thus become hierarchically structured in that their elements may be either terminals (i.e., fully specified strings) or other patterns. The ability of new patterns and equivalence classes to incorporate those added previously leads to the emergence of recursively structured units that support generalization (by opening paths that do not exist in the original corpus). Moreover, patterns may refer to themselves, which opens the door for true recursion (Figure 3, right; automatic detection of recursion is not currently implemented).

[1]The relevant publications can be found online at http://kybele.psych.cornell.edu/∼edelman/archive.html.

Two experiments in grammar induction

The results outlined next focus on the power of the ADIOS algorithm, which we assessed by examining the (so-called "weak") generativity of the representations it learns.

Experiment 1. In the first of the two studies described here, we trained ADIOS on 400 sentences produced by the context free grammar shown in Figure 1. We then compared a corpus C_{target} of $3,607,240$ sentences generated by this CFG (with up to three levels of recursion) with a corpus $C_{learned}$ of $1,916,061$ sentences generated by the patterns that had been learned by ADIOS from the 400-sentence training set. In both cases the sentences were generated randomly in batches of size $1.5{\cdot}10^7$ and merged until convergence, defined as 95% overlap between new and existing data. With these data, we obtained precision of 97%, with a recall value of 53% (as customary in computational linguistics, we define recall as the proportion of C_{target} sentences appearing in $C_{learned}$, and precision as the proportion of $C_{learned}$ appearing in C_{target}). In this demonstration, no attempt was made to optimize the two parameters that control pattern acquisition.

Experiment 2. The second experiment involved two ADIOS instances: a teacher and a student. In each of the four runs, the teacher was pre-loaded with a ready-made context free grammar (using the straightforward translation of CFG rules into patterns), then used to generate a series of training corpora with up to 6400 sentences, each with up to seven levels of recursion. After training in each run i ($i = [1 \ldots 4]$), a student-generated test corpus $C^{(i)}_{learned}$ of size 10000 was used in conjunction with a test corpus $C^{(i)}_{target}$ of the same size produced by the teacher, to calculate precision and recall. This was done by running the teacher as a parser on $C^{(i)}_{learned}$ and the student – as a parser on $C^{(i)}_{target}$. The results, plotted in Figure 4, indicate a substantial capacity for unsupervised induction of context-free grammars even from very small corpora.

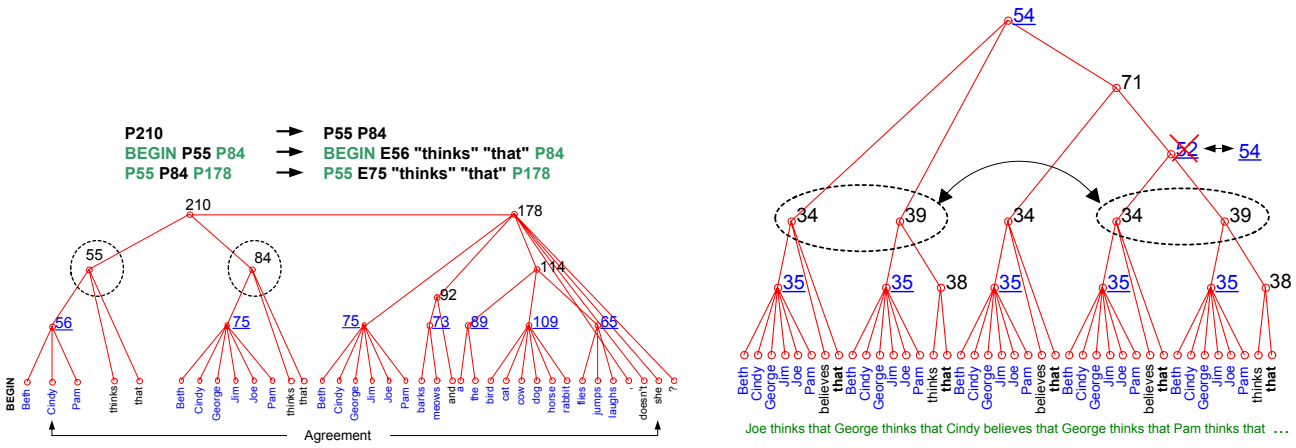


Figure 3: *Left:* because ADIOS does not rewire all the occurrences of a specific pattern, but only those that share the same context, its power is comparable to that of Context Sensitive Grammars. In this example, equivalence class #75 is not extended to subsume the subject position, because that position appears in a different context (e.g., immediately to the right of the symbol BEGIN). Thus, long-range agreement is enforced and over-generalization prevented. The context-sensitive "rules" corresponding to pattern #210 appear above it. *Right:* the ADIOS pattern representation facilitates the detection of recursive structure, exemplified here by the correspondence between equivalence classes #52 and #54.

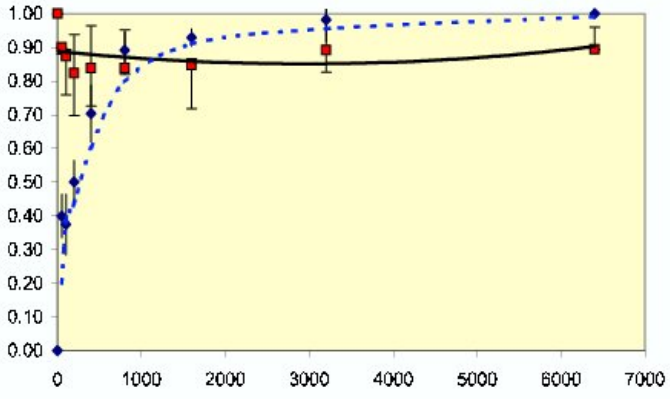


Figure 4: the results of Experiment 2; precision (squares) and recall (diamonds), plotted vs. the size of the training corpus; the error bars are std. dev. computed over four separate training/testing runs. Note that even the largest training corpus size, 6400 sentences, is a tiny proportion of the approximately $1.6 \cdot 10^8$ sentences that can be generated by the target grammar under the chosen depth constraint (7).

Related computational and linguistic formalisms and psycholinguistic findings

Unlike ADIOS, very few existing algorithms for unsupervised language acquisition use raw, unannotated corpus data (as opposed, say, to sentences converted into sequences of POS tags). The only work described in a recent review [6] as completely unsupervised — the GraSp model [13] — does attempt to induce syntax from raw transcribed speech, yet it is not completely data-driven in that it makes a prior commitment to a particular theory of syntax (Categorial Grammar, complete with a pre-specified set of allowed categories). Because of the unique nature of our chosen challenge — finding structure in language rather than imposing it — the following brief survey of grammar induction focuses on contrasts and comparisons to approaches that generally stop short of attempting to do what our algorithm does. We distinguish below between approaches that are motivated by computational considerations (Local Grammar and Variable Order Markov models, and Tree Adjoining Grammar), and those whose main motivation is linguistic and cognitive psychological (Cognitive and Construction grammars).

Local Grammar and Markov models. In capturing the regularities inherent in multiple criss-crossing paths through a corpus, ADIOS superficially resembles finite-state Local Grammars [14] and Variable Order Markov (VOM) models [15] that aim to produce a minimum-entropy finite-state encoding of a corpus. There are, however, crucial differences, as explained below. Our pattern significance criteria [3] involve conditional probabilities of the form $P(e_n|e_1, e_2, e_3, \ldots, e_{n-1})$, which does bring to mind an n'th-order Markov chain, with the (variable) n corresponding roughly to the length of the sentences we deal with. The VOM approach starts out by postulating a maximum-n VOM structure, which is then fitted to the data. The maximum VOM order n, which effectively determines the size of the window under consideration, is in practice much smaller than in our approach, because of computational complexity limitations of the VOM algorithms. The final parameters of the VOM are set by a maximum likelihood condition, fitting the model to the training data. The ADIOS philosophy differs from the VOM approach in several key respects. *First*, rather than fitting a model to the data, we use the data to construct a (recursively structured) graph. Thus, our algorithm naturally addresses the inference of the graph's structure, a task

that is more difficult than the estimation of parameters for a given configuration. *Second*, because ADIOS works from the bottom up in a data-driven fashion, it is not hindered by complexity issues, and can be used on huge graphs, with very large windows sizes. *Third*, ADIOS transcends the idea of VOM structure, in the following sense. Consider a set of patterns of the form $b_1[c_1]b_2[c_2]b_3$, etc. The equivalence classes $[\cdot]$ may include vertices of the graph (both words and word patterns turned into nodes), wild cards (i.e., any node), as well as ambivalent cards (any node or no node). This means that the terminal-level length of the string represented by a pattern does not have to be of a fixed length. This goes conceptually beyond the variable order Markov structure: $b_2[c_2]b_3$ do not have to appear in a Markov chain of a finite order $||b_2|| + ||c_2|| + ||b_3||$ because the size of $[c_2]$ is ill-defined, as explained above. *Fourth*, as we showed earlier (Figure 3), ADIOS incorporates both context-sensitive substitution and recursion.

Tree Adjoining Grammar. The proper place in the Chomsky hierarchy for the class of strings accepted by our model is between Context Free and Context Sensitive Languages. The pattern-based representations employed by ADIOS have counterparts for each of the two composition operations, substitution and adjoining, that characterize a Tree Adjoining Grammar, or TAG, developed by Joshi and others [16]. Specifically, both substitution and adjoining are subsumed in the relationships that hold among ADIOS patterns, such as the membership of one pattern in another. Consider a pattern $\mathcal{P}_i$ and its equivalence class $\mathcal{E}(\mathcal{P}_i)$; any other pattern $\mathcal{P}_j \in \mathcal{E}(\mathcal{P}_i)$ can be seen as substitutable in $\mathcal{P}_i$. Likewise, if $\mathcal{P}_j \in \mathcal{E}(\mathcal{P}_i)$, $\mathcal{P}_k \in \mathcal{E}(\mathcal{P}_i)$ and $\mathcal{P}_k \in \mathcal{E}(\mathcal{P}_j)$, then the pattern $\mathcal{P}_j$ can be seen as adjoinable to $\mathcal{P}_i$. Because of this correspondence between the TAG operations and the ADIOS patterns, we believe that the latter represent regularities that are best described by Mildly Context-Sensitive Language formalism [16]. Importantly, because the ADIOS patterns are learned from data, they already incorporate the constraints on substitution and adjoining that in the original TAG framework must be specified manually.

Psychological and linguistic evidence for pattern-based representations. Recent advances in understanding the psychological role of representations based on what we call patterns, or *constructions* [17], focus on the use of statistical cues such as conditional probabilities in pattern learning [18, 19], and on the importance of exemplars and constructions in children's language acquisition [20]. Converging evidence for the centrality of pattern-like structures is provided by corpus-based studies of the prevalence of "prefabricated" sequences of words [21], and of the entrenchment of such sequences in the lexicon [22]. Similar ideas concerning the ubiquity in syntax of structural peculiarities hitherto marginalized as "exceptions" are now being voiced by linguists [23, 24].

Cognitive Grammar; Construction Grammar. The main methodological tenets of ADIOS — populating the lexicon with "units" of varying complexity and degree of entrenchment, and using cognition-general mechanisms for learning and representation — fit the spirit of the foundations of Cognitive Grammar [9]. At the same time, whereas the cognitive grammarians typically face the chore of hand-crafting structures that would reflect the logic of language as they perceive it, ADIOS discovers the primitives of grammar empirically and autonomously. The same is true also for the comparison between ADIOS and the various Construction Grammars [17, 24], which are all hand-crafted. A construction grammar consists of elements that differ in their complexity and in the degree to which they are specified: an idiom such as "big deal" is a fully specified, immutable construction, whereas the expression "the X, the Y" – as in "the more, the better" [25] – is a partially specified template. The patterns learned by ADIOS likewise vary along the dimensions of complexity and specificity (e.g., not every pattern has an equivalence class).[2]

Related computational work on grammar induction

In natural language processing, a distinction is usually made between unsupervised learning methods that attempt to find good structural primitives and those that merely seek good parameter settings for predefined primitives. ADIOS, which clearly belongs to the first category, is also capable of learning from raw data, whereas most other systems start with corpora annotated by part of speech tags [26], or even rely on treebanks, or collections of hand-parsed sentences [4]. Of the many such methods, we can mention here only a few.

Global grammar optimization using tagged data. Stolcke and Omohundro (1994) learn structure (the topology of a Hidden Markov Model, or the productions of a Stochastic Context Free Grammar), by iteratively maximizing the probability of the current approximation to the target grammar, given the data. In contrast to this approach, which is global in that all the data contribute to the figure of merit at each iteration, ADIOS is local in the sense that its inferences only apply to the current bundle candidate. Another important difference is that instead of general-scope rules stated in terms of parts of speech, we seek context-specific patterns. Perhaps because of its globality and unrestricted-scope rules, Stolcke and Omohundro's method has "difficulties with large-scale natural language applications" [27]. Similar conclusions are reached by Clark, who observes that POS tags are not enough to learn syntax from ("a lot of syntax depends on the idiosyncratic properties of particular words." [5], p.36). His algorithm attempts to learn a phrase-structure grammar from tagged text, by starting with local distributional cues, then filtering spurious non-terminals using a mutual information criterion. In the final stage, his algorithm clusters the results to achieve a minimum description length (MDL) representation, by starting with maximum likelihood grammar, then greedily selecting the candidate for abstraction that would maximally reduce the description length. In its greedy approach to optimization (but not in its local search for good patterns or its ability to deal with untagged data), our approach resembles Clark's.

[2] Similarly to constructions, the ADIOS patterns carry semantic, and not just syntactic, information — an important issue that is outside the scope of the present paper.

Probabilistic treebank-based learning. Bod, whose algorithm learns by gathering information about corpus probabilities of potentially complex trees, observes that "[...] the knowledge of a speaker-hearer cannot be understood as a grammar, but as a statistical ensemble of language experiences that changes slightly every time a new utterance is perceived or produced. The regularities we observe in language may be viewed as emergent phenomena, but they cannot be summarized into a consistent non-redundant system that unequivocally defines the structures of new utterances." [4], p.145. This memory- or analogy-based language model, which is not a typical example of unsupervised learning, is mentioned here mainly because of the parallels between its data representation, Stochastic Tree-Substitution Grammar, and some of the formalisms discussed earlier.

Split and merge pattern learning. The unsupervised structure learning algorithm developed by Wolff between 1970 and 1985 stands out in that it does not need the corpus to be tagged. An excellent survey of his own and earlier attempts at unsupervised learning of language, and of much relevant behavioral data, can be found in [28]. His representations consist of SYN (syntagmatic), PAR (paradigmatic) and M (terminal) elements. Although our patterns and equivalence classes can be seen as analogous to the first two of these, Wolff's learning criterion is much simpler than that of ADIOS: in each iteration, the most frequent pair of contiguous SYN elements are joined together.[3] His system, however, had a unique provision for countering the usual propensity of unsupervised algorithms for overgeneralization: PAR elements that did not admit free substitution among all their members in some context were rebuilt in a context-specific manner. Unfortunately, Wolff's system has not been tested on unconstrained natural language.

Summary, prospects and challenges

The ADIOS approach to the representation of linguistic knowledge resembles the Construction Grammar in its general philosophy (e.g., in its reliance on structural generalizations rather than on syntax projected by the lexicon), and the Tree Adjoining Grammar in its computational capacity (e.g., in its apparent ability to accept Mildly Context Sensitive Languages). The representations learned by the ADIOS algorithm are truly emergent from the (unannotated) corpus data, whereas those found in published works on cognitive and construction grammars and on TAGs are hand-tailored. Thus, our results complement and extend both the computational and the more linguistically oriented research into cognitive/construction grammar.

To further the cause of an integrated understanding of language, a crucial challenge must be met: a viable approach to the evaluation of performance of an unsupervised language learner must be developed, allowing testing both (1) neutral with respect to the linguistic dogma, and (2) cognizant of the plethora of phenomena documented by linguists over the course of the past half century (see, e.g., Figure 5).

[3] An even simpler criterion, that of mere repetition, is employed by the related approach of [29], resulting in a rule set that appears to grow linearly with the size of the corpus, rather than reaching an asymptote as in our case.

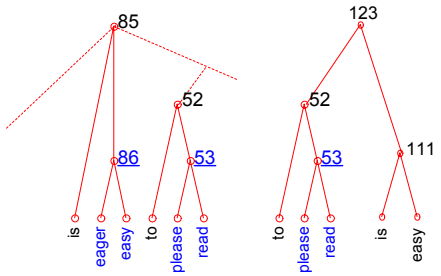


Figure 5: As a token of our intention to account, eventually, for the entire spectrum of English syntax-related phenomena described in the textbooks — agreement, anaphora, auxiliaries, *wh*-questions, passive, control, etc. [30] — we illustrate here the manner in which ADIOS treats tough movement (another phenomenon, long-range agreement, was discussed in Figure 2). When trained on sentences exemplifying "tough movement", ADIOS forms patterns that represent the correct phrases (... is easy to read, is easy to please, is eager to read, is eager to please, to read is easy and to please is easy), but does not over-generalize to the incorrect ones (*to read is eager or *to please is eager).

Unsupervised grammar induction algorithms that work from raw data are in principle difficult to test, because any "gold standard" to which the acquired representation can be compared (such as the Penn Treebank [31]) invariably reflects its designers' preconceptions about language, which may not be valid, and which usually are controversial among linguists themselves [32]. Moreover a child "...must generalize from the sample to the language without overgeneralizing into the area of utterances which are not in the language. *What makes the problem tricky is that both kinds of generalization, by definition, have zero frequency in the child's experience.*" ([28], p.183, italics in the original). Instead of shifting the onus of explanation for this "miracle" onto some unspecified evolutionary processes (which is what the innate grammar hypothesis amounts to), we suggest that a system such as ADIOS should be tested by monitoring its acceptance of massive amounts of human-generated data, and at the same time by getting human subjects to evaluate sentences generated by the system (note that this makes psycholinguistics a crucial component in the entire undertaking).

A purely empirical approach to the evaluation problem would, however, waste the many valuable insights into the regularities of language accrued by the linguists over decades. Although some empiricists would consider this a fair price for quarantining what they perceive as a runaway theory that got out of touch with psychological and computational reality, we believe that searching for a middle way is a better idea, and that the middle way can be found, if the linguists can be persuaded to try and present their main findings in a theory-neutral manner. From recent reviews of syntax that do attempt to reach out to non-linguists (e.g., [33]), it appears that the core issues on which every designer of a language acquisition system should be focusing are dependencies (such as co-reference) and constraints (such as islands), especially as seen in a typological (cross-linguistic) perspective [24].

Acknowledgment. Supported by the US-Israel Binational Science Foundation.

References

[1] Z. Solan, E. Ruppin, D. Horn, and S. Edelman. Automatic acquisition and efficient representation of syntactic structures. In S. Thrun, ed., *Advances in Neural Information Processing (NIPS)*, vol. 15, Cambridge, MA, 2003. MIT Press.

[2] Z. Solan, E. Ruppin, D. Horn, and S. Edelman. Unsupervised efficient learning and representation of language structure. In R. Alterman and D. Kirsh, eds., *Proc. 25th Conf. of the Cognitive Science Society*, Hillsdale, NJ, 2003. Erlbaum.

[3] Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised context sensitive language acquisition from a large corpus. In L. Saul, ed., *NIPS*, vol. 16, Cambridge, MA, 2004. MIT Press.

[4] R. Bod. *Beyond grammar: an experience-based theory of language*. CSLI Publications, Stanford, US, 1998.

[5] A. Clark. *Unsupervised Language Acquisition: Theory and Practice*. PhD thesis, COGS, University of Sussex, 2001.

[6] A. Roberts and E. Atwell. Unsupervised grammar inference systems for natural language. TR 2002.20, School of Computing, University of Leeds, 2002.

[7] N. Chomsky. *Knowledge of language: its nature, origin, and use*. Praeger, New York, 1986.

[8] J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. *Rethinking innateness: A connectionist perspective on development*. MIT Press, Cambridge, MA, 1996.

[9] R. W. Langacker. *Foundations of cognitive grammar*, vol.I: theoretical prerequisites. Stanford University Press, Stanford, CA, 1987.

[10] Z. S. Harris. Distributional structure. *Word*, 10:140–162, 1954.

[11] M. van Zaanen. ABL: Alignment-Based Learning. In *COLING 2000 - Proceedings of the 18th International Conf. on Computational Linguistics*, pp. 961–967, 2000.

[12] F. Pereira. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 358(1769):1239–1253, 2000.

[13] P. J. Henrichsen. GraSp: Grammar learning form unlabeled speech corpora. In *Proceedings of CoNLL-2002*, pp. 22–28. Taipei, Taiwan, 2002.

[14] M. Gross. The construction of local grammars. In E. Roche and Y. Schabès, eds., *Finite-State Language Processing*, pp. 329–354. MIT Press, Cambridge, MA, 1997.

[15] I. Guyon and F. Pereira. Design of a linguistic postprocessor using Variable Memory Length Markov Models. In *Proc. 3rd Int'l Conf. Document Analysis and Recogition*, pp. 454–457, Montreal, Canada, 1995.

[16] A. Joshi and Y. Schabes. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, eds., *Handbook of Formal Languages*, vol. 3, pp. 69–124. Springer, Berlin, 1997.

[17] A. E. Goldberg. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7:219–224, 2003.

[18] J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928, 1996.

[19] R. L. Gómez and L. Gerken. Infant artificial language learning and language acquisition. *Trends in Cognitive Science*, 6:178–186, 2002.

[20] T. Cameron-Faulkner, E. Lieven, and M. Tomasello. A construction-based analysis of child directed speech. *Cognitive Science*, 27:843–874, 2003.

[21] A. Wray. *Formulaic language and the lexicon*. Cambridge University Press, Cambridge, UK, 2002.

[22] C. L. Harris. Psycholinguistic studies of entrenchment. In J. Koenig, ed., *Conceptual Structures, Language and Discourse*, vol. 2. CSLI, Berkeley, CA, 1998.

[23] P. W. Culicover. *Syntactic nuts: hard cases, syntactic theory, and language acquisition*. Oxford University Press, Oxford, 1999.

[24] W. Croft. *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford University Press, Oxford, 2001.

[25] P. Kay and C. J. Fillmore. Grammatical constructions and linguistic generalizations: the What's X Doing Y? construction. *Language*, 75:1–33, 1999.

[26] D. Klein and C. D. Manning. Natural language grammar induction using a constituent-context model. In T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., *NIPS 14*, Cambridge, MA, 2002. MIT Press.

[27] A. Stolcke and S. Omohundro. Inducing probabilistic grammars by Bayesian model merging. In R. C. Carrasco and J. Oncina, eds., *Grammatical Inference and Applications*, pp. 106–118. Springer, 1994.

[28] J. G. Wolff. Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I. M. Schlesinger, and M. D. S. Braine, eds., *Categories and Processes in Language Acquisition*, pp. 179–215. Lawrence Erlbaum, Hillsdale, NJ, 1988.

[29] C. G. Nevill-Manning and I. H. Witten. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, 7:67–82, 1997.

[30] I. A. Sag and T. Wasow. *Syntactic theory: a formal introduction*. CSLI Publications, Stanford, CA, 1999.

[31] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[32] A. Clark. Unsupervised induction of Stochastic Context-Free Grammars using distributional clustering. In *Proceedings of CoNLL 2001*, Toulouse, 2001.

[33] C. Phillips. Syntax. In L. Nadel, ed., *Encyclopedia of Cognitive Science*, vol. 4, pp. 319–329. Macmillan, London, 2003.

Fast and Frugal Reasoning Enhances a Solver for Hard Problems

Susan L. Epstein (susan.epstein@hunter.cuny.edu)
Tiziana Ligorio (t.ligorio@verizon.net)
Department of Computer Science,
Hunter College and The Graduate Center of The City University of New York
695 Park Avenue, New York, NY 10021 USA

Abstract

This paper describes how a program that learns to solve hard problems has been enhanced with fast and frugal, recognition-based reasoning methods. The program uses these methods to help manage its own heuristics for solving constraint satisfaction problems. The result is a constraint solver that reasons quickly, much the way people appear to induce reasonable decisions in real-world situations. Empirical evidence on a variety of problems indicates that fast and frugal reasoning can accelerate the solution of very difficult problems, often without introducing additional error.

The thesis of this work is that the same *fast and frugal* reasoning which people use to formulate decisions in real-world situations (Gigerenzer, Todd, & The ABC Research Group, 1999) can accelerate autonomous decision makers without endangering their reliability. We investigate this premise with a program that learns how to combine heuristics to solve constraint satisfaction problems (*CSPs*). The principal result reported here is that, on difficult CSPs, when recognition alone is not sufficient, fast and frugal, recognition-based reasoning enhances the solver. This is particularly noteworthy because the program first learns how to apply its CSP heuristics within its problem environment, and then hones its performance using fast and frugal reasoning. We believe this to be the first exploration of fast and frugal reasoning on a large body of challenging problems whose difficulty can be explicitly characterized and whose solution can be incisively assessed.

Many large-scale, real-world problems in areas such as design and configuration, planning and scheduling, and diagnosis and testing are readily understood, represented, and solved as CSPs. CSP solution is, in general, not known to be solvable by algorithms of any but exponential complexity. Thus the effectiveness of fast and frugal reasoning on these problems is counterintuitive.

Fast and frugal reasoning assumes pre-acquired, accurate, problem-area knowledge (Gigerenzer, Todd, & The ABC Research Group, 1999). In real-world decisions, fast and frugal reasoning adaptively exploits the environment's structure. A program provided with heuristics must learn their accuracy before turning to fast and frugal decision making. The first sections of this paper provide background information on fast and frugal reasoning, and on CSP. Subsequent sections describe how we addressed these ideas in a program that learns, describe the experimental design, discuss the results, and sketch future work.

Background

As often happens in interdisciplinary work, terminology overlaps here, but meaning does not. Thus we alert the reader to the fact that, although "domain" means "problem area" in some fields, it has a different connotation (described below) in constraint solving, for which we reserve it. Similarly, fast and frugal researchers generally refer to their general problem-solving methods (e.g., Minimalist, Take the Last, Take the Best) as heuristics, but so too do CSP researchers, and again we take the CSP definition. As a result, we describe fast and frugal methods as "strategies that consult heuristics" rather than "heuristics that consult cues." Finally, the notion of recognition, which underlies the fast and frugal strategies, is itself a heuristic, albeit a more general one, defined below.

Fast and frugal reasoning

Under limited time, there exists a trade-off between decision making speed and correctness. When pressed for time, people may limit their search for information to guide them in the decision process with *non-compensatory* strategies, strategies that use a single heuristic to prefer a single option (Gigerenzer & Goldstein, 1996). People appear to work from an *adaptive toolbox*, a collection of cognitive mechanisms for inference in specific problem areas (Gigerenzer, Todd, & The ABC Research Group, 1999). This adaptive toolbox includes low-order perceptual and memory processes, including fast and frugal strategies that may be combined to account for higher-level mental processes. Such a model of cognitive heuristics is *ecologically rational*, grounded in environment-specific structure and characteristics (Goldstein & Gigerenzer, 2002; Gigerenzer & Selten, 2001; Gigerenzer, Todd, & The ABC Research Group, 1999).

In *one-reason* decision making, a set of heuristics is consulted one at a time, until some heuristic is able to *discriminate*, that is, able to select a single option. *Recognition* is the *foundation heuristic* for fast and frugal reasoning: it favors recognized options over unrecognized ones. Recognition discriminates if and only if exactly one option is recognized (Gigerenzer & Goldstein, 1996). In that case, it is sufficient for one-reason decision making, and no further computation is required.

When recognition alone does not discriminate, each strategy considered here may be thought of as a meta-heuristic that speeds the selection of the next heuristic. All three strategies initially try recognition on all the available options. If more than one option is recognized, other heuristics are consulted, one at a time, on a randomly-

selected pair of recognized options, until some heuristic does discriminate. The preferred option is then selected. The way these "other" heuristics are chosen defines the decision-making strategy. The following are drawn from Gigerenzer, Todd and the ABC Research Group (1999):

- *Minimalist*: Select a heuristic at random, until one discriminates among the options and a decision is made.
- *Take the Last*: Use the last heuristic, the one that discriminated the last time a decision was made, when recognition did not. This captures the human tendency to re-use the most recent successful strategy.
- *Take the Best*: Use the heuristic known to work best in a specific environment. The insightfulness of a heuristic on a set of problems is called its *ecological validity*.

Constraint satisfaction problems (CSPs)

CSPs are a good vehicle with which to explore fast and frugal reasoning. They arise in classes whose difficulty has a mathematical characterization, and many examples can be readily generated within each class. Furthermore, well-established criteria exist with which to gauge the performance of a program that solves them.

A CSP consists of a set of variables, each associated with a *domain* of possible values for assignment, and a set of *constraints* that specify which combinations of values are allowed. To represent a real-world problem as a CSP, one casts the entities involved as variables, and expresses the relationships required among these variables as constraints. A simple example appears in Figure 1. Real-world CSPs, however, involve many more variables, substantial domains, and a broad variety of interacting, more sophisticated restrictions as constraints. A solution for a CSP is one value for each variable such that all constraints are satisfied. Every CSP has an underlying *constraint graph* that represents each variable by a vertex. An edge in the graph represents a constraint between the corresponding variables, and is labeled by their permissible pairs of values. The *degree* of a variable is the number of edges to it in the graph. (For simplicity we restrict discussion here to binary CSPs.)

How hard a CSP is to solve is determined by how difficult it is to find values that satisfy all its constraints at once. A *class* of CSPs groups together problems with four parameters thought to estimate their difficulty. A CSP class may be described by <*n*,*k*,*d*,*t*>, where *n* is its number

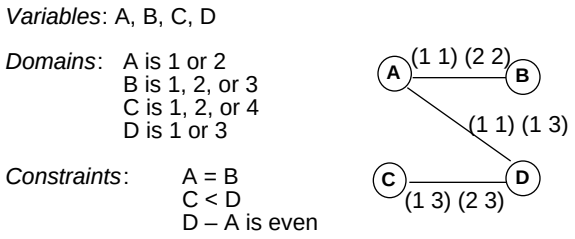


Figure 1: A simple constraint satisfaction problem and its underlying constraint graph. Edges in the graph are labeled with acceptable value pairs, computed from the domains and the constraints.

of variables and *k* its maximum domain size. The *density d* is the fraction of possible edges in the underlying constraint graph. The *tightness t* is the percentage of possible value pairs the constraints exclude. Thus in <30,8,.26,.66> every problem has 20 variables, each with domain size at most 8, and 30^8 possible value assignments. In a given class, every CSP has the same values for *n, k, d,* and *t,* and the same minimal number of decisions for solution.

To solve a CSP, one can repeatedly select a variable and assigns it a value consistent with its constraints. For example, Figure 2 represents a possible search for a solution to the problem in Figure 1. Reading from top to bottom and from left to right, each circle (*node*) represents a decision. There, Variable A was selected first, and assigned the value 2, then D was selected, and both its values (1 and 3) were tried without success. Therefore, search backed up, the value 2 was withdrawn (*retracted*) from A, and the value 1 was assigned to A instead.

The black nodes in Figure 2 represent the correct decisions, and the solid path on the right represents the solution. When a value assignment is inconsistent with the constraints, it is retracted and another assignment is tried. In Figure 2, white nodes are errors, assignments that cause subsequent decisions (the gray nodes) to be retracted, so that an error can be corrected. For example, the assignment D = 1 is *inconsistent* because it leaves no values for C. When that happens, values are retracted until all current assignments are once again consistent, and new values tried. Finding a single solution to a solvable problem this way requires at least *n* assignments.

Although CSP solution is NP-hard, some problem classes surrender readily to heuristics. For a solvable CSP, the order in which one selects variables (*variable ordering,* e.g., A, D, C, B in Figure 2) can speed solution, as can the order in which one assigns a value to a just-selected variable (*value ordering,* e.g., 2, 1 for A in Figure 2). There are dozens of variable-ordering and value-ordering heuristics in the CSP literature, but their interactions are ill-understood. The best approximation for CSP problem difficulty is currently *kappa*, which is defined as a function of *n, k, d,* and *t* (Gent, MacIntyre, Prosser, & Walsh, 1996). Nonetheless, two problems from the same class may still require different amounts of effort to solve.

During search, a CSP solver can apply a variety of inference and retraction methods. When a *partial solution* (a set of values assigned to a proper subset of the variables) is incompatible with the constraints, all the nodes that include it (e.g., gray in Figure 2) may be eliminated. An *inference* method can propagate the implications of a newly-assigned value on to the remainder of the as-yet-unassigned variables. Different amounts of inference are possible, and there are tradeoffs between inference effort and search savings. A specific inference method (the central one is arc consistency) can be carried out to varying degrees (Sabin & Freuder, 1994) and with different algorithms (Bessière & Régin, 2001). A *retraction* method re-

sponds to an inconsistent partial solution (a subtree of discarded nodes rooted at an error node in Figure 2). The standard retraction method is *chronological backtracking*, withdrawal of the most recent assignment(s).

ACE

ACE (the Adaptive Constraint Engine) is an autonomous system that learns to solve classes of CSPs (Epstein, Freuder, Wallace, Morozov, & Samuels, 2002). ACE is based on *FORR* (FOr the Right Reasons), a cognitively-oriented architecture for learning and problem solving that supports the development of expertise (Epstein, 1994). Here, "cognitively-oriented" means that FORR's reasoning structure emulates approaches readily observable in human problem solving, highly-effective approaches not always found in traditional AI artifacts (Biswas, Goldman, Fisher, Bhuva, & Glewwe, 1995; Crowley & Siegler, 1993; Klein & Calderwood, 1991; Kojima & Yoshikawa, 1998; Novick & Coté, 1992; Schraagen, 1993). FORR is based on the premise that decisions are composed from more and less trustworthy rationales. One constructs a FORR-based program by specifying heuristics that underlie decision making in a particular problem area.

ACE is an ambitious program – armed with many heuristics, it can tackle difficult problems. The version we used here begins with 40 CSP heuristics which are initially classified into a hierarchy of three *tiers* by the user. ACE moves through those tiers to make a decision. The heuristics in tiers 1 and 2 are consulted sequentially; the heuristics in tier 3 are consulted (effectively) in parallel.

Tier 1 consists of *perfect* (i.e., error-free) heuristics consulted sequentially. If any heuristic supports an action, that action is executed without reference to any subsequent heuristics. Consulting perfect heuristics first ensures that obvious correct decisions (e.g., a checkmate at chess) will be reached without devoting resources to other, less reliable heuristics. A perfect heuristic that opposes an action (e.g., "don't move into checkmate") removes that action from consideration by all subsequent heuristics, thereby preventing obvious errors. Thus placing perfect heuristics in tier 1 permits easy problems to be solved easily, a feature all too rare in complex systems. (The second tier is not applicable to the work reported here; the interested reader is referred to (Epstein, 1998)).

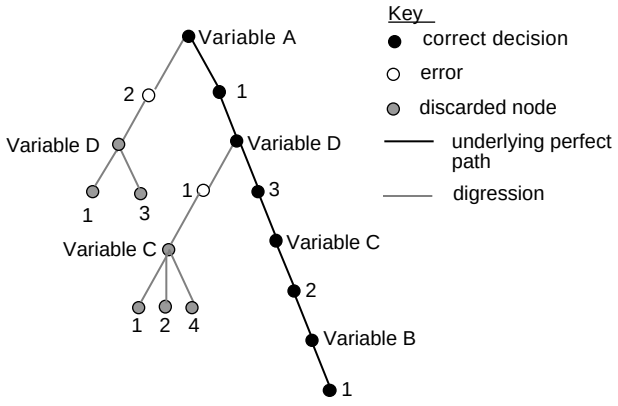


Figure 2: A search tree for solution to the simple CSP in Figure 1, depicted here as alternating variable selections and value selections. When a value selection violates some constraint, that decision is retracted, and search backtracks. New values are assigned until all variables have a value consistent with the constraints.

Tier-3 heuristics are the ordinary ones; they produce single-action *comments* that are not guaranteed to be correct. All but two of the ACE heuristics in this version lie in tier 3. Because they are fallible, their comments are combined to select the next action in a process called *voting*. Each heuristic may vote on different actions with different strengths, or it may remain silent. ACE's tier 3 heuristics were, for the most part, drawn from the CSP expert community. In every case, however, the dual of a popular heuristic was also implemented. For example, the CSP literature suggests that the next variable to have a value assigned to it should have a minimum *dynamic domain size* (set of values that would still be consistent with existing variable assignments). ACE therefore has a heuristic that comments in favor of such variables, but it also has its dual, a heuristic to maximize the dynamic domain size of a variable.

One-Reason Decision Making in ACE

ACE learns to solve CSPs efficiently by winnowing through its heuristics and balancing them appropriately. Laden with CSP knowledge, ACE's decisions are carefully reasoned but time-consuming. Fast and frugal, one-reason decision making seemed a reasonable enhancement, but its adaptation for ACE required careful thought about recognition, ecological rationality, ecological validity, and preference functions.

Recognition in ACE

Recognition is familiarity with something previously experienced. Recall that, during CSP solution, there are only two kinds of experience: variable selection and value selection. Therefore, we define recognition to be the identification of a current option as one previously selected during search *in the current problem*. Note that recognition for ACE is a selection heuristic rather than a trigger for situation assessment and re-evaluation, because mere recognition of a single option is sufficient for decision making without any consideration of the similarities of the different search states (situations) in which it previously occurred. For example, in Figure 2 assigning 1 to Variable D after Variable A is set to 1, eventually proves to be an error, since all possible assignments to Variable C are then incompatible with the constraints. In response to the detected error, the gray nodes corresponding to the selection of Variable C and the values tried for it are retracted. The decisions made in this subtree, *excluding* the error node, will be recognized during subsequent search. Effectively, later in search, a decision point that includes

options previously found attractive will find them attractive again.

A counterargument to the role of recognition in human reasoning is that recognition is often associated with other cues to obtain a given judgment, and is thus *compensatory* (Oppenheimer, 2003). When an object is recognized, other information about that object is also recalled at the moment of recognition. If this "extra" information is relevant to the judgment being made, it will be taken into account while making the decision. The relevant information may support or oppose the judgment about the recognized object. Although the additional information may be contradictory, for our purposes, the influence it has on recognition is considered as a whole. Recognition may have an attached positive or negative correlation with respect to the judgment being made.

For ACE, recognition always has a positive correlation, because it treats a previous decision as one it wants to make again. The idea here is that, if no further knowledge about ACE's tier 3 heuristics has been acquired during search on the current problem, and consulting these same heuristics previously led to making certain decisions, they should remain valid, if the option is still available. (Consistency checking may have eliminated it.) By "recognizing" such previously-computed but subsequently-retracted decisions, ACE can avoid reconsulting all its heuristics on the same (or most of the same) options. For example, in Figure 2, ACE initially assigned 2 to A and then chose Variable D. Later, when 2 is retracted and 1 assigned to A, the next variable must be selected. At that point D is recognized and so need not be recomputed. Note too that recognition can lead ACE to repeat an error. It tries 1 before 3 for D on both sides of the search tree in Figure 2.

Ecological rationality and ecological validity

ACE must have acquired problem-class-specific knowledge before it attempts speed-up through one-reason decision making. Tier-3 heuristics are ACE's version of the adaptive toolbox, its knowledge about how CSP works. ACE's heuristics, however, are not all of equal significance or reliability in a particular problem class. Therefore, ACE learns weights to combine them.

DWL (Digression-based Weight Learning) learns problem-class-specific weights for tier-3 heuristics (Epstein et al., 2002). It is specifically designed to minimize errors during solution (and therefore minimize the number of nodes in the tree of Figure 2). After ACE solves a problem, DWL examines the solution trace, and adjusts the weight of each heuristic according to whether or not it supported the correct decisions. All heuristics start with the same weight. DWL provides the ecological validity for the heuristics in a given problem class.

DWL also employs non-voting, baseline heuristics that discriminate with randomly-generated strength on randomly-chosen options. DWL uses them to gauge ACE's own heuristics, so that ACE learns to value only those heuristics that make comments more valuable than random ones. These baseline heuristics provide ACE's ecological rationality.

From preference to binary decisions

The one-reason decision-making model assumes binary heuristics, ones that vote either in favor of or against an option. Recall that, if recognition does not discriminate, the model considers other heuristics on randomly-selected pairs of options, until some heuristic discriminates. Recognition is a binary heuristic, and we translate it as such for ACE: a decision was either previously made or not. A tier-3 heuristic, however, expresses its preference for, or opposition to, an option in a comment whose *strength* lies between 0 and 10. To adapt ACE for one-reason decision making, the option with the higher strength is deemed the positive one. If a heuristic comments on both options with the same strength, or it does not comment at all, the heuristic does not discriminate, and another heuristic is consulted, depending upon the particular strategy in use.

Experimental Design

The ACE project maintains a large library of problem classes, each with many examples. In each experiment here, ACE learned by attempting to solve at most 600 problems (the *learning phase*) and then was tested on 200 different problems from the same class, with learning turned off (the *testing phase*). This learn-and-then-test approach was repeated 10 times, each time with different learning problems but the same testing problems. Fast and frugal reasoning was applied only in the testing phase.

ACE's performance here is evaluated by three standard CSP criteria: average number of nodes in the decision tree (e.g. Figure 2), average number of mistakes during solution (number of retractions), and average computation time (in seconds). Any differences identified in the following discussion are statistically significant at the .95 level. Learning was terminated early if the heuristics' weights *stabilized* (did not vary in their standard deviation by more than 0.1 over the most recent 20 problems) before 600 problems. ACE used chronological backtracking for retraction and MAC3 (Mackworth, 1977) for consistency checking. What we varied in our experiments was the problem class, and which non-compensatory search strategy was used in the testing phase.

Results

We tested ACE alone, and then with each of the fast and frugal strategies on each of three problem classes. The results appear in Table 1. The first class of problems was <30, 8, .26, .66>, an extremely difficult set of randomly-generated CSPs. (The state of the CSP art does not yet support labeling them "the most difficult" for their size, but these are certainly "exceptionally difficult.") On this class, each of the three non-compensatory strategies, combined with recognition significantly improved overall execution time. Speed-up came with a price, however. Although ACE solved every problem, it made more (albeit relatively trivial) errors, and explored more nodes during search. The most reasoned and ecologically ra-

Table 1: Performance of ACE alone and with the recognition heuristic guided by three different non-compensatory, one-reason decision making strategies, on two classes of CSPs: < 30, 8, .26, .66> and < 30, 8, .12, .5>. Time (in seconds), errors and nodes are per problem. Figures in bold represent a statistically significant improvement over ACE without recognition.

Class	Criterion	ACE		Minimalist		Take the Last		Take the Best	
30-8-.26-.66	Overall time	3.10	(3.04)	**2.72**	(2.57)	**2.85**	(2.73)	**2.37**	(2.19)
	Tier 3 time	1.41	(1.51)	**0.70**	(0.62)	**0.75**	(0.68)	**0.61**	(0.68)
	Errors	105.11	(107.64)	119.12	(134.69)	116.97	(129.27)	113.77	(127.33)
	Nodes	165.11	(107.64)	179.12	(134.69)	176.97	(129.27)	173.77	(127.33)
30-8-.12-.5	Overall time	0.76	(0.57)	**0.71**	(0.49)	**0.71**	(0.42)	**0.66**	(0.44)
	Tier 3 time	0.44	(0.38)	**0.37**	(0.31)	**0.37**	(0.28)	**0.36**	(0.34)
	Errors	13.80	(15.52)	13.38	(15.26)	14.12	(16.17)	12.97	(14.36)
	Nodes	73.80	(15.52)	73.38	(15.26)	74.12	(16.17)	72.97	(14.36)

tional strategy (Take the Best) outperformed the other two.

The next class of problems on which we tested this approach was <30, 8, .12, .5>, a somewhat easier set. Here again, all three strategies achieved significant speedup, this time without increasing the number of errors or the size of the search tree. Finally, we tested our approach on a relatively easy class, <30, 8, .1, .5> (results not shown) where no changes could be detected.

Discussion

ACE is *complete*, that is, as constructed it is guaranteed to solve any solvable CSP — eventually. Expertise, however, requires that one solves problems efficiently. (All solutions to a CSP are defined to be equally good. When CSP researchers talk about *optimal* search, they refer one that does the least work, as measured by propagation.) Although extensive computation can minimize, or even eliminate, incorrect value selection, such computation may simply not be worth the time. Indeed, a solver that makes many inexpensive mistakes may actually arrive at a solution more quickly, despite a somewhat larger search tree. In this sense, ACE is a satisficer — it makes "good enough" decisions (Simon, 1981). Even when fast and frugal reasoning introduces additional error, the program solves problems faster. Both satisficing and our implementation of recognition, it should be noted, are tolerable only on problems where errors are relatively harmless.

The results of these experiments indicate that enhancing an intelligent and ecologically rational system with fast and frugal reasoning can save computation time, but is not guaranteed to do so. Because recognition, as we have implemented it, is a consequence of previous errors on the current problem, performance on these three problem classes requires individual explanations. On the relatively easy problems of <30, 8, .1, .5>, fast and frugal reasoning does not improve performance because there are not enough retractions during search to support subsequent recognition.

Fast and frugal reasoning that is also ecologically rational (i.e., Take the Best) provided more speed-up here than the other strategies. Recognition serves as a filter for the best of the options; it recycles earlier reasoning inherently. Nonetheless, ACE still needs to select from among the recognized options the one which is likely to be the most productive, and Take the Best is one way to do that.

On the problems of medium difficulty of <30, 8, .12, .5>, fast and frugal reasoning achieves speedup because it avoids some repeated computation. It also does so without significantly introducing more error, because recognition forces persistence by attempting to restrict ACE to previously-chosen options. Even if these "recycled decisions" are wrong, there are simply not enough wrong ones to introduce many new retractions.

On the very difficult problems of <30, 8, .26, .66>, Take the Best introduces significantly less error and does less work than the other fast and frugal strategies do. The additional errors on these problems suggest that accurate decision making here is more subtle and complex than a single heuristic can support, and certainly more than recognition alone can handle. Even without fast and frugal methods, ACE makes more mistakes solving these problems, simply because the problems are harder, Therefore there are more errors that may be recycled by recognition, as well as simply more recognized options to choose from. Take the Best introduces significantly fewer errors because it uses ecological rationality to avoid recycling some of them. Here again the speedup is achieved through the tradeoff between inexpensive errors and savings in computation time.

ACE's version of recognition is not the situation-based recognition described in (Klein & Calderwood, 1991). In that work, particular features of a situation bring to mind possible solution approaches, approaches that may require adaptation for the current situation. There, recognition may be paraphrased as "I was once in a similar situation where this sequence of decisions worked well, so I will see if I can adapt it to work again here, testing it first in simulation." ACE's recognition, in contrast, may be paraphrased as "I have seen that option before, considered it carefully (perhaps with different alternatives and in a somewhat different context at the time), and have decided to prefer it once again, without considering any potential consequences." ACE's recognition applies only to a single decision, not to a sequence; it does not permit adapta-

tion; it makes no situation assessment; and it relates to previous experience in the same problem.

To make use of fast and frugal reasoning, as we have implemented it, a system needs to have a body of heuristics with which to make decisions and, if it is to take the best, it needs a metric on those heuristics. Furthermore, since recognition, as we have interpreted it, requires decisions that are made under some erroneous circumstances and then withdrawn (gray nodes in Figure 2), the system must not have perfected decision making, or there will be no prior, within-problem decisions to recycle. Thus, fast and frugal reasoning can improve mediocre or even fairly reputable performance, but cannot improve flawless performance, for without errors there can be no recognition.

Future Work and Conclusion

Previously-made decisions are here recycled through the recognition heuristic. Whether or not ACE would have remade those decisions, and in the same sequence, without recognition remains to be determined. Future work will examine ACE's decisions at the same level but in different branches of the search tree, with and without fast and frugal reasoning.

Although we did not use it in these experiments, ACE can partition its tier-3 heuristics into any number of subclasses. We intend to compare ACE's performance with different numbers of tier-3 subclasses to ACE's performance with Take the Best, which can be thought of as a "one heuristic to a subclass" partition.

Fast and frugal reasoning has been shown here to have a significant impact on an already competent CSP solver. The premise that attractive options remain attractive as problem solving progresses enables at least this program to solve problems better. Furthermore, the problems we have used here are sufficiently general to suggest that our results have a potentially broad impact. We are optimistic that, in problem areas that tolerate errors, fast and frugal reasoning, as implemented here, can make an important contribution to problem solving.

Acknowledgments

This work was supported in part by NSF IIS-0328743 and by PSC-CUNY. Thanks for their support in this work go to Gene Freuder, Anton Morozov, Rick Wallace, CUNY's ACE study group, and the Cork Constraint Computation Centre, supported by Enterprise Ireland and Science Foundation Ireland.

References

Bessière, C., & Régin, J.-C. (2001). Refining the basic constraint propagation algorithm. *JFPLC*, 1-13.

Biswas, G., Goldman, S., Fisher, D., Bhuva, B., & Glewwe, G. (1995). Assessing Design Activity in Complex CMOS Circuit Design. In P. Nichols & S. Chipman & R. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 167-188). Hillsdale, NJ: Lawrence Erlbaum.

Crowley, K., & Siegler, R. S. (1993). Flexible Strategy Use in Young Children's Tic-Tac-Toe. *Cognitive Science*, 17(4), 531-561.

Epstein, S. L. (1994). For the Right Reasons: The FORR Architecture for Learning in a Skill Domain. *Cognitive Science*, 18(3), 479-511.

Epstein, S. L. (1998). Pragmatic Navigation: Reactivity, Heuristics, and Search. *Artificial Intelligence*, 100(1-2), 275-322.

Epstein, S. L., Freuder, E. C., Wallace, R., Morozov, A., & Samuels, B. (2002). The Adaptive Constraint Engine. In P. Van Hentenryck (Ed.), *Proceedings of CP2002* (Vol. LNCS 2470, pp. 525-540). Berlin: Springer Verlag.

Gent, I. E., MacIntyre, E., Prosser, P. and Walsh, T. (1996). The Constrainedness of Search. *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (pp. 246-252).

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review*, 103(4), 650-669.

Gigerenzer, G., & Selten, R. (2001). *Bounded Rationality: The Adaptive Toolbox.* MA: MIT Press.

Gigerenzer, G., Todd, P. M. & The ABC Research Group (1999). *Simple Heuristics that Make Us Smart*. NY: Oxford University Press.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of Ecological Rationality: The Recognition Heuristic. *Psychological Review*, 109(1), 75-90.

Klein, G. S., & Calderwood, R. (1991). Decision Models: Some Lessons from the Field. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(5), 1018-1026.

Kojima, T., & Yoshikawa, A. (1998). A Two-Step Model of Pattern Acquisition: Application to Tsume-Go. *Proceedings of the First International Conference on Computers and Games.*

Mackworth, A. K. (1977). Consistency in Networks of Relations. *Artificial Intelligence*, 8, 99-118.

Novick, L. R., & Coté, N. (1992). The Nature of Expertise in Anagram Solution. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Bloomington, IN.

Oppenheimer, D. M. (2003). Not so Fast! (and not so Frugal!): Rethinking the Recognition Heuristic. *Cognition*, 90, B1-B9.

Sabin, D., & Freuder, E. C. (1994). Contradicting Conventional Wisdom in Constraint Satisfaction. *Proceedings of the Eleventh European Conference on Artificial Intelligence*, Amsterdam.

Schraagen, J. M. (1993). How Experts Solve a Novel Problem in Experimental Design. *Cognitive Science*, 17(2), 285-309.

Simon, H. A. (1981). *The Sciences of the Artificial* (second ed.). Cambridge, MA: MIT Press.

Application of a Novel Neural Approach to 3D Gaze Tracking: Vergence Eye-Movements in Autostereograms

Kai Essig[1], Marc Pomplun[2] and Helge Ritter[1]

[1]Neuroinformatics Group, Faculty of Technology, Bielefeld University,
P.O.-Box 10 01 31, 33501 Bielefeld, Germany
Email: (kessig, helge)@techfak.uni-bielefeld.de

[2]Department of Computer Science, University of Massachusetts at Boston,
100 Morrissey Boulevard, Boston, MA 02125-3393, USA
Email: marc@cs.umb.edu

Abstract

Vergence eye-movements occur not only in real environments, but also in virtual 3D environments. Autostereograms can cover large visual angles without requiring vergence beyond natural parameters and are thus well-suited for the investigation of vergence movements in virtual 3D images. We developed an anaglyph-based 3D calibration procedure and used a *parametrized self-organizing map* (PSOM) to approximate the 3D gaze point from a subject's binocular eye-movement data. Besides analyzing the general pattern of vergence eye-movements in autostereogram images, the present study examined the influence of image granularity on these movements. Unlike previous research on random-dot stereograms, we found substantially overshooting convergence eye-movements, especially for medium granularities. Moreover, divergence movements were completed more quickly for coarse than for fine granularities. Results are discussed in the context of granularity effects on autostereogram perception and the dissociation between convergence and divergence eye-movements.

Vergence Eye-Movements in Autostereograms

In everyday life, we employ vergence eye-movements to successively fixate objects at different distances. It is interesting to note that these movements are also produced in virtual 3D environments. Some work has been done on the analysis of vergence movements occurring while viewing *Random-Dot Stereograms* (*RDS*). For *RDS* two slightly different images for the left and the right eye seen through a special device, a so-called stereoscope, lead to the perception of 3D information. Mowforth, Mayhew and Frisby (1981) found that *RDS* are perceptually filtered by spatial frequencies. Low frequencies enable the 3D perception of *RDS* with larger disparities and lead to faster vergence movements than do high frequencies. Furthermore, vergence velocities were found to be faster for convergent than for divergent eye movements.

In another study, Rashbass and Westheimer (1961) found reaction times of about 160 ms for both convergent and divergent eye movements in response to suddenly introduced target vergence changes. The authors reported start velocity varying with stimulus amplitude, and vergence reaching asymptotically its final level after approximately 800 msec without any overshoots or a prolonged period of oscillations. They also found turnabouts in the response before the target vergence change reaches zero, showing the anticipatory behavior of the eye-vergence system.

For our experiment, however, instead of *RDS* we used *Single Image Random Dot Stereograms (SIRDS)*, also called *autostereograms*. The difference between *RDS* and *SIRDS* is that in *SIRDS* the two slightly different images of the *RDS* are combined into a single image. No stereoscope is necessary to perceive the 3D information in a *SIRDS*. To achieve the 3D perception of a *SIRDS*, observers have to combine the information of the two images for the left and the right eye to one depth perception by focusing a point before (*cross-eyed method*) or behind (*wide-eyed method*) the image plane, whereby our preliminary studies (Essig, 1998) demonstrated that the later method leads to a more stable perception of the depth information and was therefore used for the reported experiments. Once the observers perceive the 3D scene, they may move their gaze to any position in the image without losing the 3D impression. Most important in the present context, *SIRDS* can cover large visual angles without requiring vergence beyond natural parameters, which makes them appropriate stimuli for studying vergence movements in virtual 3D images.

In our natural environment, and when looking at *RDS*, we move our eyes inward (*convergence*) to inspect near objects and outward (*divergence*) to inspect distant objects. An obvious question is: Do these vergence movements also occur during the observation of autostereogram images? Belopolskii and Logvinenko (1994) found that vergence movements are not *necessary* to get the 3D perception of an autostereogram. However, in recent experiments (Essig, 1998) we found that in fact vergence eye-movements are executed during the examination of autostereogram images. These results encouraged us to conduct further experiments investigating stereogram parameters that are likely to influence vergence movements.

A distinguishing feature of autostereogram images is their grain size (granularity). Figure 1 shows examples for autostereogram images with small (upper panel) and big grain sizes (lower panel). The reported experiment used autostereogram images with different grain sizes to systematically investigate the influence of this parameter on the vergence movements. According to Mowforth et al. (1981), overshoots and oscillations should be more likely for autostereogram images with coarser granularities, i.e. lower spatial frequencies, which, however, is in-

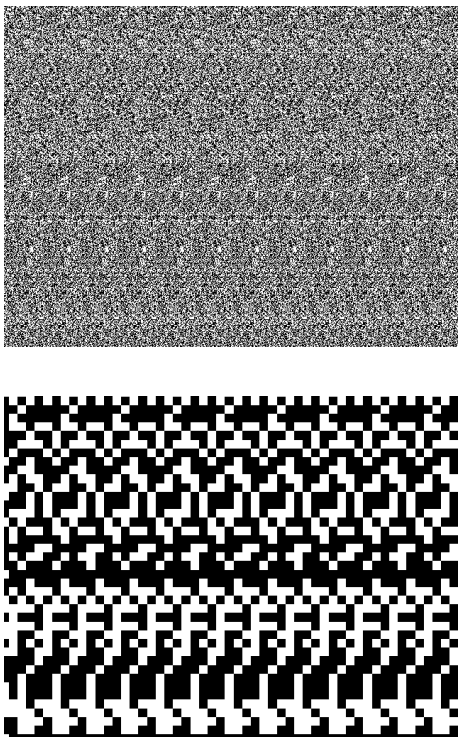


Figure 1: Autostereogram images with granularities 1 (upper panel) and 12 (lower panel).

consistent with the findings of Rashbass and Westheimer (1961). For a clarifying analysis of this issue, we measured vergence eye-movements with a modern binocular eye tracker with a sampling rate of 250 Hertz. For increased precision of vergence and 3D gaze-position measurement, we developed and applied a novel neural-net based calibration interface, which is described in the following section.

A Neural Approach to 3D Gaze-Point Calculation

During the examination of autostereograms shown on the computer screen, the intersection of the viewing axes is in general in front of or behind the screen plane, depending on the 3D gaze position. As a consequence, the correct coordinates for the actual gaze position are different from the screen coordinates which the eye tracker provides, because the system uses a 2D calibration to calculate the relation between the pupil positions and the gaze point on the screen. We tackled this problem by developing a 3D calibration that precedes the experiment and uses a *parametrized self-organizing map* (PSOM) to approximate the 3D gaze point from a subject's binocular eye-movement data.

Since the experimental conditions change from subject to subject (e.g. subjects have physical differences, their gaze characteristics are individually different, and the eye-tracker setup varies across sessions), it is clearly advantageous to use a method which is able to "learn" these individual parameters. Additionally, we have to take into consideration that the mapping from pupil to screen coordinates is non-linear. Hence neural nets are suitable for the solution of these problems, because they are able to "learn" nonlinear functions. Kohonen (1990) developed the so-called *self-organizing maps* (SOMs), which could learn the correct mapping from the pupil to the screen coordinates. However, SOMs have two disadvantages: They supply only the position of the most stimulated neuron in a "neuron lattice" instead of a continuous output, and they usually require thousands of training examples. Therefore, we use a variant of SOMs, namely a so-called *parametrized self-organizing map* (PSOM) (Ritter, 1993). This variant does not have the disadvantages mentioned above, because it provides the demanded continuous output and gets only some selected input/output pairs as parameters, i.e. the coordinates of the calibration points and the related positions of the pupils measured by the eye tracker.

A 2D version of the PSOM was already used in (Pomplun, Velichkovsky & Ritter, 1994), reducing the average calibration error to about 30% of its previous value. In opposition, our "new" PSOM does not only enhance the accuracy of measurement, it also approximates the subject's 3D gaze position from the two 2D coordinates of the eye tracker system.

A PSOM can be considered as a recursive neural net that realizes a distinct, mostly multi-dimensional projection **f**. The input data of this projection consist of the "correct" coordinates of 27 calibration points $k \in \mathbf{A}$ (standardized in the interval from 0 to 2 by the PSOM), where $\mathbf{A} = \{k_{xyz} | k_{xyz} = x\hat{e_x} + y\hat{e_y} + z\hat{e_z}; x, y, z = 0...2\}$

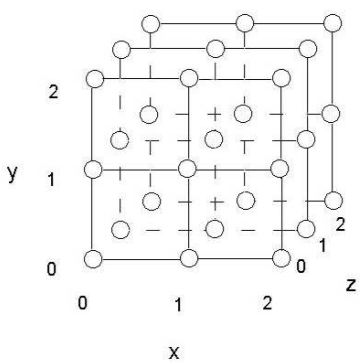


Figure 2: The calibration points in the virtual 3 x 3 x 3 cuboid. For the PSOM the coordinates are standardized in such a way that they only have the values 0, 1, and 2.

(see Figure 2) arranged in a 3 x 3 x 3 grid and presented to the subjects during the calibration procedure. In order to create a virtual 3D perception, the calibration grid was presented as an anaglyph image, viewed through red-green glasses. The calibration points were drawn on the planes of a cuboid, where the planes before and behind the screen plane formed its surfaces. This made it perceptually easier for the subjects to locate the stimuli on one of the three planes of the cuboid. For each of these

27 calibration markers, we have the associated gaze coordinates measured by the eye tracker and the x-divergence between the gaze positions of the left and right eye on the screen calculated from this data. Thus, the reference vector is $w_k = (x_{plk}, y_{plk}, x_{prk}, y_{prk}, x_{dk}) \in \boldsymbol{R}^5$. x_{plk} is the x-coordinate for the left eye, and y_{plk} the corresponding y-coordinate for the left eye belonging to the vector w_k. For the right eye the corresponding values are x_{prk} and y_{prk}. The fifth element of w_k results from $x_{dk} := x_{prk} - x_{plk}$.

We introduce the divergence (x_{dk}) as the fifth dimension of w_k because the z-coordinate of the 3D gaze-position mainly depends on this divergence. Since the differences in the divergence are smaller than those in the x- and y-directions, the divergence has to be weighted by a specific factor. This method leads to a faster termination of the PSOM-calculations. Hence, every fixation point $k \in \mathbf{A}$ is associated with a corresponding reference vector $w_k \in \boldsymbol{R}^5$. With the reference vectors w_k we could already construct a SOM which could enable the projection from the 2D coordinates of the system to the "correct" 3D coordinates. This SOM would only enable a crude approximation to the real 3D coordinates, because it could only choose one of the calibration points as a possible output. In the present context, however, we are looking for a projection which can interpolate between the vectors w_k. This projection can be done by a PSOM.

The desired interpolation function $\mathbf{f}$(s) can be constructed from the superposition of a suitable number of simpler basis functions H(.,.) as follows:

$$\mathbf{f}(s) = \sum_{k \in A} H(s,k) w_k \tag{1}$$

The values of the basis functions are within the interval [0,1] depending on different values s, so that the coordinates of a calibration point, which is close to the desired gaze position, is weighted strongly (near 1) and points which are far away are weighted weakly (near 0). The basis functions $H : \boldsymbol{R}^3 \times A \rightarrow \boldsymbol{R}$ have to comply with the requirement

$$H(s,k) = \delta_{s,k} \qquad \forall \;\; s,k \in A \tag{2}$$

where δ represents the Kronecker symbol. It is defined as: $\delta_{ij} = \begin{cases} 1 & : \quad i = j \\ 0 & : \quad i \neq j \end{cases} \quad \forall \;\; i,j = 0,1,2$
(in this special case). This ensures that

$$\mathbf{f}(s) = w_s \qquad \forall \;\; s \in A$$

Thus, it is guaranteed that the interpolation function passes through the given points. But how can we choose suitable functions H that obey equation (2), that are smooth and simple to handle?

One convenient solution is to make a product ansatz and to derive the suitable function by combining three 1D functions (one in each case for every coordinate direction x, y, and z). The new 1D functions must then have the property $H^{(1)} : \boldsymbol{R} \times \{0,1,2\} \rightarrow \boldsymbol{R}$:

$$H^{(1)}(q,n) = \delta_{q,n} \quad \forall \;\; q \in \boldsymbol{R}, n \in \{0,1,2\} \tag{3}$$

Because n has only three possible values (0, 1, and 2), it is sufficient to choose an account of three basis functions $\boldsymbol{R} \rightarrow \boldsymbol{R}$. For this purpose, polynomials of second degree are especially suitable, because they have no redundant degrees of freedom.

Now we have built a function which projects 3D coordinates onto 2D gaze-positions for both eyes. Our final aim is to find a function which does just the opposite though, namely to approximate the 3D gaze-position of the subject from the system's 2D measurements. Therefore, we have to calculate the inverse function $\mathbf{f}^{-1}$ of $\mathbf{f}$. The non-linearity of $\mathbf{f}$ forces us to use a numerical procedure. Thus we have to create an error function $E(s)$, which is defined as:

$$E(s) = \frac{1}{2}(\mathbf{f}(s) - \mathbf{f}_{et})^2 \tag{4}$$

i.e. the deviation of the pupil coordinates provided by the eye tracker ($\mathbf{f_{et}}$) from the eye tracker data calculated by the PSOM $\mathbf{f}(\mathbf{s})$ for the actually assigned two screen points.

If this difference exceeds a specified threshold, the coordinates of these points are modified in an iterative gradient-descent procedure until the results of Equation (1) closely approximate the actual eye tracker data provided by the system:

$$s(t+1) = s(t) - \epsilon \cdot \frac{\delta E(s)}{\delta s} \quad , \text{with } \; \epsilon > 0 \tag{5}$$

This means that the iteration process stops if $E(s(t))$ falls below the threshold, which we should adapt to the screen resolution. In this way, an exact 3D gaze position of the subject during the examination of a 3D stimulus is assigned to the 2D eye tracker data for the left and right eye.

We conducted an experiment in which the accuracy of the PSOM gaze-point calculation is compared to one provided by a geometrical solution. Subjects had to visually track a dot that appeared at positions of a 4 x 4 x 4 grid in 3D space in a random sequence. In the geometrical method the equations for the right and left visual axes are calculated on the basis of both the measured gaze-positions on the screen and the subjects' (assumed) constant head position. The gaze point is determined as the point where the visual axes are closest to one another. It results as the center of the shortest straight link between the visual axes.

The results show that the neural net calculates gaze positions from the eye tracker data which are nearly 46% closer to the actual gaze positions than those of the geometrical solution. The average total error for all subjects, separated for the individual coordinates, is shown in Table 1 (the values for the geometrical method and the neural net show the average total error and the standard error).

It is obvious that the z-error is always higher than the $x-$ and $y-$errors, because the z-coordinate is much more sensitive to small changes in the binocular gaze position than the $x-$ and $y-$coordinates. We also found that the measurement errors decrease from the back to the front

Table 1: Average total error for individual coordinates.

	both methods	**geometrical method**	**neural net**
coordinate	average total error	average total error	average total error
x	$0.97cm$	$1.41cm \pm 0.04cm$	$0.52cm \pm 0.05cm$
y	$1.03cm$	$1.24cm \pm 0.05cm$	$0.82cm \pm 0.05cm$
z	$4.16cm$	$5.79cm \pm 0.36cm$	$2.53cm \pm 0.11cm$

plane. The neural net compensates the errors in the back plane better than the geometrical method, because small changes in the gaze position at the back plane lead to severe inaccuracies. Obviously, the precision of our novel method of 3D gaze-position measurement provided an appropriate basis for the autostereogram study.

The Autostereogram Experiment

Method

Subjects. Eight paid subjects (1 female, 7 male), all of them were students at the University of Bielefeld, participated in this experiment. They had normal or corrected-to-normal visual acuity and were experienced in viewing autostereogram images.

Stimuli. The stimuli in this experiment were autostereogram images whose depth impression arose from two horizontally divided half planes, which were recognized at different virtual distances. The autostereogram images used in the experiment varied in their granularities. Using the autostereogram generation program *rdsgen V1.2b* by Frederic N. Feucht[1], we produced different autostereogram images in which one dot of the autostereogram image corresponded to one, two, four, eight, or twelve screen pixels in height and width. The subjects always got the spatial perception that the lower plane was nearer than the upper one. The depth difference between the two planes corresponded to a vergence angle difference of 0.859^o.

By using *rdsgen* we created two autostereogram images for each of the granularities 1, 2, 4, 8, and 12[2]. These 10 autostereogram images were presented to the subjects in a random order on a 20-inch screen and a spatial resolution of 640x480 pixels. The images consisted of black and blue (luminance: 24 cd/m^2) points, where one pixel corresponded to 4.17 min arc. The distance from the subject to the screen was 50 cm.

[1] http://www.bcc.cc.nc.us/graphics/g2c.html.

[2] We first changed the distance between the repeating patterns of the autostereogram image and magnified it with the corresponding "enlargement factor" (grain size) to keep the distance between the repeating patterns (and therefore the depth impression) constant.

Apparatus. We used an SMI EyeLink eye tracker for the experiments. This system employs a headset with two cameras to enable binocular eye-movement recording. Further features of the EyeLink system are a high sampling rate of 250 Hz and an average on-screen gaze position error between 0.5^o and 1.0^o.

Procedure. Prior to the experiment, subjects were presented with autostereogram images without using the eye tracker. This way the subjects could practice the perception of the depth information. A lamp fixed behind the subjects to create reflections on the screen helped the subjects to fixate a virtual point behind the screen plane, facilitating the 3D perception of the autostereogram image. When the subject got the 3D impression, the light from behind was switched off.

At the beginning of every trial the light was switched on as soon as the autostereogram image appeared on the screen. When the subject got the 3D impression of the image, the experimenter switched off the light and started the eye-movement recording. It was the subjects' task to change their view from the front plane to the back plane and vice versa (every few seconds). Between these eye movements subjects kept their view on one plane for a while before changing their view to the other one. The vergence movements were recorded over a duration of 1 minute per image. If subjects "lost" the 3D impression during the recording interval, they had to press the right mouse button, recover the 3D impression, and press the left button to continue recording. Ten autostereogram images were shown to the subject in a random order so that every granularity appeared twice.

Data Analysis. In general, the vergence angle α is used to make statements about a subject's vergence. The bigger the vergence angle is, the stronger the eyes converge, i.e. the nearer the (virtual) surface fixated by the subject. The time course of vergence movements was the most important data to be analyzed in this experiment, especially during the period right before and after subjects changed their gaze point from one plane to another. We distinguished between eye movements from the near to the far plane and vice versa. For the evaluation process we summarized the data of all subjects and calculated the arithmetic mean and the standard error of the vergence angle as a function of time relative to gaze transitions between the two depth planes.

There were two criteria for the presence of a gaze transition between planes: First, there had to be a change in the measured y-value that indicated a crossing of the horizontal boundary. Second, any such change from plane A to plane B had to be preceded by a contiguous sequence of 50 measured gaze positions (200 msec) on plane A, and succeeded by a contiguous sequence of 250 gaze positions (1000 ms) on plane B. To account for a small fraction of measurement errors, we allowed a maximum of 4 violations of these conditions among the 300 measurements. After detecting a plane transition, the time course of vergence from 200 ms before to 1000 ms after the transition was calculated.

Results and Discussion

Our first observation was that subjects had problems to achieve a stable 3D perception of autostereogram images with large grain size (granularities 8 and 12). Once achieved, it was difficult for them to maintain the stable depth impression when changing their gaze point from one plane to the other. Subjects classified the autostereogram images with the granularities 2 and 4 as most pleasant to view. Figure 3 shows the temporal progress of the convergence and divergence movements during the examination of autostereogram images across granularities. The vergence angles are standardized, i.e. 0 corresponds to the back and 1 to the front plane. In the figures the value 0 on the time axis signals the time of plane transition. Each panel shows the vergence data from 200 ms before to 1 s after an eye movement between the two planes. Figure 3 also illustrates that, for all granularities, the convergence movements are obviously faster than the divergence movements. This is in line with the observations in Mowforth, Mayhew and Frisby (1981). The final values are reached after a period of about 800 ms, as stated in Rashbass and Westheimer (1961). As can clearly be seen, both convergence and divergence movements already start before the onset of an eye movement between planes, suggesting that the imagination of the target distance is already sufficient for the execution of vergence eye-movements (proximal vergence). Similar results were already obtained by Yarbus (1967).

The high overshoot of the convergence movements (Figure 3, upper panel) for the granularities 2 and 4 is very obvious. Both show an overshoot of 1.8 and approach the end value at nearly 900 ms. The similarity between those vergence movements is in line with the subjects reporting that it is easier to get and stabilize the 3D illusion of these two types of autostereogram images. One reason could be that the perception of the depth planes is particularly stable for these granularities. Subjects can perceive the near plane very well, even if they fixate a point on the back plane. As a consequence, the visual system might perform a particularly fast, overshooting vergence movement towards an angle that would project the pattern of the near plane to corresponding points on the retina. This mechanism allows us a fast focussing of near or approaching objects, which is an important capability to react in dangerous situations.

The vergence movements during the observation of the autostereogram image with granularity 1 are similar to those for granularities 2 and 4. It is obvious, though, that the overshoot of the convergence movements for granularity 1 is smaller (1.5) than for granularities 2 and 4 (granularity 1 seems too fine for a stable extrafoveal perception). Also the approaching to the end value is shorter (nearly 750 ms instead of approx. 900 ms). These results further support the finding that vergence movements can clearly be driven by relatively high spatial frequencies (fine granularities), as stated in Mowforth, Mayhew and Frisby (1981).

The data for the coarse granularities 8 and 12 are clearly different from those for 1, 2, and 4. Hence it

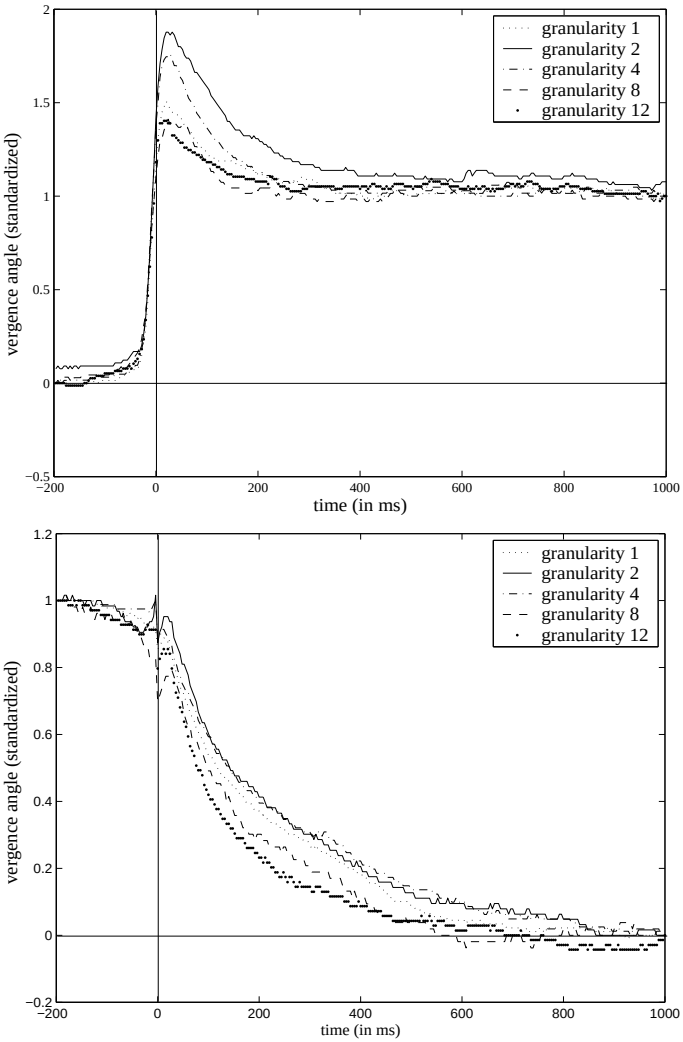


Figure 3: Convergence (upper panel) and divergence movements (lower panel) occurring during the examination of autostereogram images with different granularities.

is much more difficult for the subjects to perceive both planes simultaneously in those autostereogram images, which is consistent with the subjects' impressions. An important point is that the overshoot of the convergence movements is substantially weaker (1.4) than for those occurring in the images with finer granularities. The standard error for the convergence as well as for the divergence movement is high. This can be interpreted as indicating an instable perception of the depth planes. Furthermore, the progress of the divergence movements is faster and finishes sooner than 600 ms after the gaze shift.

The finding of overshoots for convergence movements is inconsistent with the results obtained by Rashbass and Westheimer (1961). It is possible that either the apparatus used in their study did not allow the detection of these overshoots, or that vergence eye-movements in *RDS* – in contrast to those in *SIRDS* – do not show

overshoots. Rashbass and Westheimer describe the vergence system as a damped system, because it does not show any oscillations. In the present study, however, we found overshoots, but no following undershoots or any kind of oscillations. It seems therefore correct to speak of a damped system but with a strong tendency towards overshoots for convergence movements in *SIRDS*.

With regard to divergence movements (Figure 3, lower panel), we found faster approximation of the target angle for the autostereogram images with coarser granularities than for those with finer granularities. This is just the opposite of the results we got for the convergence movements. A possible explanation is a fundamental difference in nature between convergence and divergence movements, as can be seen from Figure 3. Convergence movements are fast and impulse-like, whereas divergence movements are slower and smoother. From an evolutionary standpoint, since distant or vanishing objects are less dangerous, there is no need for fast divergence eye-movement mechanisms. Instead, this divergence resembles more a relaxation process. And this relaxation might be facilitated by easily losing the perception of the near plane. In other words, for the finer granularities 1, 2, and 4, the continuous stable perception of the front plane might impede the subjects' effort to focus attention – and consequently vergence – on the far plane.

All in all, some findings for the *RDS* were confirmed in our experiments with *SIRDS*, in particular that convergence is faster than divergence, that the vergence mechanism can clearly be driven by relatively high spatial frequencies, and that vergence response occurs even before the disparity reaches zero. Inconsistent with previous studies, however, our experiment demonstrates substantial differences in vergence movements between *SIRDS* and previously investigated *RDS*, especially with regard to the overshoot for convergence movements. This difference, however, might have been caused by the possibly low resolution of the apparatus used in the *RDS* study (Rashbass & Westheimer, 1961).

Furthermore, our experiments show a clear influence of granularity on vergence movements. The pattern of influence seems to be more complex than described in Mowforth, Mayhew and Frisby (1981), because, according to our results, lower frequencies (coarser granularities) do not always lead to faster vergence movements. One reason might be the more difficult simultaneous perception of different depth planes in the stimuli of very low frequencies. At any rate, these complex granularity effects on vergence eye-movements in autostereogram images indicate that there may be rather non-intuitive factors that can have a significant impact on the coordination of both eyes. Future research will have to address further issues, such as the timing of coordinated binocular saccades and its dependence on various pattern features. This is a research area that so far has been explored very little, and the rather recent possibility of fast and accurate binocular eye tracking, combined with appropriate techniques for 3D gaze measurement like the approach presented here, will greatly contribute to its advancement.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 360, *Situierte Künstliche Kommunikatoren*).

References

Essig, K. (1998). *Messung von binokularen Augenbewegungen in realen und virtuellen 3D-Szenarien.* Master Thesis, Faculty of Technology, Bielefeld University.

Julesz, B. (1971). *Foundations of Cyclopean Perception.* Chicago: University of Chicago Press.

Kohonen, T. (1990). The Self-Organizing Map. *Proceedings of IEEE*, *78*, 1464–1480.

Littmann, E. (1989). Vergenz und Kontrast. *Optometrie*, *3*, 15–30.

Logvinenko, A. D. & Belopolskii, V. L. (1994). Convergence as a Cue for Distance. *Perception*, *23*, 207–217.

Mowforth, P., Mayhew, J. E. W. & Frisby J. P. (1981). Vergence Eye Movements Made in Response to Spatial-Frequency-Filtered Random-Dot Stereograms. *Perception*, *10*, 299–304.

Ning Qian (1997). Binocular Disparity and the Perception of Depth. *Neuron*, *18*, 359-368.

Pomplun, M., Velichkovsky, B.M. & Ritter, H. (1994). An artificial neural network for high precision eye movement tracking. In Nebel, B. & Dreschler-Fischer, L. (Eds.), *Lecture notes in artificial intelligence: AI-94 Proceedings*, 63-69. Berlin: Springer Verlag.

Rashbass C. & Westheimer G. (1961). Disjunctive Eye Movements. *J. Physiol.*, *159*, 339-360.

Reading, R.W. (1983). *Binocular Vision. Foundations and Applications.* Boston: Butterworths.

Ritter, H. (1993). Parametrized Self-Organizing Maps. *ICANN93-Proceedings*, 568–577, Berlin: Springer.

Schor, C. M. & Ciuffreda, K. J. (1983). *Vergence Eye Movements: Basic and Clinical Aspects.* Boston: Butterworths.

Yarbus, A. (1967). *Eye Movements and Vision.* New York: Plenum Press.

Design, Adaptation and Convention: The Emergence of Higher Order Graphical Representations

Nicolas Fay (nfay@atr.jp)
ATR Media Information Science Labs, 2-2-2 Hikaridai, Keihanna Science City,
Kyoto 619-0228, Japan

Simon Garrod (simon@psy.gla.ac.uk)
Tracy MacLeod (tracym@psy.gla.ac.uk)
Department of Psychology, University of Glasgow,
Glasgow, G12 8QB, Scotland

John Lee (j.lee@ed.ac.uk)
Jon Oberlander (j.oberlander@ed.ac.uk)
HCRC, University of Edinburgh, 2 Buccleuch Place,
Edinburgh, EH8 9LW, Scotland

Abstract

To study the development of graphical conventions we had members of a simulated community play a series of graphical interaction games with partners drawn from the same pool (Experiment 1). Once the community was established, a conventional graphical referring scheme emerged that facilitated high levels of semantic coordination, with reduced communicative effort. Next, a forced choice reaction time study (Experiment 2) demonstrated that the graphical conventions developed in the simulated community offer distinct processing advantages when compared with those developed by isolated pairs (i.e. participants who always interact with the same partner). This is interpreted as evidence that the graphical conventions that evolve within a closed community constitute higher order cognitions, the whole being greater than the sum of its parts.

Background

Vygotsky (1981) claims that higher order cognition is a product of social interaction, that novel structures emerge as a consequence of interpersonal, as opposed to intrapersonal, communication. Hutchins (1995) shares this view, arguing that higher order cognition is a cultural product, a consequence of interaction (human-environment and human-human) that is distributed across time and space. According to Hutchins, higher order cognitions emerge from "an adaptive process that accumulates partial solutions to frequently encountered problems" (p.354). Lewis (1969, 1975) defines conventions in a related way, as arising from situations where a community faces the recurrent problem of coordination.

If we agree that conventions are cultural products, should we accept that they represent higher order cognitions? Using Chinese characters as an example (Figure 1), we argue that conventions are culturally evolved higher order cognitions.

Over several thousand years the original Chinese character that represents mountain (left) has evolved into its current, less complex, form (right). We argue that this change is not arbitrary; it is a result of global coordination that took place over time and space, culminating in a refined, conventional form that promotes rapid communication with reduced effort. This is an example of an evolutionary process where the whole is greater than the sum of its parts.

Figure 1. The changing form of the Chinese character that represents mountain (Vacarri & Vacarri, 1961; cited in Arbib, in press)

Having partners collaborate on a graphical referential communication task, Fay, Garrod, Lee and Oberlander (2003) studied the influence of interaction upon representational form. The task requires pairs of participants to graphically communicate a series of recurring concepts. Figure 2 illustrates the changing representation of the concept 'Clint Eastwood' over 6 games, where partners' drawing and identifying roles alternated from game to game.

What is initially a designed, iconic representation of Clint Eastwood, develops, through a process of adaptation and entrainment, into a simplified, symbolic form (an arrow pointing East). Although there are obvious similarities between this process and the evolution of Chinese characters, the derived representation of Clint Eastwood does not constitute a convention in Lewis' terms. According to Lewis, a convention must be common knowledge within the

wider community. At best, Figure 2 illustrates the development of a 'local' convention.

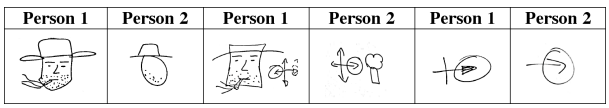


Figure 2. Partners' changing representation of the concept 'Clint Eastwood' over 6 games

Garrod & Doherty (1994) distinguish this 'local' process from the 'global' coordination process that produces conventions. To study the development of linguistic conventions, Garrod et. al. had members of a simulated community play a series of computerized maze games with partners drawn from the same pool. After several games, community members demonstrated higher maze description scheme convergence and more closely coordinated linguistic entrainment when compared with interacting pairs, or participants drawn from a non-community (i.e. interacting partners not drawn from the same pool).

This effect was interpreted as indicating the establishment of the community, and the development of a robust referring convention. Garrod et. al. argue that referring conventions emerge on account of a global coordination constraint based upon two factors; pressure to converge upon the most popular description scheme with each new partner, and the consequent polarization of this scheme throughout the community. This is contrasted with the local coordination process evident among pair and non-community members, a process based upon the less stable heuristics of salience (pressure to choose the most salient description scheme) and precedence (pressure to choose the previously used scheme).

In much the same way that Garrod et. al. studied the development of linguistic conventions, we investigate the development of graphical conventions (Experiment 1). Furthermore, we demonstrate that these cultural products represent higher order cognitions (Experiment 2).

Experiment 1

Experiment 1 investigates the development of graphical conventions within a simulated community of drawers.

Task and Procedure

Fay et. al.'s (2003) graphical referential communication task was employed. This task requires pairs of participants to communicate a series of concepts using only graphical means. Like the game 'Pictionary', participants are not allowed to speak or use text in their drawings. Concepts are drawn from a list of 16 items that are known to both partners. The list was designed to contain a set of graphically confusable concepts (theatre, art gallery, museum, parliament, Robert De Niro, Arnold Schwarzenegger, Clint Eastwood, drama, soap opera, cartoon, television, computer monitor, microwave, loud, homesick, poverty).

Participants played six consecutive games, using the same item set, with their partner. On each game the director, or drawer, depicted the first 12 items from an ordered list (12 items plus 4 distracters) such that their partner, the matcher, could identify each drawing from their unordered list. Item order was randomized on each game. Partners' roles, as drawer or matcher, alternated from game to game, although participants were permitted to draw in either role. Drawing took place on a standard whiteboard. The completed drawings were recorded on digital camera for later analysis.

Subjects

The community was composed of 8 undergraduate students who were paid to participate in the study.

Community Design

A simulated community was created through a series of one-to-one interactions among partners drawn from the same pool. Over 7 rounds, each participant interacted with the other members of the community. As discussed, participants completed a series of 6 consecutive graphical interaction games with each partner. The structure of the community is illustrated in Figure 3.

It was so designed that community could first establish itself at Round 4. This was the earliest point the community could converge upon a conventional graphical description scheme. For example, if person 1 influences person 2 (Round 1), person 2 influences person 3 (Round 2), and person 3 influences person 8 (Round 3), person 1 and 8 will share some interactive history upon meeting in Round 4. Thus, Rounds 1-3 represent pre convergence games, whereas Rounds 4-7 represent post convergence games.

ROUND	PAIR	PAIR	PAIR	PAIR
1	1&2	3&4	5&6	7&8
2	1&4	3&2	5&8	7&6
3	1&6	3&8	5&2	7&4
4	1&8	3&6	5&4	7&2
5	1&3	2&4	5&7	6&8
6	1&5	2&6	3&7	4&8
7	1&7	2&8	3&5	4&6

Figure 3. Structure of the simulated community

Three independent measures were employed to determine the establishment of a conventional graphical referring scheme: identification accuracy (i.e. participants' ability to more successfully identify conventional graphical representations), graphical complexity (i.e. the reduced effort required to negotiate

the meaning of conventional graphical representations) and graphical convergence (i.e. the greater uniformity of a conventional graphical description scheme).

Results

Identification Accuracy. Figure 4 details the identification rate (proportion of items correctly identified by matchers) of pre convergence (Rounds 1-3) and post convergence representations (Rounds 4-7) over the six games played by each pair. In pre convergence rounds there is a steady improvement in identification accuracy from games 1 to 3. After this identification rates reach ceiling level. In contrast, post convergence identification rates begin from, and are maintained at, ceiling level across games. Analysis of Variance (ANOVA) confirms these observations.

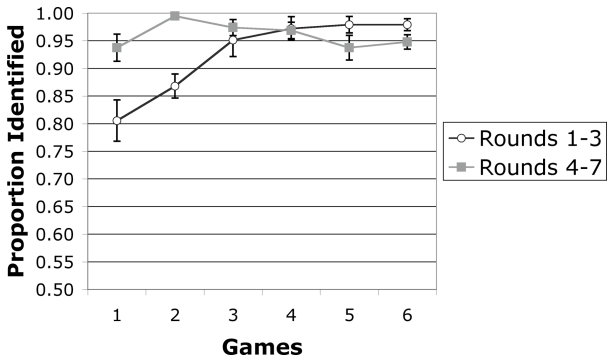


Figure 4. Mean proportion of items correctly identified by matchers in games 1 to 6 during Rounds 1-3 (pre convergence) and 4-7 (post convergence)

Proportion scores were entered into a 2 x 6 ANOVA. Analyses were conducted by subject (F_1) and by item (F_2). By subject tests used a mixed design, treating Game (1 to 6) as a within subject factor and Round (1-3 and 4-7) as between. Item tests used a within subject design. A main effect of Game $F_1(5, 130) = 6.78$, $F_2(5, 55) = 4.51$, and Round $F_1(1, 26) = 8.69$, $F_2(1, 11) = 5.97$, $p<.05$, was qualified by their interaction $F_1(5, 130) = 7.14$, $F_2(5, 55) = 6.68$ (for all results reported $p<.01$ unless otherwise stated). Tests of simple effects corroborate the observations made above; identification accuracy improves in the pre convergence rounds (1-3), $F_1(5, 130) = 11.07$, $F_2(5, 55) = 8.17$, but not in the later post convergence rounds (4-7) where identification rate is maintained at ceiling from game 1 onwards, $Fs<1.48$.

The consistently high identification rate in Rounds 4 to 7 indicates the establishment of the community and the emergence of a robust referring scheme.

Graphical Refinement. Through interaction partners minimize their collaborative effort, stripping away unnecessary graphical information, leaving only the salient properties of the image (Fay et. al., 2003). As a result, what is initially an iconic representation becomes increasingly symbolic (see Figure 2). Comparable graphical refinement was evident in the community members' drawings.

Drawing complexity was measured using the Perimetric Complexity formula developed by Pelli, Burns, Farrell and Moore (accepted with minor revisions) to measure the visual complexity of letters, Complexity = Perimeter^2/Ink. This measure compares favorably with human judgments of drawing complexity (Fay et. al., 2003).

Figure 5 illustrates the mean complexity of drawings made in games 1 to 6 during Rounds 1-3 and 4-7. Mean complexity scores were calculated after the removal of scores 2.5 standard deviations (SD) from the condition mean. Extreme values were replaced with values corresponding to the mean plus or minus 2.5 SDs. Such cases accounted for 2.4% of the data.

In both the pre and post convergence rounds (Rounds 1-3 and 4-7 respectively) the complexity of community members' drawings is reduced across games. However, this effect is more marked in the early, pre convergence rounds. This observation is corroborated by ANOVA.

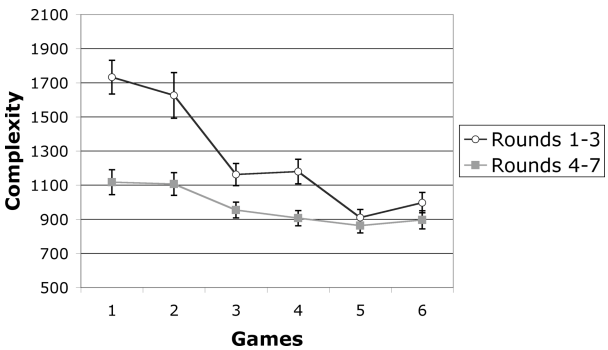


Figure 5. Mean Complexity (Perimeter^2/Ink) of pre and post convergence drawings (Rounds 1-3 and 4-7 respectively) in games 1 to 6

As before, complexity scores were entered into a 2 x 6 ANOVA. This returned a main effect of Game $F_1(5, 130) = 58.57$, $F_2(5, 55) = 34.48$, and Round $F_1(1, 26) = 13.01$, $F_2(1, 11) = 39.11$, that was qualified by their interaction $F_1(5, 130) = 15.35$, $F_2(5, 55) = 12.43$. Tests of simple effects show that graphical complexity is reduced across games in both the pre convergence rounds $F_1(5, 130) = 58.33$, $F_2(5, 55) = 39.34$, and post convergence rounds $F_1(5, 130) = 8.48$, $F_2(5, 55) = 4.24$, (Rounds 1-3 and 4-7 respectively). Between condition differences in drawing complexity in games 1 to 4, $ps<.05$, and the comparable complexity of drawings at games 5 and 6, $Fs<1.07$, explain the interaction.

Results support Garrod and Doherty's (1994) distinction between local and global coordination

processes, signaled by the emergence of a conventional graphical referring scheme during Rounds 4-7. Unlike the initial exchanges in Rounds 1-3, where partners must negotiate a common description scheme, the establishment of a conventional referring scheme from Round 4 requires considerably less local negotiation.

Graphical Convergence. Graphical convergence concerns the degree to which community members' drawings converge, or become more similar, as a consequence of their interaction. To investigate the emergence of a common referring scheme, the similarity of the first drawings of each concept produced by community members at Rounds 1, 4 and 7 (i.e. pre convergence, convergence and post convergence rounds) was compared.

Figures 6, 7 and 8 illustrate the 8 community members' changing representation of the concept 'cartoon' at Rounds 1, 4 and 7 respectively. In addition to the reduction in graphical complexity across rounds, observe the community members' convergence upon a conventional description scheme for cartoon, in this case a Mickey Mouse-like depiction, characterized two large circular ears above the head.

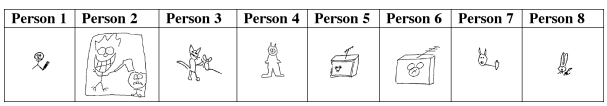


Figure 6. Community members' drawings of the concept 'cartoon' at Round 1

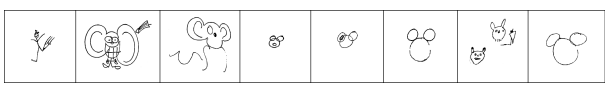


Figure 7. Community members' drawings of the concept 'cartoon' at Round 4

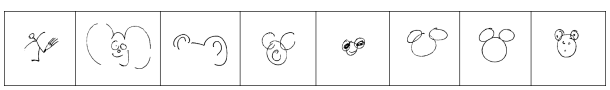


Figure 8. Community members' drawings of the concept 'cartoon' at Round 7

Graphical convergence was analyzed by having 12 subjects, who had no experience of the graphical communication task, rank sets of images in terms of similarity. Each subject individually ranked three sets of 8 images (e.g. Round 1, 4 and 7 drawings of the concept 'cartoon' produced by each of the 8 community members) in terms of similarity. This was done for each of the 12 target items drawn by community members. The set of images thought to be most similar was given a rank of 1; those deemed least similar a rank of 3. The presentation order of item type (e.g. television, cartoon etc.) and round (Rounds 1, 4 or 7) was randomized.

Graphical convergence was measured by calculating the proportion of Round 1, 4 and 7 images ranked as most similar. As can be seen from Figure 9, graphical convergence increased across Rounds. A substantially higher proportion of Round 4 images were ranked as most similar when compared with Round 1 images. In addition, more Round 7 images were ranked as most similar when compared with Round 4 images.

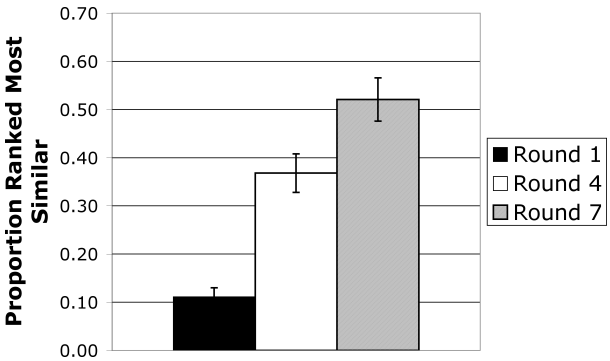


Figure 9. Mean proportion of drawings ranked most similar in Rounds 1, 4 and 7 (pre convergence, convergence and post convergence rounds respectively)

Proportion scores were entered into two repeated measures t-tests. Results were as predicted; Round 4 images were ranked as most similar more often than Round 1 images $t_1(11) = 5.82$, $t_2(11) = 2.91$, $p<.05$. This effect was less clear when Round 4 and Round 7 images were compared; by subject tests revealed a marginal effect $t_1(11) = 1.84$, $p<.10$, whereas there was no effect of round when tested by item, $p>.10$.

Consistent with the previous identification and complexity analyses, convergence tests indicate the emergence of a conventional referring scheme at Round 4. Tests show a large jump in drawing convergence from Rounds 1 to 4 and a smaller, marginally significant, increase in graphical convergence from Rounds 4 to 7. This indicates the establishment of a conventional graphical referring scheme at Round 4, and its continued development in the later rounds.

Experiment 2

Having detailed the emergence of a conventional graphical description scheme in Experiment 1, Experiment 2 demonstrates that these graphical productions constitute higher order cognitions.

Experiment 2 contrasts representations that emerge as a product of design, local coordination and global coordination processes. Pre interaction drawings (i.e. subjects' first drawing of each item) represent pure design, an intrapersonal process whereby the drawer designs a representation to meet the needs of his/her partner. This is consistent with the notion of 'audience

design' (Isaacs & Clark, 1987). Local coordination, or adaptation, is illustrated in the interacting pair's final drawing of Clint Eastwood in Figure 2. In this example partners' drawings serve their local needs. Global coordination, or evolution, is exemplified by the development of a conventional referring scheme that meets the needs of the wider community (see Figure 8).

A forced choice reaction time (RT) experiment was designed to compare the processing efficiency of graphical representations that are a product of design, local coordination and global coordination processes. If graphical conventions constitute higher order cognitions they will provide a processing advantage when compared with designed or locally developed representations.

Task and Procedure

The RT experiment required subjects to make binary judgments regarding a set of learnt images. Community (Experiment 1) and isolated pair images (from Fay et. al., 2003) were used as stimuli, presented on a computer screen using PsyScope (Cohen, MacWhinney, Flatt & Provost, 1993).

24 undergraduate students, who were unfamiliar with the graphical communication experiment, were paid to act as subjects. Prior to taking part in the RT experiment, subjects learnt the identity of 48 images (50% community generated and 50% pair generated). Each image set (community and pair) was composed of 50% pre interaction images (i.e. the first drawing of each item) and 50% post interaction images (i.e. the last drawing of each item). Pre and post interaction drawings (matched by drawer) were sampled quasi-randomly from the community and isolated pair conditions. Although pre interaction drawings were more complex than post interaction drawings, there was no difference between community and isolated pair drawings (Community, *M* pre = 1796, *M* post = 906; Pair, *M* pre = 1669, *M* post = 932). Participants were judged to have learnt the images when each drawing could be identified on three consecutive presentations of the set.

Each experimental trial consisted of the following sequence; a fixation point (a small cross presented in the middle of the screen for 25 msecs), the learnt image (50 msecs), a mask (the screen was blacked out for 25 msecs) and a forced choice (text that either matched or mismatched the presented image, e.g. a cartoon image followed by the text 'drama'). The time required to make a match/mismatch judgment, by key press, was recorded. Subjects completed 384 trials, with each drawing appearing 4 times in each condition.

Results

There was a 4.8% error rate on match/mismatch questions, suggesting that participants had adequately learnt the task materials. Mean RTs were calculated after the removal of times 2.5 SDs from the population mean. These extremes were replaced with values corresponding to the mean plus or minus 2.5 SDs. This accounted for 2.5% of the data.

Mean RTs for matching image and text judgments are shown in Figure 10. As predicted, the community generated graphical conventions (post interaction) were processed more rapidly than the designed (pre interaction) or locally developed pair representations (post interaction). The same pattern is evident in the mismatching text condition, but at a slower response rate (*M* match = 781 msecs; *M* mismatch = 884 msecs).

RTs were entered into a 2 x 2 x 2 ANOVA, treating Communication (Community or Pair), Image (Pre and Post) and Text (Match or Mismatch) as within subject factors. Analyses returned a main effect of Text $F_1(1, 23) = 43.79$, $F_2(1, 11) = 65.26$, indicating subjects' faster response times when the image and text stimuli matched. There was also a reliable interaction between Communication and Image $F_1(1, 23) = 10.08$, $F_2(1, 11) = 3.16$, $p = .10$. Tests of simple effects confirmed that community generated representations were processed more rapidly than those developed in isolated pairs $F_1(1, 23) = 5.63$, $p<.05$, $F_2(1, 11) = 6.34$, $p<.05$. Pre interaction images and post interaction pair representations were processed at a comparable rate, $Fs<1$.

The RT experiment distinguishes between representations that emerge as a product of design, local coordination and global coordination processes. The graphical conventions that evolved within the simulated community, a society composed of the pairwise interactions of its members, outperformed those produced by pairs who interacted in isolation. These graphical productions provide a clear example of higher order cognition, where the whole is greater than the sum of its parts.

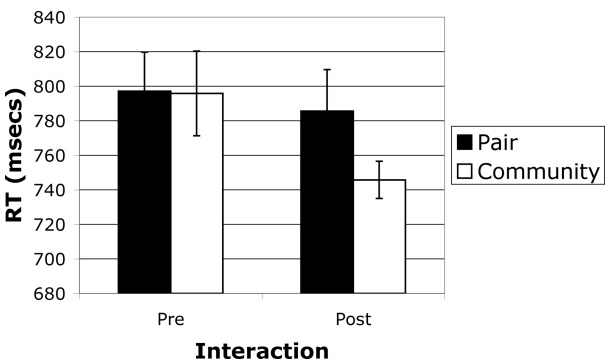


Figure 10. Mean time required to process pre and post interaction community and pair representations in the matching text condition (image and text match)

Discussion

Experiment 1 details the emergence of a conventional graphical description scheme within a simulated community of drawers. Once established, community members exhibit near perfect semantic coordination (identification accuracy) with reduced communicative effort (graphical complexity). This is a consequence of the development of a conventional graphical description scheme (convergence). Experiment 2 distinguishes between graphical productions that develop as a product of design, local coordination and global coordination. The graphical productions that evolved within the simulated community offer substantial processing advantages when compared with designed or locally developed representations. Thus, conventional graphical representations constitute higher order cognitions, the whole being greater than the sum of its parts.

However, there are two potentially confounding factors in Experiment 2. Community members played more games with more partners (42 games; 6 games with each of 7 partners) than pair members (6 games with 1 partner), exposing them to a greater variety of description schemes. Number of games played can be discounted for the simple reason that isolated pairs rapidly negotiate and maintain a locally stable description scheme (Fay et. al., 2003; Garrod & Doherty, 1994). Community members' exposure to a greater number of exemplars is likely to have a profound effect, as people are better able to learn a prototype when exposed to its variants (Posner & Keele, 1968). However, number of exemplars alone is not enough; global coordination is necessary to derive a stable underlying representation (Garrod & Doherty, 1994).

So what is 'better' about the graphical conventions developed in the simulated community? We believe there are two factors at play; iconicity and systematicity. Like the present day Chinese character for mountain (Figure 1), community representations retain a degree of iconicity, i.e. once told what a drawing represents it can be 'seen' as such. Again, like Chinese characters, community drawings exhibit a degree of systematicity that makes them easily differentiable. We believe these factors explain the RT advantage for graphical conventions found in Experiment 2. At present this is pure conjecture. Further testing is required.

As the saying goes, 'A picture is worth a thousand words'. That pictures have advantages over words is supported by research showing that meaning is extracted more quickly from pictures than from text (Smith & McGee, 1980). The current study demonstrates that some pictures do this better than others.

Acknowledgments

This research was supported in part by the National Institute of Information and Communications Technology, Japan, and the ESRC and EPSRC (grant L323253003). We are grateful to Yasuhiro Katagiri and three anonymous reviewers for their helpful comments on an earlier draft of this paper.

References

Arbib, M. A. (in press). From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics. To appear in *Behavioral Brain Sciences*.

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments & Computers*, 25, 257-271.

Fay, N., Garrod, S., Lee, J., & Oberlander., J. (2003). Understanding interactive graphical communication. *Proceedings of the 25th Annual Conference of the Cognitive Science Society, 2003,* pp. 384-389.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53, 181-215.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Isaacs, E. A., & H. H. Clark. (1987). References in conversations between experts and novices. *Journal of Experimental Psychology: General,* 116, 26-37.

Lewis, D. K. (1969). *Convention: a philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, D. K. (1975). Language and languages. In K Gunderson (Ed.), *Language Mind and Knowledge Minnesota studies in the Philosophy of Science* (Vol. 7). Minneapolis: University of Minnesota Press.

Pelli, D. G., Burns, C. W., Farrell, B., & Moore, D.C. (accepted with minor revisions). Identifying letters. *Vision Research*.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353-363.

Smith, M.C. & McGee L.E. (1980). Tracing the time course of picture-word processing. *Journal of Experimental Psychology: General,* 109, 373-392.

Vaccari, O., & Vaccari, E. E. (1961). Pictorial Chinese-Japanese Characters. Fourth Edition, Tokyo: Charles E. Tuttle Co.

Vygotsky. L. S. (1981). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Verbal Working Memory in Sentence Comprehension

Evelina Fedorenko (evelina9@mit.edu)
Department of Brain and Cognitive Sciences, 77 Mass Ave, NE20-437d
Cambridge, MA 02139 USA

Edward Gibson (egibson@mit.edu)
Department of Brain and Cognitive Sciences, 77 Mass Ave, NE20-459
Cambridge, MA 02139 USA

Douglas Rohde (dr@tedlab.mit.edu)
Department of Brain and Cognitive Sciences, 77 Mass Ave, NE20-437
Cambridge, MA 02139 USA

Abstract

This paper investigates the nature of verbal working memory (WM) in sentence comprehension and provides evidence for overlapping pools of verbal WM resources between on-line sentence comprehension and other verbally-mediated tasks. We report the results of two dual-task experiments. In Experiment 1, participants simultaneously performed a self-paced reading task and a self-paced arithmetic addition task in a 2x2 design crossing syntactic complexity (low, high) and arithmetic complexity (low, high). In addition to two main effects, the most interesting result was a significant interaction between syntactic and arithmetic complexity during the critical region of the linguistic materials: participants processed the complex/complex condition more slowly than would be expected if the two tasks relied on independent resource pools. To address a potential confound of shared attentional resources, Experiment 2 was conducted, where participants simultaneously performed a self-paced reading task and a self-paced spatial-rotation task in a similar 2x2 design crossing syntactic complexity with the complexity of the spatial task. As in Experiment 1, there were two main effects of complexity in the critical region. However, in contrast to Experiment 1, these effects were strictly additive, with no trace of interaction. The results of the two experiments therefore support a WM framework where on-line linguistic processing and on-line arithmetic processing rely on overlapping pools of verbal WM resources.

Introduction

A major question in psycholinguistic research concerns the nature of the working memory (WM) resources used in language processing. Empirical research has suggested that different pools of WM resources are used for processing visuo-spatial information and verbal information (e.g., Baddeley & Hitch, 1974; Baddeley, 1986; Vallar & Shallice, 1990; Hanley et al., 1991; Jonides et al., 1993; Shah & Miyake, 1996). Some researchers (Caplan & Waters, 1999; cf. Just & Carpenter, 1992) have hypothesized that the verbal WM pool can be further divided into two sub-pools: (1) verbal WM for natural language comprehension and production; and (2) verbal WM for non-linguistic verbally-mediated cognitive tasks. This paper attempts to empirically evaluate this hypothesis.

One way to address this question is via dual-task paradigms in which participants perform two tasks simultaneously: (1) on-line sentence processing, and (2) a non-linguistic verbally-mediated task. The underlying assumption is that we should observe a super-additive interaction when the complexity of both tasks is high only if the two tasks rely on overlapping pools of resources.

Previous dual-task experiments found either no interaction or only a suggestion of one (e.g. King & Just, 1991; Just & Carpenter, 1992; Caplan & Waters, 1999; Gordon et al., 2002). In all of the previous experiments, however, the secondary task involved storage of words or digits across the sentence-processing task. Although storage – in a very general sense of keeping track of previously encountered information – plays an important role in on-line sentence comprehension (e.g., Chomsky & Miller, 1963; Kimball, 1973; Gibson, 1991; 1998; Lewis, 1996), it may be qualitatively different from the kind of storage involved in the secondary tasks in the earlier experiments.

According to one recent resource-based theory of on-line syntactic processing, the dependency locality theory (DLT; Gibson, 1998, 2000), there are two working memory components to sentence comprehension: storage and integration. The *storage* component involves keeping track of partially processed syntactic dependencies that are still awaiting their second element in order for the sentence to be grammatical, whereas the *integration* component involves connecting a newly input word into the structure that has been built so far. Critically, the storage component of on-line sentence comprehension is unlike the storage involved in keeping track of a list of unconnected items. Consequently, it is possible that the lack of on-line interactions between syntactic complexity and memory load in earlier studies could be a result of the distinct nature of the storage processes involved. Moreover, there have been no previous attempts to explore the potential interaction between integration processes in sentence comprehension and secondary verbally-mediated tasks, which involve similar but non-linguistic on-line integration processes. In the current paper, we propose a novel paradigm to address this issue.

Experiment 1

This experiment had a dual-task design, in which participants read sentences phrase-by-phrase, and at the same time were required to perform simple additions. The on-line addition task is similar to on-line sentence comprehension in that an incoming element – a number – must be integrated into (i.e., added to) the representation constructed thus far: the working sum. Both tasks had two levels of complexity, resulting in a 2x2 design. Critically, there was no difference in linguistic complexity between the easy and hard arithmetic conditions: the complexity of the arithmetic task was manipulated in terms of the difficulty of the arithmetic operations (by making the addends larger), while keeping the linguistic form of the two conditions identical (number + number + number, etc.). Therefore, if we observe a super-additive interaction between the two tasks when the complexity of both tasks is high, then we may infer that the verbal WM resources that are involved in performing the arithmetic task overlap with those that are involved in syntactic integration processes. In contrast, if language processing relies on an independent verbal WM resource pool, there should be no such interaction.

Methods

Participants Forty participants from MIT and the surrounding community were paid for their participation. All were native speakers of English and were naive as to the purposes of the study.

Design and materials The experiment had a 2x2 design, crossing syntactic complexity (subject-extracted relative clauses (RCs), object-extracted RCs) with arithmetic complexity (simple additions (low initial addend, consequent addends between 1 and 3) vs. complex additions (higher initial addend, consequent addends between 4 and 6)).

The language materials consisted of 32 sets of sentences, having four different versions as in (1):

(1) a. *Subject-extracted, version 1*:
The janitor | who frustrated the plumber | lost the key | on the street.

b. *Subject-extracted, version 2*:
The plumber | who frustrated the janitor | lost the key | on the street.

c. *Object-extracted, version 1*:
The janitor | who the plumber frustrated | lost the key | on the street.

d. *Object-extracted, version 2*:
The plumber | who the janitor frustrated | lost the key | on the street.

As described above, there were only two levels of syntactic complexity – subject- and object-extractions – but there were four versions of each sentence in order to control for potential plausibility differences between the subject- and object-extracted versions of each sentence. As a result, no independent plausibility control is needed in this design. Each participant saw only one version of each sentence, following a Latin-Square design.

The numbers for the addition task were randomly generated online for each participant with the following constraints: (1) the value of the initial addend in the easy-math condition varied from 1 to 10, whereas the value of the initial addend in the hard-math condition varied from 11 to 20, and (2) the addends varied from 1 to 3 in the easy-math condition and from 4 to 6 in the hard-math condition.

In addition to the target sentences, 40 filler sentences with various syntactic structures other than relative clauses were included. The length and syntactic complexity of the filler sentences was similar to that of the target sentences. The stimuli were pseudo-randomized separately for each participant, with at least one filler separating the target sentences.

Procedure The task was self-paced phrase-by-phrase reading with a moving-window display (Just, Carpenter & Woolley, 1982). The experiment was run using the Linger 2.85 software by Doug Rohde. Each experimental sentence had four regions (as shown in (1a)-(1d)): (1) a noun phrase, (2) an RC (subject-/object-extracted), (3) a main verb with a direct object (an inanimate noun phrase) and (4) an adjunct prepositional phrase. The addends for the addition task were presented simultaneously with the sentence fragments, above and aligned with the second character of each fragment. The first sentence region had a number above it (e.g. "12") and all the consequent regions had a plus sign followed by a number (e.g. "+4"), as shown in Figure 1.

Time 1:	12 The janitor	--- ---------- --- -------	---- --- ---	-- --- ------.
Time 2:	+4 --- ------- who frustrated the plumber		---- --- ---	-- --- ------.
Time 3:	+5 --- -------	--- ---------- --- -------	lost the key	-- --- ------.
Time 4:	+4 --- -------	--- ---------- --- -------	---- --- ---	on the street.

Figure 1: Sample frame-by-frame presentation of an item.

Each trial began with a series of dashes marking the length and position of the words in the sentence. Participants pressed the spacebar to reveal each region of the sentence. As each new region appeared, the preceding region disappeared along with the number above it. The amount of time the participant spent reading each region and performing the accompanying arithmetic task, was recorded as the time between key-presses.

To make sure the participants performed the arithmetic task, a window appeared at the center of the screen at the end of each sentence and the participants were asked to type in the sum of their calculations. If the answer was correct, the word "CORRECT" flashed briefly on the screen, if the answer differed by up to 2 from the correct sum, the word "CLOSE" flashed briefly, and if the answer was off by more than 2, the word "INCORRECT" flashed briefly on the screen. To assure that the participants read the sentences for meaning, two true-or-false statements were presented

sequentially after the sum question, asking about the propositional content of the sentence they just read. Participants pressed one of two keys to respond "true" or "false" to the statements. After a correct answer, the word "CORRECT" flashed briefly on the screen, and after an incorrect answer, the word "INCORRECT" flashed briefly.

Participants were instructed not to concentrate on one task (reading or additions) more than the other. They were asked to read sentences silently at a natural pace and to be sure that they understood what they read. They were also told to answer the math and sentence questions as quickly and accurately as they could, and to take wrong answers as an indication to be more careful.

Before the experiment started, a short list of practice items and questions was presented in order to familiarize the participants with the task. Participants took approximately 35 minutes to complete the experiment.

Results

Arithmetic accuracy Participants answered the arithmetic sum correctly 88.7% of the time. A two-factor ANOVA crossing arithmetic complexity (easy, hard) and syntactic complexity (easy, hard) on these question-answering data revealed a main effect of arithmetic complexity ($F1(1,39)=9.45$; $MSe=0.120$; $p < .005$; $F2(1,31)=7.21$; $MSe=0.087$; $p < .02$), but no other significant effects.

Comprehension question performance There were two comprehension questions following each experimental trial. Participants answered the first question correctly 80.2% of the time, and the second question 78.1% of the time. The percentages of correct answers by condition were very similar for the two questions, so we collapsed the results in our analyses. A two-factor ANOVA crossing arithmetic complexity (easy, hard) and syntactic complexity (easy, hard) on the responses to the two comprehension questions revealed a main effect of syntactic complexity ($F1(1,39)=9.8$; $MS=0.1$; $p < .005$; $F2(1,31)=4.04$; $MS=0.074$; $p=.05$) and a main effect of arithmetic complexity in the participants analysis ($F1(1,39)=4.31$; $MS=0.047$; $p <.05$; $F2(1,31)=2.9$; $MS=0.042$; $p =.10$), but no significant interaction ($Fs < 1$).

Reaction times Because participants had to answer three questions (one math, two language) for each sentence, the odds of getting all three correct were not very high overall (55.6%). As a result, we analyzed all trials, regardless of how the comprehension questions were answered. The data patterns were very similar in analyses of smaller amounts of data, in which we analyzed (1) trials in which one or both of the language comprehension questions were answered correctly, or (2) trials in which the math question was answered correctly. To adjust for differences in word length as well as overall differences in participants' reading rates, a regression equation predicting reading times from word length was derived for each participant, using all filler and target items (Ferreira & Clifton, 1986; see Trueswell, Tanenhaus & Garnsey, 1994, for discussion). At each word position, the reaction time predicted by the participant's regression equation was subtracted from the actual measured reaction time to obtain a residual reaction time. The statistical analyses gave the same numerical patterns for analyses of raw reaction times. Reaction time data points that were less than 100 msec in the raw data (indicating erroneous key presses) or more than 2.5 standard deviations away from the mean residual RT for a position within a condition were excluded from the analysis, affecting 3.3% of the data. Figure 2 presents the mean residual RTs per region across the four conditions of the experiment.

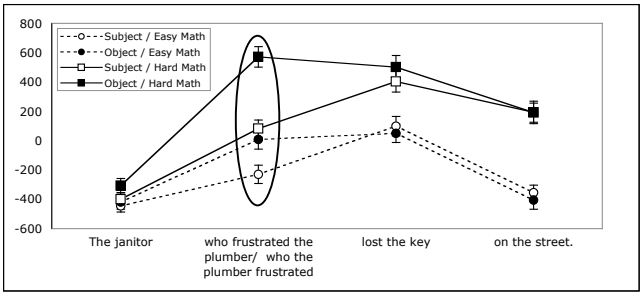


Figure 2: Reaction times per region in the four conditions of Experiment 1. The critical region is circled.

We present the analysis of the critical region (Region 2) first, followed by the analyses of the other regions. The critical region included the RC ("who frustrated the plumber" / "who the plumber frustrated"). A 2x2 ANOVA (easy-math / hard-math, subject-extracted RC / object-extracted RC) in this region revealed two significant main effects and a significant interaction. First, the hard-math conditions were read significantly slower than the easy-math conditions ($F1(1,39)=47.26$; $MSe=7641827$; $p < .001$; $F2(1,31)=42.58$; $MSe=5880083$; $p < .001$). Second, the syntactically more complex object-extracted RC conditions were read significantly slower than the subject-extracted conditions ($F1(1,39)=38.74$; $MSe=5283587$; $p < .001$; $F2(1,31)=33.4$; $MSe=4072481$; $p < .001$). Third, and most interestingly, there was a significant interaction, such that in the hard math conditions, the difference between subject- and object-extracted RCs was larger than in the easy math conditions ($F1(1,39)=4.74$; $MSe=623599$; $p < .05$; $F2(1,31)=7.15$; $MSe=526415$; $p < .02$). This interaction is predicted by the hypothesis whereby sentence processing and arithmetic processing rely on overlapping pools of resources, but not by the hypothesis that the pools of resources are independent.

In Region 1, consisting of the main clause subject (e.g., "The janitor") together with the initial addend, a 2x2 ANOVA revealed a main effect of arithmetic complexity (marginal in the items analysis), but no other significant effects. The hard-math conditions were read slower than the easy-math conditions ($F1(1,39)=5.08$; $MSe=245326$; $p < .05$; $F2(1,31)=3.62$; $MSe=149836$; $p = .067$). In Region 3, the top-level verb and its object ("lost the key"), a 2x2 ANOVA revealed a main effect of arithmetic complexity ($F1(1,39)=30.21$; $MSe=5726294$; $p < .001$; $F2(1,31)=33.32$; $MSe=3978352$; $p < .001$), but no other effects. Finally, in

Region 4, the sentence-final prepositional phrase ("on the street"), there was again an effect of arithmetic complexity ($F1(1,39)=72.58$; $MSe=13066602$; $p < .001$; $F2(1,31)= 105.06$; $MSe=10545386$; $p < .001$), but no other effects.

Discussion

The results of Experiment 1 are consistent with a WM framework where online sentence comprehension and arithmetic processing rely on overlapping resource pools. Most importantly, there was an interaction between syntactic complexity and arithmetic complexity in the critical region of the linguistic materials, where syntactic complexity was manipulated between subject-extracted RCs (low complexity) and object-extracted RCs (high complexity). There was no evidence of any interaction of this kind in any of the other three regions. Critically, linguistic complexity was not varied in the arithmetic task, so the observed interaction is not due to an overlap in the linguistic processes that are involved in the two tasks.

It should be noted, however, that there is an alternative explanation for the observed pattern of results in terms of attentional resources required for the simultaneous performance of the two tasks. In dual-task paradigms, resources are needed in order to direct attention to one task or another. It is possible that in the difficult conditions, more attention switches are required, or the switches between tasks are more costly. The observed interaction could therefore be a result of additional task-switching costs in the high syntactic complexity / high arithmetic complexity condition. Experiment 2 was designed to address this issue.

Experiment 2

This experiment used a similar dual-task paradigm as the first experiment. In contrast to Experiment 1, however, the secondary task was a spatial-rotation task matched for difficulty with the addition task used in Experiment 1. In this task, participants were instructed to visually imagine adding different-size sectors of a circle and to keep track of the angle subtended by the combined segments. The most natural way to solve this task is to mentally rotate each incoming sector until it abuts the estimated sum of the previous sectors. The on-line spatial-rotation task is similar to the addition task in that an incoming element – a sector – must be integrated into, or added to, the representation constructed thus far. Critically though, the spatial-rotation task does not rely on verbal WM resources, and should not therefore interact with the sentence-processing task if the cause for the observed interaction in Experiment 1 is an overlap in the use of verbal WM resources. However, if the attentional costs are responsible for the interaction, we should observe a similar interaction, regardless of the nature of the secondary task.

Methods

Participants Twenty-four participants from MIT and the surrounding community were paid for their participation. All were native speakers of English and were naive as to the purposes of the study. None of the participants took part in Experiment 1.

Design and materials The experiment had a 2x2 design, crossing syntactic complexity (subject-/ object-extracted RCs) with the complexity of the spatial-rotation task (simple rotations with small-angle sectors/ complex rotations with larger-angle sectors). The language materials were exactly the same as those used in Experiment 1.

The sectors for the spatial-rotation task were randomly generated online for each participant in the following way: the size of the sectors for the easy condition varied from 5 to 90 degrees, whereas the size of the sectors for the hard condition varied from 30 to 180 degrees. As a result, it was more likely in the hard condition for the sum of sectors to be more than 360 degrees, thus "wrapping around" the circle. Pilot testing of the pie task by itself suggested that the task is easier to perform with smaller sectors.

As in Experiment 1, 40 filler sentences with various syntactic structures other than relative clauses were included, and the stimuli were pseudo-randomized separately for each participant, with at least one filler separating the target sentences.

Procedure The procedure was identical to that of Experiment 1, except for substituting the spatial-rotation task for the arithmetic task. Above each sentence fragment, participants saw a small circle. They were instructed to think of it as a plate for a pie. On each "plate", there was a "pie-slice" shown in blue. The size of the "pie-slices" varied (as described in Materials and Design above), but they all started at the 12:00 position, as shown in Figure 3.

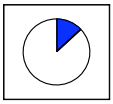


Figure 3: Sample figure of the spatial-rotation task.

Participants were instructed to visually imagine adding each new "pie-slice" to the previous one(s) by mentally "putting" them next to each other. To assure that the participants performed the task, at the end of each trial a large blank circle appeared at the center of the screen with a vertically-pointing radius. Participants were instructed to drag this radius (by using the mouse) to the end-point where all the "pie-slices" they just saw would come to when placed next to each other. If the answer was within 10 degrees of the correct answer, the words "Very Close!" flashed briefly on the screen; if the answer was within 35 degrees, the words "Pretty Good" flashed briefly; if the answer was within 90 degrees, the words "In The Ballpark" flashed briefly; finally, if the answer was not within 90 degrees, the words "Not Very Good" flashed briefly on the screen. The participants were warned that sometimes the "pie-slices", when added together, would form more than a complete pie. In such cases, they were told to assume that the slices "wrapped around" and to ignore the complete portion of the pie.

As in Experiment 1, this task was followed by two comprehension questions about the content of the sentences.

Results

Spatial-rotation task accuracy On average, participants' estimates were 30.3 degrees off of the correct answer. A two-factor ANOVA crossing spatial-rotation task complexity (easy, hard) and syntactic complexity (easy, hard) revealed a main effect of complexity of the spatial-rotation task ($F1(1,23)=18.36$; $MSe=2676$; $p < .0005$; $F2(1,31)=22.28$; $MSe=3568$; $p < .0005$), but no other significant effects. It is worth noting that this pattern of results for the spatial-rotation task accuracy is parallel to that of the results for the arithmetic task accuracy in Experiment 1.

Comprehension question performance There were two comprehension questions following each experimental trial. The percentages of correct answers by condition were very similar for the two questions, so we collapsed the results in our analyses. Across conditions, participants answered the questions correctly 83% of the time. A 2x2 ANOVA crossing spatial-rotation task complexity (easy, hard) and syntactic complexity (easy, hard) on the responses to the comprehension questions revealed no significant effects or interactions ($Fs<1$). This pattern of results differs slightly from that in Experiment 1 in that there was no effect of syntactic complexity in Experiment 2. Note, however, that overall, subjects performed better on comprehension questions in Experiment 2 (83% across conditions), compared with Experiment 1 (79% across conditions). This accuracy difference across the experiments may have resulted from greater interference of the secondary task in Experiment 1 with subjects' memory of the propositional content of the sentences, due to its verbal nature. The lack of syntactic complexity effect in Experiment 2 could then be explained by a possible ceiling effect in the comprehension performance: without a verbally interfering task, people perform well on both the subject- and object-extracted relative clause sentence types.

Reaction times As in Experiment 1, we analyzed all trials, regardless of how the comprehension questions were answered. Also, as in Experiment 1, reaction time data points that were more than 2.5 standard deviations away from the mean residual RT for a position within a condition or less than 100 msec in the raw data were excluded from the analyses, affecting 3.7% of the data. Figure 4 presents the mean reaction times per region across the four conditions in the experiment.

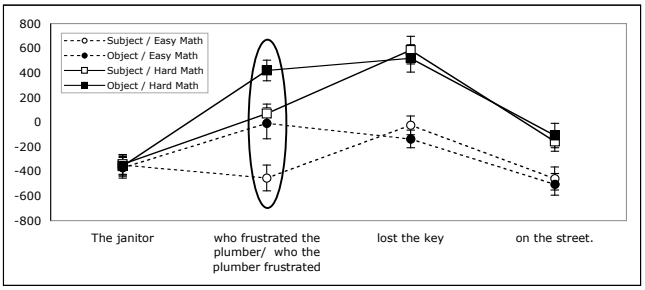


Figure 4: Reaction times per region in the four conditions of Experiment 2. The critical region is circled.

We first present the analysis of the critical region, Region 2, which included the RC ("who frustrated the plumber" / "who the plumber frustrated"). A 2x2 ANOVA conducted on this region revealed two significant main effects. First, the hard-spatial-task conditions were read significantly slower than the easy-spatial-task conditions ($F1(1,23)=22.98$; $MSe=5451605$; $p < .001$; $F2(1,31)=40.08$; $MSe=6428277$; $p < .001$). Second, the syntactically more complex object-extracted RC conditions were read significantly slower than the subject-extracted RC conditions ($F1(1,23)=15.59$; $MSe=3791349$; $p < .001$; $F2(1,31)=22.94$; $MSe=4675397$; $p < .001$). Critically, there was no trace of an interaction between syntactic complexity and the complexity of the spatial task ($Fs<1$). Moreover, the effect of syntactic complexity in the hard-spatial-task conditions was numerically *smaller* than that in the easy-spatial-task conditions. This result rules out the attentional explanation of the interaction that was observed in Experiment 1.

In Region 1, consisting of the main clause subject (e.g., "The janitor") together with the initial "pie-slice", a 2x2 ANOVA revealed no significant effects. In Region 3, the top-level verb and its object ("lost the key"), a 2x2 ANOVA revealed a main effect of spatial task complexity ($F1(1,23)=39.36$; $MSe=9601145$; $p < .001$; $F2(1,31)=62.5$; $MSe=12710598$; $p < .001$), but no other effects. Finally, in Region 4, the sentence-final prepositional phrase ("on the street"), there was again an effect of spatial task complexity ($F1(1,23)=16.1$; $MSe=2925378$; $p < .001$; $F2(1,31)=45.2$; $MSe=4061993$; $p < .001$), but no other effects.

Discussion

The attentional account of the interaction between syntactic and arithmetic complexity that was observed in Experiment 1 predicted a similar interaction between syntactic and spatial-rotation complexity in Experiment 2. No such interaction was observed. In fact, the numerical trend was in the reverse direction. The lack of such an interaction therefore argues against the attentional account of the interaction observed in Experiment 1.

In general, the lack of an interaction between the complexity of two tasks could arise for at least two different reasons: (1) independent resource pools required for each task; or (2) ceiling or floor effects on one or both of the tasks, such that resources are either abundant or insufficient. Hence, in order to argue that the results of Experiment 2 are

due to independent resource pools for the two tasks, we need to be confident that the secondary task is neither too complex nor too simple. It is unlikely that the spatial-rotation task is too simple, because we observed a highly significant complexity effect for this task. Neither is it likely that the spatial-rotation task is too complex for the following reasons. First, the performance on the spatial-rotation task was extremely good, averaging only 30.3 degrees off from the target position. Second, the range of the reaction times across conditions for the two experiments was almost identical, suggesting that the arithmetic and spatial-rotation tasks were comparable in difficulty.

Conclusions

In summary, using a dual-task paradigm, we have demonstrated an on-line interaction between syntactic complexity and arithmetic complexity in Experiment 1 suggesting that these two cognitive functions rely on overlapping pools of verbal WM resources. Furthermore, in Experiment 2, we have ruled out an attentional account of the observed interaction by showing that a spatial task, which does not rely on verbal WM resources, does not interact with on-line sentence comprehension. These results therefore support a WM framework in which sentence processing and arithmetic processing overlap in the use of verbal WM resources. The results are not consistent with the hypothesis whereby sentence processing relies on an independent pool of verbal WM resources (Caplan & Waters, 1999).

An open question that we have not yet addressed is the exact nature of the overlap in verbal working memory resources for sentence and arithmetic processing. One possibility is that both syntactic and arithmetic processes involve a subservant mechanism for integrating verbal symbolic information units. In this mechanism, the difficulty of integrating linguistic elements depends on the distance between elements to be connected. Relatedly, the difficulty of adding numbers depends on the distance between the initial addend and the resulting sum in the computation on the number line. We leave it to future work to distinguish this hypothesis from other possibilities.

Acknowledgments

DLT Rohde was supported by NIH NRSA 1-F32-MH65105-02.

References

Baddeley, A. D. & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). New York: Academic Press.

Baddeley, A. D. (1986). *Working Memory*. New York: Oxford University Press.

Caplan, D. & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Brain & Behavioral Sciences, 22,* 77-126.

Chomsky, N. & Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. 2. Wiley, New York, pp. 269–321.

Ferreira, F. & Clifton, C., Jr. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348-368.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68,* 1-76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., & O'Neil, W. (Eds.), *Image, language, brain.* MIT Press, Cambridge, MA.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27(6),* 1411-1423.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science, 13,* 425-430.

Hanley, J. R., Young, A., & Pearson, N. A. (1991). Impairment of the visuo-spatial scratchpad. *Quarterly Journal of Experimental Psychology, 43A,* 101-125.

Jonides, J., Smith, E. E., Koeppe, R. A., Awh, E., Minoshima, S., & Mintun, M. A. (1993). Spatial working memory in humans as revealed by PET. *Nature, 363,* 623-625.

Just, M.A. & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122-149.

Just, M.A., Carpenter, P.A., & Woolley, J.D. (1982). Paradigms and processing in reading comprehension. *Journal of Experimental Psychology: General, 111,* 228-238.

Kimball, J. (1973). Seven Principles of Surface Structure Parsing in Natural Language. *Cognition, 2*, 15-47.

King, J. & Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language, 30,* 580-602.

Lewis, R., (1996). A theory of grammatical but unacceptable embeddings. *Journal of Psycholinguistic Research, 25*, 93–116.

Shah, P. & Miyake, A. (1996). The Separabillity of Working Memory Resources for Spatial Thinking and Language Processing: An Individual Differences Approach. *Journal of Experimental Psychology: General, 125 (1),* 4-27.

Trueswell, J.C., Tanenhaus, M.K. & Garnsey, S.M. (1994). Semantic influences on parsing: use of thematic role information in syntactic disambiguation. *Journal of Memory and Language, 33*, 285-318.

Vallar, G. & Shallice, T. (Eds.). (1990). *Neuropsychological Impairments of short-term memory*. New York: Cambridge University Press.

Talking about space: A cross-linguistic perspective

Michele I. Feist (m-feist@ northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Abstract

What do people attend to when describing the locations of objects in space? This paper describes a study of the ways in which speakers of seventeen languages describe static spatial relations, delving into the meanings of two kinds of spatial relational terms evident cross-linguistically: specific spatial terms and generalized spatial terms. The findings provide support for the importance of geometry, function, and qualitative physics to the meanings of specific spatial terms and suggest an interplay between semantic and pragmatic elements of meaning for generalized spatial terms.

Introduction

Multiple times each day, speakers make use of a relatively small set of spatial relational terms (Landau & Jackendoff, 1993) in order to localize themselves and the entities with which they interact. Use of these terms is practically automatic; from the point of view of the native speaker, they are simple, clear, and obvious. However, the difficulty that spatial terms present to second language learners belies this apparent simplicity. Furthermore, the prodigious cross-linguistic variability in spatial terms (cf, Levinson, Meira, & The Language and Cognition Group, 2003) suggests that they are anything but simple, clear, and obvious.

The variability evident in spatial language takes on many different forms. As Bowerman and her colleagues have shown, distinctions that are drawn in one language may not be drawn in another. For example, while English distinguishes between support and contact, on the one hand, and containment, on the other, this distinction does not appear in Korean spatial terms (Bowerman & Choi, 2001). Instead, Korean distinguishes between a tight fit and a loose fit between two objects, a distinction not evident in English spatial terms. Thus, in Korean, the act of putting a Lego *onto* a stack of Legos would be described by the same term as the act of putting a book *into* its sleeve: both are instances of tight fit. Further, the act of putting a Lego *onto* a stack of Legos is distinguished from the act of putting a book *onto* a desk: while the former is a tight fit relation, the latter represents loose fit.

Even closely related languages are not immune from such differences in the distinctions drawn between spatial relational terms. For example, as Bowerman has pointed out (Bowerman, 1996; Bowerman & Pederson, 1992, 1996; Gentner & Bowerman, 2000), Dutch makes a three-way distinction where English does not: between a cup *on* a table (Dutch *op*), a picture *on* a wall (Dutch *aan*), and a ring *on* a finger (Dutch *om*).

Even if two languages appear to draw the same distinction, the boundaries between the contrasting categories often differ. For example, both English and Finnish mark a distinction between a very intimate relation such as containment and a less intimate relation such as surface contact, but the set of configurations placed in each group differs dramatically between the two languages (Table 1): rather than categorizing a handle *on* a pan as an instance of the less intimate relation, along with a cup *on* a table and a picture *on* a wall (as English does), Finnish places this configuration in the more intimate category along with an apple *in* a bowl (Bowerman, 1996).

A similar example comes from a comparison of English and Berber spatial terms. Spatial relational terms in Berber fail to make a distinction between inclusion and contact with/support via an external surface of the Ground (Bowerman & Choi, 2001) akin to the English *in-on* distinction. Rather, reminiscent of the case in Finnish, the distinction is between "being loosely in contact" and "being 'incorporated' into" the Ground, with "incorporation" including both being inside and being tightly attached to an external surface or point (Bowerman & Choi, 2001).

Table 1: English and Finnish categorizations of some Figure-Ground relations (adapted from Bowerman, 1996, Figure 4).

	Apple in bowl	Handle on pan	Bandaid on leg	Ring on finger	Fly on door	Picture on wall	Cup on table
English	In	On	On	On	On	On	On
Finnish	Inessive case	Inessive case	Inessive case	Inessive case	Inessive case	Adessive case	Adessive case

Despite this variability, all humans have the same ability to perceive spatial relations. As suggested by the cross-linguistic variability, spatial relational terms encode a variety of factors of the scenes they are used to describe (Bowerman, 1996; Levinson, 1996; Sinha & Thorseng, 1995), including geometric, functional, and qualitative physical factors. Furthermore, there is evidence from studies of English spatial prepositions that the influences of the factors interact, leading to complex meanings (Feist, 2002; Feist & Gentner, 2003). What are the roles of these factors in the spatial terms of a diverse set of languages? Are there some factors that recur in spatial meaning across languages? If so, these factors could underlie our conception of space in general.

How Do Speakers of Different Languages Talk About Space?

In a wide-reaching cross-linguistic survey, Bowerman and Pederson (1992, 1996) presented a set of carefully drawn pictures to speakers of thirty-four languages. Each picture depicts a spatial relation, with the Figure colored in yellow and the Ground in black and white. Informants provided descriptions of the pictures, including the spatial relational term that would most naturally be used to describe the relation depicted.

Bowerman and Pederson examined the ways in which the languages in their survey grouped the spatial relations in their pictures, as defined by description by the same term. This led to the discovery of a "similarity gradient" (Bowerman & Choi, 2001) along which they could arrange the scenes from their study. At one end of the gradient lie configurations in which a Figure is supported from below by a Ground (e.g., a cup on a table); at the other end lie configurations in which a Figure is completely included within a Ground (e.g., a pear in an otherwise empty bowl). In between lie configurations bearing similarities to both endpoints, arranged according to whether they are more similar to support from below or to complete inclusion.

Although Bowerman and Pederson found variation in how linguistic terms grouped spatial configurations, this variation was systematic. In particular, all of the languages in their sample respected the similarity gradient that Bowerman and Pederson had identified, only describing non-adjacent configurations with the same term if all configurations that lie between them are also described by the term. In other words, the ranges of use of spatial terms were found to be continuous with respect to the similarity gradient.

I borrowed Bowerman and Pederson's pictorial elicitation technique for collecting spatial terms, asking speakers of seventeen languages to describe a single set of simple pictures taken from the larger set used by Bowerman and Pederson. This resulted in the elicitation of a narrow set of spatial relational terms across a diverse set of languages. The ranges of use of the elicited terms illuminated the importance of a few attributes of spatial scenes instantiating relations along Bowerman and Pederson's similarity gradient, providing clues to the likely organizing dimensions of spatial terminology.

Method

Informants Twenty-eight speakers of seventeen languages (representing twelve language families, with 1-4 participants per language) were recruited from around the Northwestern University/Evanston community; one additional informant was recruited from the New York area. Informants ranged in age from 18 to 69 and were all native speakers of the languages in which they participated.

Stimuli The stimulus set consisted of twenty-nine line drawings, each depicting two objects in a simple spatial relation. The relations depicted span the similarity gradient described by Bowerman and Pederson (1992, 1996; Bowerman & Choi, 2001). Following their methodology, one of the objects in each picture, the Figure, was colored yellow; and the other, the Ground, was black and white. Twenty-seven of the twenty-nine drawings were borrowed from Melissa Bowerman and Eric Pederson's Topological Picture Series (cf., Bowerman & Pederson, 1992, 1996; Gentner & Bowerman, 1996, 2000; Levinson et al., 2003); one of the remaining two, a picture of an address on an envelope, was modified from a picture in the Topological Picture Series, and the other, a picture of flowers in a vase, was borrowed from an example in Coventry (1998).

Procedure Each informant participated individually in a session lasting an average of one hour. In the first part of the session, informants were shown each picture in the set individually. They were asked to provide a description in their native language of the location of the yellow object with respect to the other object. Responses were both tape-recorded and phonetically transcribed. After all of the pictures had been described, informants provided as close to a morpheme-by-morpheme translation as could be elicited.[1] Finally, informants for languages using the same orthography as English were asked to provide a written transcription of their responses.

[1] Variation in the exactness of the morpheme-by-morpheme translations resulted from informants' inability and/or unwillingness to provide translations below the level of the word.

Picture coding In previous work, there are numerous arguments for the importance of geometry (e.g., Bennett, 1975; Feist & Gentner, 1997, 1998, 2003; Herskovits, 1986; Landau, 1996; Miller & Johnson-Laird, 1976), function (Coventry, 1998; Coventry, Carmichael, & Garrod, 1994; Feist & Gentner, 1998, 2003; Vandeloise, 1991, 1994), and qualitative physics (Bowerman & Choi, 2001; Feist & Gentner, 2003; Forbus, 1983, 1984; Talmy, 1988) to spatial relational meaning. In order to determine whether there are attributes of spatial scenes that figure in the meanings of spatial relational terms across a range of languages[2], I coded each of the pictures for whether it matched each of a small set of attributes related to geometry, function, and qualitative physics.

For geometry, I coded for *a difference in vertical position* (important to terms such as *above, below, over,* and *under* (O'Keefe, 1996)), *contact* (important to terms such as *on* (Cienki, 1989; Herskovits, 1986; Miller & Johnson-Laird, 1976)), and *inclusion* (important to terms such as *in* (Cienki, 1989; Herskovits, 1986; Miller & Johnson-Laird, 1976)), as well as for *relative size*. Although not argued for in previous work, *relative size* was coded because a larger Ground might facilitate the matching of other attribute values, such as *support* or *inclusion* of the Figure, thus influencing spatial term use.

For function, I coded for the *functional relatedness* of the Figure and the Ground – the likely interaction resulting from an object's function (cf, Coventry, 1998; Coventry et al., 1994; Vandeloise, 1991, 1994 on the importance of function in general). For example, a lamp and a table are functionally related; a cloud and a mountain are not.

For qualitative physics, I coded for *support* by the Ground (important to terms such as *on* (Bowerman & Pederson, 1992, 1996; Herskovits, 1986; Miller & Johnson-Laird, 1976)). In addition, I coded for *animacy*[3] and the ability of the Ground to constrain the location of the Figure, which both influence what predictions can reasonably be made about the qualitative physics of a scene. Specifically, if a Ground can constrain the location of a Figure, the configuration may seem less subject to outside forces and thus more likely to remain as pictured. In addition, *animacy* of the Figure and the Ground was found in past research to influence speakers' choice between English *in* and *on* (Feist & Gentner, 1998, 2003). In addition, *constraint of location* may be important to some functional relations (such as functional containment; see Coventry et al., 1994), prompting its inclusion as an independent factor.

Results

Two kinds of spatial relational terms appeared in the elicited descriptions. The first, *specific spatial terms*, occur only in limited contexts and impart relatively specific information about the location of the Figure[4]. This kind of term can be exemplified by the English prepositions *in* and *on* and by the terms in examples (1) (from Croatian) and (2) (from Swedish).

(1) Jabuka je *v* zdjeli.
apple is *in* bowl
The apple is in the bowl.

(2) Koppen står *på* bordet.
cup-definite stands *on* table-definite
The cup is on the table.

The second kind of term, *generalized spatial terms*, occur in all spatial descriptions and impart no specific information about the location of the Figure. Rather, these terms just serve to relate the Figure to the Ground. Such terms do not occur in English, but can be exemplified by terms such as Japanese *ni* (example (3)) and Indonesian *di* (example (4)), both glossed simply as LOC (locative).

(3) Kaban no naka[5] *ni* haite iru hako.
bag genitive inside *LOC* put-in is box
The box is in the bag.

(4) Cincin itu *di* jari.
ring that *LOC* finger
The ring is on the finger.

Although generalized spatial terms are often glossed as *at, in*, or *on*, such glosses are hardly appropriate characterizations of the meanings of the terms, as will become clear below. In particular, generalized spatial terms appear in environments where the glosses would be unacceptable, raising questions about whether the glosses can capture the true meaning of the generalized spatial term.

Specific spatial terms For each specific spatial term collected, I grouped together the pictures that the term had been used to describe. Then, for each group, I isolated the attributes that were common to all of the pictures in the group. Only four of the attributes coded

[2] Here I consider only those terms whose distribution suggests that they are specific spatial terms (see below).

[3] I used a fairly broad definition of animacy, namely, things that are capable of self-determination (e.g., human legs, cats) were taken as animate, while objects incapable of self-determination (e.g., jackets, doors) were not. Looking across languages, this is not the only way to look at the notion of animacy.

[4] I include here spatial nominals and locative cases along with adpositions, as both occur (in languages using them) as answers to *where*-questions, and neither is expected to display semantic patterns different from those of adpositions (Levinson et al., 2003).

[5] Specific spatial terms like the Japanese spatial nominal *naka* often appear in spatial descriptions with generalized spatial terms, as will be discussed further below.

appeared as unifying factors for the ranges of use of the terms I collected: *difference in vertical position of the Figure and the Ground, contact, support of the Figure by the Ground, and inclusion of the Figure within the Ground*. This is exemplified by the terms in Table 2[6]; each term listed is marked with a plus under those attributes that must be true of scenes described by the term and a minus under those that may not be true. Attributes with neither a plus nor a minus may, but need not, be true of scenes described by the term. For example, the Polish term *na* requires that the Figure and the Ground be in contact and that the Ground support the Figure, regardless of which, if either, is higher, and regardless of whether the Figure might be included in the Ground[7].

These four attributes highlight the importance of geometry, function, and qualitative physics (cf., Feist & Gentner, 2003). The first two, *difference in vertical position of the Figure and the Ground* and *contact*, both encode information about the geometry of the relation between the Figure and the Ground. The next attribute, *support*, encodes information about the physics of the interaction (the Ground is constraining the location of the Figure in one dimension) and about the function of the Ground. Lastly, *inclusion* provides information about geometry, function, and physics, as the typical situation when a Figure is geometrically included in a Ground is that the Ground functions as a container for the Figure and thereby constrains the location of the Figure in more than one dimension.

In addition to highlighting the importance of geometry, function, and qualitative physics across a sizable sample of languages and spatial terms, this data demonstrates that important similarities co-exist with cross-linguistic variation (cf., Bowerman & Pederson, 1992, 1996; Levinson et al., 2003; Regier, 1996).

Generalized spatial terms The criterial factor for identifying generalized spatial terms is that they occur in all spatial descriptions. These terms can either appear alone or in combination with a more specific term, as exemplified by the Indonesian examples in (5) – (7).

(5) Buku itu *di* meja.
book that *LOC* table
The book is on the table.
⇒ *di atas* may be substituted for *di*

Table 2: Example terms and the attributes characterizing them.

Example terms	Figure higher than Ground	Contact	Ground supports Figure	Inclusion
ue (Japanese)	+			
taas (Tagalog)	+			
[nad][8] (Russian)	+	-		
[upar] (Hindi)	+	-		
na (Polish)		+	+	
på (Swedish)		+	+	
sur (French)		+	+	
auf (German)		+		
an (German)		+		
u (Croatian)				+
-bVn (Hungarian)				+
iqinde (Turkish)				+

(6) Parmen itu *di* kotak.
candy that *LOC* box
The candy is in the box.
⇒ *di dalam* may be substituted for *di*

(7) Meja itu *di* *bawah* lampu.
table that *LOC* beneath lamp
The table is under the lamp.
⇒ *di* may not occur alone

As mentioned earlier, generalized spatial terms such as Indonesian *di* are often glossed as *at, in*, or *on* (e.g., Macdonald, 1976). However, examination of the range of uses evident for *di* reveals that no single English preposition can occur in the entire range. Although *di* may occur alone in situations where English uses *at, in,* or *on*, it also appears in combination with locational nouns in situations where the English glosses are unacceptable (e.g., example (7)). Thus, glosses – which are a function of both the scene and the sentence as a whole (Ameka, 1995) – fail to capture their meaning.

What then is the meaning of the generalized spatial term? To account for both classes of use, I propose one basic element of meaning appropriate to all uses (8) and two pragmatically licensed elements of meaning (9).

(8) di = location of the Figure in the region of interaction of the Ground

[6] Due to space constraints, I only present a representative subset of the terms collected.

[7] Although the norm when the Figure is included in the Ground is to use a term marking inclusion, such as Polish *w*, terms such as *na* may be used to describe configurations such as a face on a stamp, in which the Figure may be conceived of as included in the Ground.

[8] Terms from languages that do not use the English alphabet are presented as phonetic transcriptions in square brackets.

(9) (a) the Figure is in contact with the Ground.
(b) the Figure-Ground relation is canonical.

The element of meaning in (8) serves to unite the disparate range of uses of *di* without falsely implying equivalence between *di* and its English prepositional glosses. Additionally, the pragmatically licensed elements of meaning proposed in (9) make a clear prediction about when *di* felicitously appears alone and when the addition of a locational noun is preferred. The default assumption when *di* appears alone is that the elements in (9) are true, although this is not always the case (see (11) and (12)). The use of a locational noun emphasizes the specifics of the relation and highlights any deviations from this assumption.

To test this analysis, I created spatial descriptions that violate each of the proposed elements of meaning (example (10) violated (8); (11) violated (9a); and (12) violated (9b)). Each sentence involved a use of *di* without the addition of a more specific term.

(10) *Buku itu *di* meja, tapi bukan dekatnya.
Book that *LOC* table but not near-possessive
The book is on the table but not near it.

(11) Buku itu *di* meja tapi tidak menyentuh.
Book that *LOC* table but not touching
The book is on the table but it's not touching it.

(12) Buku itu *di* meja tapi menempel dengan aneh.
Book that *LOC* table but stuck manner weird
The book is on the table but it's attached in a weird manner.

Eleven speakers of Indonesian were asked to assess the acceptability of the created sentences. The different violations resulted in quite different acceptability judgments: violations of (8) were rarely accepted (9%); violations of (9), while odd, were more acceptable (55% for (9a); 73% for (9b)), $F(2,32) = 6.09$, $p < .01$. These data support the hypothesis that (8) is part of the semantics of *di*, while the elements in (9) are pragmatically licensed.

Discussion

Analysis of the terms used to describe spatial locations across seventeen languages revealed spatial terms that fall into two classes: specific spatial terms, which provide semi-precise information about the location of a Figure; and generalized spatial terms, which simply serve to locate the Figure in the region of interaction of a named Ground.

By investigating spatial semantics in many languages, we can gain insights into the range of attributes of spatial scenes to which humans attend when localizing objects. By grouping pictures described by each individual specific spatial term, it was possible to isolate those attributes of the spatial scenes described that are important to the use of each term. Across seventeen languages, four attributes recurred in the meanings of specific spatial terms: a difference in vertical position, contact, support, and inclusion. This finding corroborates recent work by Levinson and his colleagues (Levinson et al., 2003) suggesting the importance of *attachment, a difference in vertical position (both super- and subadjacency), proximity,* and *containment* to spatial meanings across languages. These attributes together highlight the importance of geometric, functional, and qualitative physical factors, all of which have been argued in previous work to be important to spatial relational meaning in a small set of languages, to the meanings of spatial relational terms more generally.

Whereas specific spatial terms have received a fair amount of attention in linguistics and cognitive science, the attention accorded generalized spatial terms has been far sparser. However, an understanding of the uses of generalized spatial terms, both with and without accompanying specific spatial terms, is integral to an understanding of the range of spatial meanings evident in human language. The analysis of Indonesian *di* presented here provides a step towards a comprehensive account. Further testing of this analysis, including a study of the applicability of sentences without specific terms as descriptions of a variety of pictures, will be necessary to solidify the conclusions reached here. In addition, in order to arrive at a descriptively adequate account of generalized spatial terms, these studies will need to be repeated for generalized spatial terms of further languages.

Although more work, including the examination of a wider range of languages, is necessary to completely understand the factors influencing how humans talk about locations in space, I have presented here two windows into spatial relational meaning. The view from these windows, illuminating both cross-linguistic variation and commonalities in specific spatial terms through one, and a coherent meaning for a previously misunderstood set of terms through the other, furthers our understanding of cross-linguistic variation and linguistic universals in the semantics of space.

Acknowledgments

This work was supported by NSF-LIS award SBR-9720313 and NSF-ROLE award 21002/REC-0087516. I thank Florencia Anggoro for help in translating materials and conducting the Indonesian study. I am grateful to Melissa Bowerman and Eric Pederson for the use of their Topological Picture Series. I also thank Dedre Gentner, Beth Levin, Judith Levi, Terry Regier, Florencia Anggoro, and Jason Jameson for comments on the ideas presented here.

References

Ameka, F.K. (1995) The Linguistic Construction of Space in Ewe. *Cognitive Linguistics 6,* 139-181.

Bennett, D. C. (1975). *Spatial and temporal uses of English prepositions: An essay in stratificational semantics*. London: Longman.

Bowerman, M. (1996). The origins of children's spatial semantic categories: Cognitive vs. linguistic determinants. In Gumperz, J. and Levinson, S. (Eds.), *Rethinking Linguistic Relativity*. Cambridge, England: Cambridge University Press.

Bowerman, M. & Choi, S. (2001). Shaping meanings for language: Universal and language specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge, UK: Cambridge University Press.

Bowerman, M., & Pederson, E. (1992). *Cross-linguistic perspectives on topological spatial relationships*. Paper presented at the 91st Annual Meeting of the American Anthropological Association, San Francisco, CA.

Bowerman, M., & Pederson, E. (1996). *Cross-linguistic perspectives on topological spatial relationships*. Manuscript in preparation.

Cienki, A. J. (1989). *Spatial cognition and the semantics of prepositions in English, Polish, and Russian*. Munich, Germany: Verlag Otto Sagner.

Coventry, K.R. (1998). Spatial prepositions, functional relations, and lexical specification. In Olivier, P. and Gapp, K-P. (Eds.), *The representation and processing of spatial expressions.* Mahwah, NJ.: Lawrence Erlbaum.

Coventry, K., Carmichael, R., & Garrod, S. C. (1994). Spatial prepositions, object-specific function, and task requirements. *Journal of Semantics, 11,* 289-309.

Feist, M. I. (2002). Geometry, function, and the use of *in* and *on*. Paper presented at the *Sixth Conference on Conceptual Structure, Discourse, and Language*, Houston, TX.

Feist, M. I., & Gentner, D. (1997). *Animacy, control, and the IN/ON distinction*. Paper presented at the Fourteenth National Conference on Artificial Intelligence, Workshop on Language and Space, Providence, RI.

Feist, M. I., & Gentner, D. (1998). On plates, bowls, and dishes: Factors in the use of English IN and ON. *Proceedings of the Twentieth Annual meeting of the Cognitive Science Society*, 345-349.

Feist, M. I., & Gentner, D. (2003). Factors involved in the use of *in* and *on*. *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*.

Forbus, K. D. (1983). Qualitative reasoning about space and motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Lawrence Erlbaum.

Forbus, K. D. (1984). Qualitative process theory. *Journal of Artificial Intelligence, 24,* 85-168.

Gentner, D., & Bowerman, M. (1996). *Crosslinguistic differences in the lexicalization of spatial relations and effects on acquisition.* Paper presented at the Seventh International Congress for the Study of Child Language, Istanbul, Turkey.

Gentner, D., & Bowerman, M. (2000). Not all *on*'s are equal: Crosslinguistic differences in spatial relations and their effects on acquisition. Unpublished mss.

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in English.* Cambridge, England: Cambridge University Press.

Landau, B. (1996). "Multiple geometric representations of objects in languages and language learners." In Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. F. (Eds.), *Language and Space*. Cambridge: MIT Press.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences, 16,* 217-265.

Levinson, S.C. (1996). Relativity in spatial conception and description. In Gumperz, J. and Levinson, S. (Eds.), *Rethinking Linguistic Relativity*. Cambridge, England: Cambridge University Press.

Levinson, S., Meira, S. & The Language and Cognition Group. (2003). "Natural concepts" in the spatial topological domain – adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language, 79 (3),* 485-516.

Macdonald, R. R. (1976). *Indonesian Reference Grammar.* Washington: Georgetown University Press.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception.* Cambridge, MA: Belknap Press of Harvard University Press.

O'Keefe, J. (1996). The spatial prepositions in English, vector grammar, and the cognitive map theory. In Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. F. (Eds.), *Language and Space*. Cambridge: MIT Press.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.

Sinha, C. & Thorseng, L. A. (1995). A coding system for spatial relational reference. *Cognitive Linguistics, 6-2/3*, 261-309.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12,* 49-100.

Vandeloise, C. (1991). *Spatial prepositions: A case study from French.* Chicago: University of Chicago Press.

Vandeloise, C. (1994). Methodology and analyses of the preposition *in*. *Cognitive Linguistics, 5 (2),* 157-184.

Explorations of the (Meta)Representational Status of Desire in the Theory-Theory of Mind Framework

Leo Ferres (lferres@ccs.carleton.ca)
Human-Oriented Technology Laboratory
Carleton University, 1125 Colonel By Drive
Ottawa, ON K1S 5B6 Canada

Abstract

Some researchers have proposed that what accounts for children's earlier ability to reason by means of desire compared to reasoning by means of belief is the fact that desires do not necessarily invoke the ability to metarepresent. In this paper, I argue that this is a misconception stemming from the confusion between desire ascription and simple desire states. In other words, there would be no way to entertain a thought about someone's desire without metarepresenting, in Leslie's (1991) terms. I provide some empirical evidence in the fashion of Bartsch and Wellman (1995) that also points in this direction.

The problem

Although the concept of desire is at least as important as the concept of belief for describing, explaining and predicting the behavior of different entities (Fodor, 1987; Dennett, 1978b, 1987), substantially more attention has been paid to the development of the concept of belief in Theory-Theory of Mind research (henceforth, TToM) (Wellman, Cross, & Watson, 2001; Astington, 2001). In fact, quite often, metarepresentation (defined here as the internal representation of an epistemic relation (Leslie, 1991), a "second-order" representation (Sperber, 1999)) has been somewhat fused with the child's ability to entertain/ascribe a belief that stands for a counterfactual state of affairs (Dennett, 1978a; Davies & Stone, 1995), without any reference to the explanatory power of the concept of desire or to its being an epistemic relation itself.

This incipient collapsing of metarepresentation and reasoning by false beliefs (beliefs that stand for counterfactual states of affairs) in TToM research is due, in part, to the assumption that the concept of belief, and especially that of false belief, taps the child's metarepresentational capacities; while, arguably, the concept of desire does not. This is a common assumption despite the fact that beliefs and desires share several important characteristics (such as defining opaque contexts, being intentional in the philosophical sense, being subject and object specific (Wellman & Woolley, 1990), etc. to name but a few). In the developmental literature, more often than not, desires have been understood as a special case of mental state ascription; namely, one that does not demand of the agent doing the ascription (a child in our case), that he or she be able to metarepresent.

The previous assumption appears to have been brought about by a body of evidence that suggests that the concept of desire is acquired around a year before the concept of belief (Wellman, 1991; Tan & Harris, 1991; Astington & Gopnik, 1991; Harris, 1996; Bartsch & Wellman, 1995). Thus, in order to make it fit the evidence available from research on the concept of belief, it has been claimed that at least until three or four years of age –that is, until they acquire the concept of belief– desire ascriptions are to be thought of as non-metarepresentational. The reasoning behind this claim, I presume, goes along the following lines: on the one hand, a) reasoning by means of beliefs taps an organism's (children's, for example) metarepresentational capacities; on the other hand, b) the capacity to reason by means of beliefs is acquired at age X. Thus, if (a) and (b) hold, then c) metarepresentation should be acquired at age X, and not before X. Now, *because* of this conclusion, the rest of the argument tells us that if (c) is the case, then e) reasoning by means of desires taps an organism's metarepresentational capacities if and only if f) the capacity to reason by means of beliefs is acquired at age X. However, it is not the case that (f). Therefore, the argument goes on to say, reasoning by means desire does not tap an organism's metarepresentational capacities.

However, characterizing reasoning by means of the concept of desire as non-metarepresentational simply because it is acquired before the concept of belief and the latter is, only by assumption (as we saw a couple of paragraphs above), the flagship of the child's metarepresentational capacities is, at all extents, an *ad hoc* solution. In these pages, I will take position against the conclusion of the argument that reasoning by means of desire does not tap on metarepresentational capacities. In other words, it is, I will argue, far from clear that reasoning by means of the concept of desire is non-metarepresentational in nature, even though it is acquired a year before the concept of belief.

Desires as Metarepresentational

For reasoning by means of desires to be non-metarepresentational at ages younger than four years means that a child at those ages does not represent other people (or even themselves) as representing the desired object as part of the desire relation. Thus, these children are supposed to be merely in some sort of "connection"

to the desired object. According to this view, then, when a two-year-old child says something like "Peter wants a car", he or she is not ascribing Peter (representing Peter as entertaining) a desire for a given car, but merely putting Peter in some sort of connection to either the car he wants (Wellman, 1990; Wellman & Bartsch, 1994; Bartsch & Wellman, 1995) or to a hypothetical situation in which Peter has the car (Perner, 1991). Consider the following proposition as a state of affairs in the world; that is, a proposition that holds:

Loreto wants to be in Chile. [1]

Within the general Theory of Mind (ToM) framework, there are two ways to interpret this proposition, and they have usually been confounded. In a first, trivial interpretation, Loreto is just in a state such that she wants to be in Chile. That merely means that she has tokened the proposition "I am in Chile" in her desire box (Fodor, 1975) and will take action to bring it about that she is in Chile. This interpretation is useless at the time of explaining behavior because the agent trying to explain Loreto's behavior (maybe Loreto herself) may not know that *that* is the proposition she is tokening in her desire box. The second interpretation, however, is the non-trivial interpretation that to be able to explain someone else's (even one's own) behavior in terms of desires, one must entertain a belief about the organism's desire state (Davies & Stone, 1995). To explain or predict behaviors and actions, it is not enough that we are able to be in desire states (unlike, for example, the case of Simulation-Theory of Mind, see Gordon (1995)). What is a conditio sine qua non is that we are able to entertain beliefs about an organism's desire states (Dennett, 1987; Sperber, 1999). For example, one way to explain why Loreto is buying a ticket to Chile this morning is to entertain a belief with the embedded proposition in [1] above. Thus, in order to engage in folk psychological practice (Davies & Stone, 1995), we need to entertain a thought along the following lines:

$$B_u[D(Loreto, P) \wedge \neg P] \quad [2]$$

where B_u stands for the agent's belief state at the time of ascription of the desire state, D stands for the "desire" predicate which takes two arguments, the organism to which the agent is ascribing the desire ($Loreto$, in [2]) and the organism's desired state of affairs (the variable P in [2] or "I am in Chile" or any other proposition). For this quasi-formalization of desire ascription to work, P should also be part of the belief state as a proposition that does not hold; since, for something to be a desire, it is by definition that the conditions are false. Simplifying the issue slightly, it would indeed be a contradiction to desire something that one already has.

Notice further that there is in fact no way to formalize the first (trivial) interpretation of [1] above in the TToM framework. You may be able to formalize it for logical purposes as something like $D(Loreto, P) \wedge \neg P$, but that will be of no use to someone trying to explain behavior. The proposition in [1] is independent of any folk psychological theory-theory because it is not tokened as a belief about the world in the mind of a particular agent engaging in folk psychological practice. It is just a true proposition (of the external world) at time T. Thus, part of the argument here is that if it is so difficult for us as adults to imagine a non-metarepresentational account of desire at early ages, then it might be the case that this non-metarepresentational characterization is wrong (Astington & Gopnik, 1991).

There seems to be no obvious alternative formalizing of desire ascription (not desire states) except for [2] above. Thus, even at younger ages, every time children talk about their own or other people's desires, they should be entertaining a thought along the lines of [2]. It is hard to characterize the thoughts the child is entertaining when explaining or reporting behaviors by means of desires when the latter are merely understood as "subjective connections" to objects. Suppose for the sake of argument, that children do in fact see desires as a "subjective connection" between the organism they are trying to explain the behavior of and the object that this organism desires. This could be relatively easy to see for desire ascription to organisms other than self. However, it would be hard to believe that when talking about their own desires and explaining their own behaviors by means of desires ("because I wanted to go to the park"), children think of *themselves* as just holding a subjective connection to a state of affairs that does not hold. It is in fact very hard to believe that when reasoning about their own behaviors by means of desires, children are not representing themselves as wanting something in particular; to be in the park, for example. But suppose further that they do not representing themselves as wanting something in particular. It is undeniable that the very act of communicating those desires involve a metarepresentation of both the communicator and the person the speaker is talking to. When communicating, and more so when communicating mental states, there should be mutual metarepresentation of the communicator and the addressee (Sperber, 1999).

Given the arguments above, it is hard to take desire ascriptions to either other people ("Peter wants to have a car") or to oneself ("I want to be in Chile") as non-metarepresentational (at least) in Leslie's (1991) terms. Now, assuming desire talk stands proxy for desire reasoning about the behavior of other people (see, for example, Dennett (1978a), Bretherton and Beeghly (1982), Tager-Flusberg (1993), Bartsch and Wellman (1995) and the literature spawned by these studies), then we would expect desire talk to actually tap on the child's metarepresentational capacities, albeit indirectly. One way to look at this is to follow Wellman and Bartsch (1994) and Bartsch and Wellman (1995). If we were able to tell genuine psychological references to desire apart from mere communicative uses of particular words associated to the expression of desire, then we should find difference between talk about desires and talk about other communicative uses of these words. Specifically, while genuine psychological references to desire should change

as a function of age (because they are presumably tapping on metarepresentational abilities), communicative uses should not. They should not because there is no need to be in a belief state about an epistemic state when using a mental state term for mere communicative purposes. If a child repeats an adult's utterance, for example, it may be that he or she is just repeating it for the sake of not being silent during an interaction, but without having analyzed the utterance itself. More arguably, when a child says something like "I want a cookie" while the cookie is in plain view, the child might be actually saying something like "pass me the cookie", without imputing a mental state either to self or to someone else. Instead of entertaining a thought such as $B_u[D(self, IHaveACookie) \wedge \neg IHaveACookie]$, the child is simply in a desire state, maybe, such that $D_u(self, IHaveACookie) \wedge \neg IHaveACookie$. But that does not qualify as a metarepresentational state in our terms here. The following study tests the prediction that there should be a difference between communicative uses of "want" (the desire term par excellance (Wellman & Bartsch, 1994; Bartsch & Wellman, 1995)) and its genuinely psychological uses.

Method

Data. A total of 14,896 child utterances were taken from the Wells corpus (Wells, 1981) in the CHILDES database (MacWhinney, 2000). These data come from the longitudinal observation of spontaneous speech production of 12 children (6 boys and 6 girls) whose ages ranged from 18 months at the time of the first observation to 60 months at the time of the last observation and who were acquiring English as their mother tongue. Each child was observed a total of 10 times, for about 40 minutes each, in 3-month intervals. Since the objective of Well's (1981) research was to obtain spontaneous speech samples, a timing mechanism was devised to set off a tape recorder – connected to a wireless microphone in the child's garment – at different times between 9am and 6pm, to prevent parents from planning activities, for example. Twenty-four 90-second samples were recorded in each observation. These were later transcribed into several files using normal English orthography. Table 1 in page 3 gives some general information about the samples, where $Ages(mo.)$ means ages in months, N means number of participants in each age group (all twelve participants are the same children at different ages), $\#TotUtt$ means the total number of utterances in the samples, $MLU(\overline{x})$ means the average mean length of utterance for that age group, and $MLU(SD)$ means the standard deviation of the mean for the MLU values for that group.

Procedure. The first step was to identify all and only the instances of the term "want" in all and only the target child's exchanges. Once the "want" utterances were identified and cleaned for false positives, they were coded as belonging to one of five mutually exclusive categories: genuine psychological references to desire (GPRDs), behavioral requests(BRs), direct repetitions (DRs), idiomatic expressions (IEs) and uncodable utterances (UUs). While GPRDs refer to mental states, the other three categories do not, they fulfill a mostly communicative function.

Table 1: Information on the samples.

Ages (mo.)	N	#TotUtt	MLU($\overline{x}$)	MLU(SD)
18-24	12	3483	1.480	0.252
25-28	12	1830	1.697	0.430
29-32	12	2694	2.246	0.566
33-36	12	2972	2.709	0.446
37-40	12	2041	2.981	0.390
41-44	12	1876	3.202	0.425

Genuine psychological references to desire (henceforth, GPRDs) are instances of children's unequivocally referring to themselves or other people as being in a mental state of desire. Behavioral requests, in turn, (henceforth, BRs) are 'unadorned' instances in which the child uses a desire term to fulfill an immediate goal, like receiving something that is beyond her reach but in plain view. Bartsch and Wellman (1995) take these instances to mean nothing more than "give me x". Direct repetitions (henceforth, DRs) are dialog turns in which the child merely repeats the adult (or his own) utterance. Idiomatic expressions (henceforth, IEs) are high-frequency collocation of words in the particular language. This is the case of Spanish "I don't want to" or "I want more", when they appear without an object. Uncodable utterances (henceforth, UU) are instances of the desire term "want" for which categorization was impossible, due mainly to failure in retrieving contextual information from the dialog.

To code each of them into one of the five mutually exclusive categories, child utterances containing "want" were not taken in isolation, but embedded in a window of the four previous and the four following utterances of the whole sample. However, sometimes this short context did not help defining which category the utterance belonged to. Thus, the whole transcript had to be analyzed in order to assign a category to the utterance in question. An independent rater, unaware of the hypothesis of the study rated a subset of the data (10%=60 utterances, Cohen's κ=.85). Disagreements were resolved by discussion and, in the light of the discussions, there was a second coding pass to the whole data set.

Results

There were a total of 602 "want" utterances in the analyzed corpus. 347 (57.35%) were GPRDs, 145 (23.96) were other communicative uses of "want". Of those 145 communicative uses, 37 (6.11% of total "want" utterances) were behavioral requests, 92 (15.20%) were direct repetitions and 16 (2.65%) were idiomatic expressions. Figure 1 below shows the average frequency of talk about genuine desires as a percentage of the total number of utterances for each particular child at each particular age. It is evident that the developmental picture I have obtained resembles the one in Bartsch and Wellman (1995) very closely, even the ranges of the percentages are similar (see Bartsch and Wellman (1995),

p. 73, Figure 4.2B). Talk about genuine desire seems to be present at the first age analyzed (AGE1, 18-24 months), to increase slowly by AGE2 (25-28 months of age) and then more drastically again at AGE3 (29-32 months). The frequency of genuine talk about desires seems to peak at around AGE4 (33-36 months) to drop and stabilize thereafter. In order to test for significant differences in children's talk about desires at any given age group, a repeated measures ANOVA with age as a 6-level variable was used. There was an overall significant main effect $F(5, 55)=6.72$, p<.000. Post hoc analyses using the Bonferroni criterion for significance indicated that the average frequency of GPRDs for AGE1 (M=0.51, SD=0.55) and AGE2 (M=0.96, SD=0.80) were significantly lower than frequency of talk about GPRDs at AGE4 (M=4.28, SD=2.24) and AGE6 (M=3.54, SD=2.10).

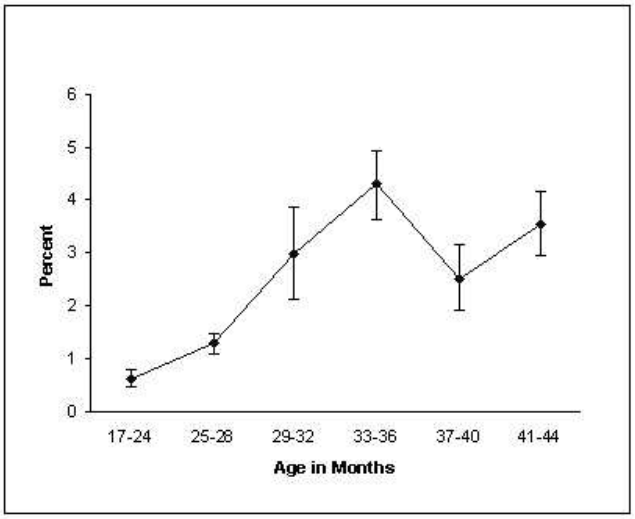


Figure 1: GPRDs by age as a percentage of the total number of utterances for that child at that age, error bars are standard errors.

Figure 2 below shows the development of children's communicative uses of "want" as a function of age. Except for direct repetitions, both idiomatic expressions and behavioral requests do not appear in the first age stage sampled (AGE1=18-24). By AGE2, all three categories of communicative uses are present, although not extremely different from the previous age. The frequency of DRs seems to grow and separate from the main trend at AGE3 and then again at AGE4, while at AGE3 both BRs and IEs are at the same level. Something indeed seems to happen at AGE4, when all three categories seem quite different in their frequencies, with DRs leading the frequency count, followed by BRs and IE in the last place. Both AGE5 and AGE5 seem to show the new convergence of these categories. It seems then that after AGE4, all three kinds of communicative uses of "want" stabilize. To test for significant effects, a repeated measures ANOVA with age as the within-subject variable was carried out for each communicative use of "want". The tests show that, taken one by one, there is no significant effects for age and each of the communicative uses of "want", $p > .05$. However, a repeated measures ANOVA with age and communicative uses as within-subjects variables yielded a significant effect for both age $F(2.96, 32.61)$[1]$=3.174$, p=.038 and communicative uses $F(2, 22)=12.708$, p<.000. No main effect was found for the interaction between age and communicative uses $F(3.36, 36.97)$[1]$=0.682$, p>.05, n.s.

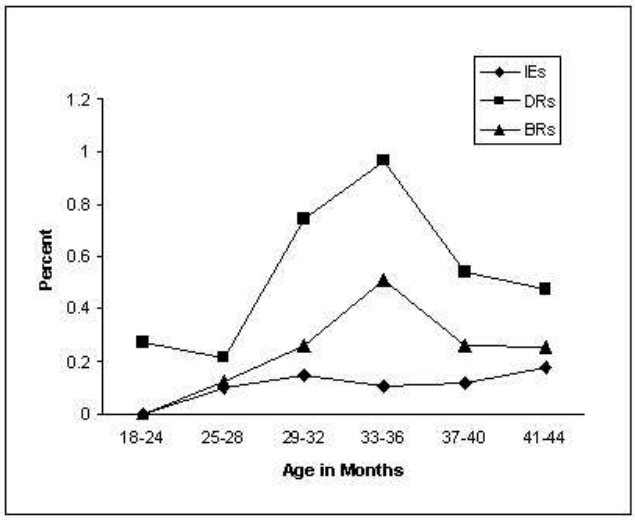


Figure 2: Communicative uses of "want" by age as a percentage of the total number of utterances for that child at that age.

Figure 3 shows quite clearly that although both GPRDs and all communicative uses start at roughly the same frequency, it is only GPRDs that increase the frequency significantly, while the other communicative uses of "want" stay roughly the same across ages. A repeated measures ANOVA with AGE as a 6-level variable (Ages 1 through 6) and communicative uses as a 4-level variable (GPRDs, BRs, IE, DRs) was used to test for differences. As expected from the previous analyses, there was an overall significant main effect of communicative uses, $F(3, 33)=59.545$, p<.000, a significant main effect for age, $F(5, 55)=7.444$, p<.000 and a significant interaction of Age and Communicative uses, $F(3, 33)=4.861$, p<.000.

Discussion

From the theoretical discussion above, we concluded that if children undergo some metarepresentational change of the concept of desire as a function of age, then while the developmental picture of GPRDs reflect this change, communicative uses of "want" should stay relatively the same across ages.

The analyses carried out yield some results that point towards this direction. Although there are differences among the communicative uses themselves (that is, there are differences between DRs and IEs, for instance, at 33 months, see Figure 2), there is no main effect for

[1] F corrected for sphericity by Greenhouse-Geisser.

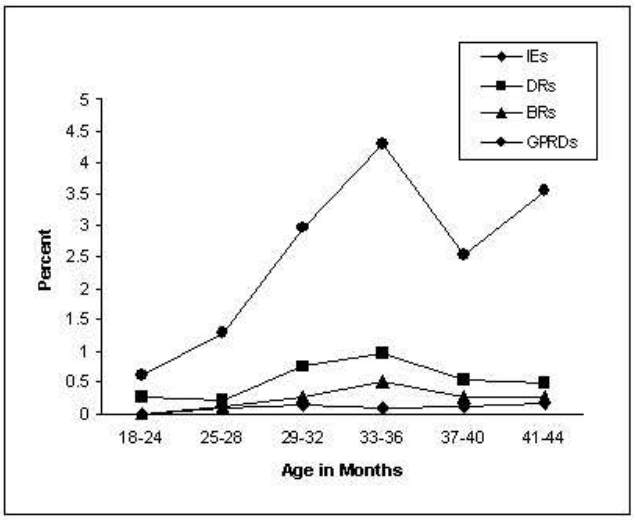


Figure 3: Communicative uses of "want" by age as a percentage of the total number of utterances for that child at that age.

age for each of these communicative uses taken in isolation. However, there is a main effect for GPRDs between the first age analyzed (18-24 months) and AGE4 (33-36 months) and AGE6 (41-44 months). This, obviously, draws a difference between the communicative uses of "want" and genuine psychological references to desire. This difference, I am inclined to say, may be related to metarepresentational issues during the acquisition of the concept of desire. If it were not about matters metarepresentational, it would be very difficult to explain why the other communicative uses (BRs, in particular), which are morphosyntactically very similar to GPRDs, do not provide a main effect for age.

This paper is not calling into question the hypothesis that belief (and particularly false belief) taps an organism's representational capacities, nor that belief is acquired at whatever age (probably, on all conservative accounts, at 4 years of age). What is being questioned here is the working assumption that if belief taps on representation and belief is acquired at 4 years of age, then metarepresentation is acquired at 4 years of age.

The argument that uses the premise above seems to be, at least, enthymematic. It could be said that reasoning by means of belief is tapping certain kinds of metarepresentational abilities, the kinds for which many computational resources have to be in place (Wimmer, Hogrefe, & Perner, 1988; Leslie, 1988; Davies & Stone, 1995). As an analogy, you can take the difference that exists between a belief attribution such as "Loreto believes there's a blue car outside school" and "Loreto believes Namic thinks there's a blue car outside school". Terminological differences aside, children seem to acquire the ability to solve problems like the latter by around 6 years of age (Perner & Wimmer, 1985), two years after they have allegedly acquired the ability to metarepresent (metarepresent by false belief, that is). However, just because of this empirical fact, one would not argue that by passing this more complicated task, the child has now acquired another ability, one different from the metarepresentational abilities acquired two years earlier. The same argument holds for desires: just because children talk and reason by means of desired a year before they do so with belief, that does not mean that a new ability has been acquired. That children are able to calculate this double-embedding of *the same belief concept* a couple of years later than 4 years of age points in the direction of problems with some of the computational mechanisms that help children deal with metarepresentation (Fodor, 1992), but not with the ability to metarepresent itself.

General conclusions and future work

The main point of this paper is that it is extremely hard to consider reasoning by means of desire (at any age stage) as a non-metarepresentational endeavor. This hypothesis has been analyzed in two ways: a) by means of a logical analysis of what is involved during desire reasoning and communication and b) by providing some preliminary empirical evidence that even talk (as a proxy for reasoning) about desire shows a clear developmental trend when compared to communicative uses of the same words used to talk about desire ("want", in this case).

If the main point of this paper is right, then the most pressing issue to deal with is the lag between the acquisition of the concept of belief and that of desire. In other words, if metarepresentation lies in the nature of both belief and desire but children have less difficulty understanding the representational nature of the latter while failing to understand the equivalent metarepresentational character of the former (Astington & Gopnik, 1991), then again it may be the case that something other than metarepresentation is at stake. Of course, much more work is needed in this area. Nonetheless, I would like to propose that the answer to this riddle lies in the computational mechanisms dealing with metarepresentation at the different stages. Not with metarepresentational abilities themselves. In other words, I would like to propose that the ability to metarepresent is acquired as soon as children start talking about and reliably communicating their own and other people's mental states, starting with desire at around the 30th month of life. This is somewhat earlier than previously thought, but it would help explain and make sense of the whole philosophical tradition of belief and desires as belonging to roughly the same theoretical arena as the rest of the propositional attitudes.

Acknowledgments

This paper was written while the author was holding a postdoctoral fellowship at the Human-Oriented Technology Laboratory at Carleton University. I would like to thank, particularly, John Logan, Andrew Brook and Gitte Lindgaard for insightful comments and support.

References

Astington, J. (2001). The future of theory-of-mind research: Understanding motivational states, the role of language, and real-world consequences.

Commentary on "Meta-analysis of theory-of-mind development: The truth about false belief.". *Child Development*, *72*(3), 685-687.

Astington, J., & Gopnik, A. (1991). Developing understanding of desire and intention. In A. Whiten (Ed.), *Natural theories of mind* (p. 39-50). Oxford, UK: Blackwell.

Bartsch, K., & Wellman, H. (1995). *Children talk about the mind.* New York: Oxford University Press.

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology*, *18*(6), 906-921.

Davies, M., & Stone, T. (Eds.). (1995). *Mental simulation.* Oxford: Blackwell.

Dennett, D. (1978a). Beliefs about beliefs:. *Behavioral and Brain Sciences*, *1*(4), 568-570.

Dennett, D. (1978b). *Brainstorms: Philosophical essays on mind and psychology.* Montgomery, VT.: Bradford Books.

Dennett, D. (1987). *The intentional stance.* Cambridge, MA.: MIT Press.

Fodor, J. (1975). *The language of thought.* New York, NY: Crowell.

Fodor, J. (1987). *Psychosemantics.* Cambridge, MA: MIT Press.

Fodor, J. (1992). A theory of the child's theory of mind. *Cognition*, *44*, 283-296.

Gordon, R. (1995). Folk psychology as simulation. In M. Davies & T. Stone (Eds.), *Folk psychology.* Oxford: Blackwell.

Harris, P. (1996). Desires, beliefs and language. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind.* Cambridge: Cambridge University Press.

Leslie, A. (1988). Some implications of pretense for mechanisms underlying the child's theory of mind. In J. Astington, P. Harris, & D. Olson (Eds.), *Developing theories of mind* (p. 19-46). Cambridge University Press.

Leslie, A. (1991). The theory of mind impairment in autism. In A. Whiten (Ed.), *Natural theories of mind* (p. 63-78). Oxford, UK: Blackwell.

MacWhinney, B. (2000). *The childes project : Tools for analyzing talk* (3rd ed.). Hillsdale, N.J.: L. Erlbaum.

Perner, J. (1991). On representing that: The asymmetry between belief and intention in children's theory of mind. In D. Frye & C. Moore (Eds.), *Children's theories of mind* (p. 139-155). Hillsdale, NJ: Erlbaum.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that": Attribution of second-order beliefs by 5-to 10-year-old children. *Journal of Experimental Child Psychology*, *39*, 437-471.

Sperber, D. (1999). Metarepresentation. In R. A. Wilson & F. C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences.* Cambridge, MA.: MIT Press: Bradford Books.

Tager-Flusberg, H. (1993). What language reveals about the understanding of minds in children with autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism.* Oxford: Oxford University Press.

Tan, J., & Harris, P. (1991). Autistic children understand seeing and wanting. *Development-and-Psychopathology*, *3*(2), 163-174.

Wellman, H. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Wellman, H. (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading* (p. 19-38). Oxford: Blackwell.

Wellman, H., & Bartsch, K. (1994). Before belief: Children's early psychological theory. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind.* Hove: Lawrence Earlbaum.

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development: The truth about false-belief. *Child Development*, *72*(3), 655-684.

Wellman, H., & Woolley, J. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*, 245-275.

Wells, C. (1981). *Learning through interaction: The study of language development.* Cambridge, UK: Cambridge University Press.

Wimmer, H., Hogrefe, J., & Perner, J. (1988). Children's understanding of informational access as source of knowledge. *Child Development*, *59*, 386-397.

Categorization and Memory: Representation of Category Information Increases Memory Intrusions

Anna V. Fisher (fisher.449@osu.edu)
Department of Psychology & Center for Cognitive Science
Ohio State University
208B Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210 USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210 USA

Abstract

False recognition of verbal information has long been established with word lists. Current research examines the phenomenon of false recognition with pictorial stimuli. Experiment 1 demonstrated that similar to word-lists, pictorially presented information elicits memory intrusions, and that rates of intrusions differ across stimuli sets. Experiment 2 investigated the effects of focusing on category-level versus item-specific information on the rates of false recognition. Results of Experiment 2 suggest that memory accuracy decreases dramatically when participants perform category-based processing compared to item-based processing. Experiment 3 confirmed that processing manipulations rather than other extraneous factors influence levels of false recognition in Experiment 2.

Introduction

People strive for accurate and reliable memories; however their memories often get distorted. Although forgetting is one of the most obvious types of memory distortions, it is not the only one. There is much research demonstrating that people often distort memories in systematic and predictable ways. For example, prior knowledge has previously been implicated in memory distortions: people often falsely recognize new information when it is consistent with their knowledge (e.g., Alba & Hasher, 1983). For instance, after reading a story describing a famous person, participants tended to falsely recognize statements that were not part of the story they had read, but were thematically related to this person (Sulin & Dooling, 1974).

Systematic memory distortions are not limited to sentence information, and are often found with word lists. These types of memory distortions were first demonstrated by Deese (1959), who presented participants with word-lists (e.g., "bed", "rest", and "awake") consisting of associates of a single non-presented word (e.g., "sleep"). When asked to recall the words from the list, participants often erroneously recalled words consistent with the overall theme of the list, which was never actually presented. Deese's fundings were followed up by Roediger and McDermott (1995), who replicated Deese's results, demonstrating that memory intrusions of non-presented words persist in recall as well as in recognition, thus giving the name of DRM (for Deese-Roediger-McDermott) to this phenomenon. However, the nature of the phenomenon is still unclear.

According to one explanation, during recognition, participants perceive both studied items, and semantically related critical lures, to be more familiar than unrelated distracters. Because familiarity strongly affects the decision criterion for accepting items as studied or "old", those items that have elevated familiarity are more likely to be accepted both correctly and erroneously. This increased familiarity may stem from a summary or "gist" representation that reflects the general meaning of the list (in addition to representing individual items in the list), with critical lures being consistent with the gist (Brainerd, Reyna, & Mojardin, 1999). Therefore, on a recognition test, item-specific representations drive hits or correct acceptance of studied items, whereas "gist" representations drive both hits and false alarms on critical lures (i.e., erroneous acceptance of items semantically related to studied items).

According to another explanation, processing of items (either at study or at test) activates a critical lure, an item strongly associated with studied items. However, during recognition participants fail to monitor the source of this activated information, and as a result of confusing internally generated and externally presented information, participants falsely recognize critical lures (Gallo & Roediger, 2002; Koutstaal & Schacter, 1997; Roediger, et al., 2001).

If participants can represent both the "gist" and individual items (Brainerd, et al., 1999), and distortions (or false alarms) are driven by the gist representations, then it should be possible to facilitate the formation of either representation by focusing participants on the overall theme or on individual items. If our contention is correct, then a manipulation focusing on a gist representation should lead to elevated memory distortions (due to an elevated level of false alarms on critical lures).

This manipulation can generate evidence capable of distinguishing between the two theoretical positions because

the source confusion explanation does not predict these effects.

Note, however that much of DRM-based research has been based on word-lists rather than pictures (see Koutstaal & Schacter, 1997; Seamon Luo, Schlegel, Greene, & Goldenberg, 2000, for notable exceptions). At the same time, pictures are well suited for this manipulation: participants could be focused on an entire category (e.g., *Cats*) or on individual items, such as a picture of a particular cat.

Another advantage of pictorially presented information is that pictures can drastically decrease the tendency to make source monitoring errors: it is highly unlikely that one would spontaneously generate a particular unique picture serving as a critical lure (see Koutstaal & Schacter, 1997, for related arguments). Therefore, persistence of memory intrusions with pictures would further suggest that these intrusions do not stem solely from source monitoring errors.

It has been previously demonstrated that pictures do generate memory intrusions (Koutstaal & Schacter, 1997; Seamon, et al., 2000). However some of these findings are based on a procedure that used large 120-study-item stimuli sets, and a 3-day delay between the study and the recognition phases. In this research, we will use the procedure that follows more closely the DRM procedure with word lists: we present participants with a reasonably small stimulus set, and impose no delays between the study and recognition phases. We have demonstrated elsewhere (Sloutsky & Fisher, in press) that false recognition of information presented pictorially can be obtained in a procedure that closely follows the original DRM task. However, these results were obtained with a single picture set, and it is unclear how well they can be generalized to a greater number of categories.

Overall, the reported research has two goals: (1) to examine whether or not pictorially-presented information can generate DRM-type phenomena and, if yes, then (2) whether false recognition in DRM stems from source-monitoring errors or from gist representation of information.

The goal of Experiment 1 was to replicate these findings using multiple categories. Experiment 2 had a more theoretically important goal: to generate evidence capable of distinguishing among the proposed theoretical accounts of the DRM-effect. Recall that finding increased memory intrusions as a result of processing manipulations would support the position that DRM-type memory intrusions stem from gist-type representations, while weakening the position that these intrusions stem from source-monitoring errors.

Experiment 1

Method

Participants Participants were introductory psychology students at a large Midwestern university (N = 103, *M* age= 19.7 years, *SD* = 1.8 years; 51 women and 52 men) who received a partial course credit for participation.

Design, Materials and Procedure Materials were 90 color photographs of animals presented against a white background. The photographs represented five different animal categories (*Cats*, *Bears*, *Squirrels*, *Fish*, and *Birds*) with 18 photographs per category.

The task consisted of a study and a recognition phase. During the study phase participants were presented with 30 pictures from three different animal categories: 10 items from the Target category, and 20 items from the two Filler categories. Participants were instructed to remember the presented pictures as accurately as possible for a f uture recognition test. During the recognition phase, which immediately followed the study phase, participants were presented with 28 pictures: 14 previously studied pictures (7 from the Target category and 7 from one of the Filler categories), and 14 new pictures (7 new pictures from the Target category, and 7 pictures from a novel category, which served as control items). Participants were asked to determine whether each picture presented during the recognition phase was "old" (i.e., exactly the same as previously seen in the study phase) or "new".

The categories that were designated to be Targets were rotated such that *Cats*, *Bears*, *Squirrels*, *Fish*, and *Birds* served as Targets in one of five between-subject conditions. All participants were tested individually, and had all instructions and stimuli presented to them on a computer screen in a self-paced manner.

Results and Discussion

Some participants did not reliably reject control items (i.e., at least 5 out of 7 correct), and their data were excluded from further analyses; 11 participants were excluded overall. The rest of the participants were very accurate in recognizing previously studied items (on average over 88% of correct recognitions across the target categories) and in rejecting items from novel categories (over 97% of correct rejections across categories).

Most importantly, participants often mistakenly recognized new items from the Target category, or critical lures. Proportions of "old" responses to previously studied items from the Target category (Hits), to new items from the Target category (False Alarms), and to novel items from a novel category (Control Items) are presented in Figure 1.

Data in the figure indicate that (a) although for all categories, the proportion of Hits was significantly higher than the proportion of False Alarms (FA), all paired-sample *t*s > 6, *p*s < .0001, some pictorially presented categories (e.g., *Bears*) elicited sizable memory intrusions (Hits = .86, FA = .54, Hits –FA = .32), and (b) proportions of memory intrusions varied across the categories, ranging from relatively high for *Bears* to almost non-existent for *Birds*. To examine the significance of differences across categories, accuracy measures (i.e., Hits – FA) were subjected to a one-way between-subjects ANOVA with Target category as a factor. The results point to significant differences in accuracy across the Target categories, F (4, 91) = 27.3, MSE = 0.04, p < .0001, with the following

pattern of accuracy *Birds* > *Fish* = *Squirrels* = *Cats* > *Bears*, post-hoc Tukey test, for all differences *p*s < .05. Therefore, major DRM phenomena that were previously found with word-lists are replicable with pictures. First, pictures generated substantial levels of memory intrusions. And second, similar to word-lists, pictures elicited different levels of memory intrusions across different target conditions: while little false recognition occurred for the Target category *Birds,* other Target categories (*Bears, Cats, Fish, and Squirrels*) elicited sizeable levels of false recognition.

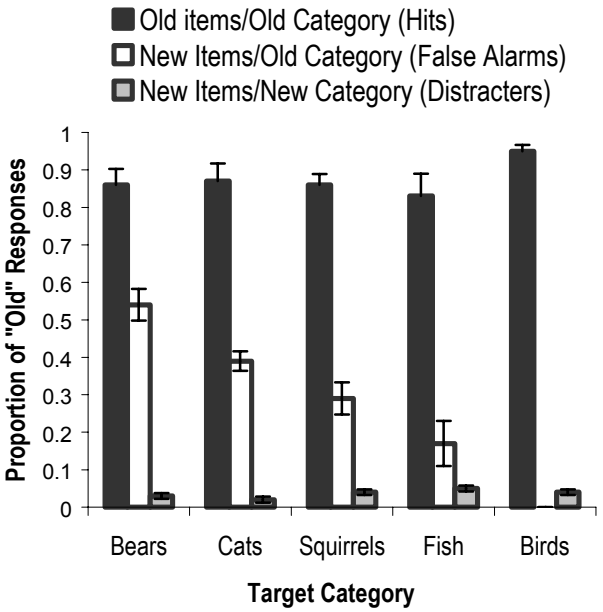


Figure 1: Mean proportions of "Old" responses across target types in the recognition memory test of Experiment 1

It could be argued, however, that Birds were an odd category that stood out in the context of mammals, which resulted in a more accurate processing of this odd category. To test this alternative, we conducted an additional experiment, using *Cats* as the Target category presented in the context of reptiles, with *Frogs* and *Alligators* used as Filler Categories. When *Cats* were an odd category, the level of false recognition of critical lures was statistically equivalent to that in Experiment 1 (42% versus 39%, respectively).

It is also possible that during the study phase, participants spontaneously labeled species of birds (but not cats, squirrels, fish, or bears), which dramatically reduced memory intrusions for the Target category *Birds*. To rule out this possibility we asked 23 undergraduates to label pictures of birds used in Experiment 1: no more than 4 out of 18 birds received unique labels, which was comparable to the labeling of cats.

Overall, these results suggest that DRM-type memory distortions are a robust phenomenon independent of the mode of presentation. More importantly, the fact that patterns of memory intrusions are similar for verbally and pictorially presented stimuli suggests that DRM-type memory intrusions do not stem solely from source monitoring errors. The goal of Experiment 2 was to examine directly whether a task that facilitates category-level processing will lead to an increase in memory intrusions compared to the baseline of Experiment 1.

Experiment 2

Method

Participants Participants were introductory psychology students at a large Midwestern university (N = 134, *M* age= 20.26 years, *SD* = 2.5 years; 64 women and 70 men) who received a partial credit for participation.

Design, Materials and Procedure Materials in Experiment 2 were identical to Experiment 1, however the study phase of the experiment was different. Participants were first presented with a picture of an animal from a Target category, and informed that the animal had "beta-cells inside its body". Participants were then presented with 30 pictures (10 from the Target category and 20 from two Filler categories), and asked to determine whether each presented animal also had beta-cells inside. Participants were provided with feedback, which indicated that only animals from the Target category had the property in question, whereas animals from the Filler categories did not. Participants were not warned about an upcoming recognition test.

Similar to Experiment 1, during the recognition phase participants were presented with 28 pictures: 14 previously studied pictures (7 from the Target category and 7 from one of the Filler categories), and 14 new pictures (7 new pictures from the Target category, and 7 pictures from a novel category). Participants were asked to determine whether each picture presented during the recognition phase was "old" (i.e., exactly the same as previously seen in the study phase) or "new".

The categories that were designated to be Targets were rotated such that *Cats*, *Bears*, *Squirrels*, *Fish,* and *Birds* served as Targets in one of five between-subject conditions. All participants were tested individually, and had all instructions and stimuli presented to them on a computer screen in a self-paced manner.

Results and Discussion

Some participants did not reliably reject control items (i.e., at least 5 out of 7 correct), and their data were excluded from further analyses; 35 participants were excluded overall. The rest of the participants were very accurate in recognizing previously studied items (on average over 83% of correct recognitions across the target categories) and in rejecting items from novel categories (over 97% of correct rejections across categories).

However, the rates of false recognition in each target condition increased substantially, compared to the baseline

in Experiment 1. Proportions of hits (i.e., correct recognitions), false alarms on Target distracters (FA), and accuracy scores (Hits – FA) for each target category are presented in Table 1.

Table 1: Proportions of false alarms (FA), and Accuracy scores (Hits – FA) across target categories in Experiment 2.

Target Category	Hits	FA	Accuracy (Hits – FA)
Birds	.84	.50	.34
Fish	.80	.47	.33
Squirrels	.80	.61	.19
Cats	.80	.62	.18
Bears	.88	.79	.09

Overall results of Experiments 1 and 2 are presented in Figure 2. Data in the figure indicate that recognition accuracy markedly decreased in Experiment 2 compared to Experiment 1. This differential accuracy was the result of a processing manipulation introduced in Experiment 2 (i.e., an induction task) that focused participants on the category-level properties of stimuli, as opposed to item-specific properties in Experiment 1. Data in the figure also suggest that the decrease in accuracy in Experiment 2 was not proportional to the level of performance in Experiment 1. This task by condition interaction, F (4, 181) = 3.7, MSE = .21, $p < .05$, suggests that when participants are focused on category-level properties, they form mainly category-level representations, as opposed to mainly item-level representation in the Baseline. However, it is possible that decrease in memory accuracy obtained in Experiment 2 can be explained by increased task demands of Experiment 2 compared to Experiment 1 (performing an induction task versus no task during the study phase). It is also possible that overall accuracy in Experiment 2 decreased because participants were not warned about a subsequent memory test. Experiment 3 was designed to test these alternative explanations.

Experiment 3

The goal of Experiment 3 was to eliminate potential confounds of Experiment 2 by introducing a task that would force participants to engage in item-based processing. Similar to Experiment 2 participants were not warned about a subsequent memory test. Therefore, if accurate memory performance is obtained in Experiment 3, both of alternative explanations for the results of Experiment 2 will be eliminated.

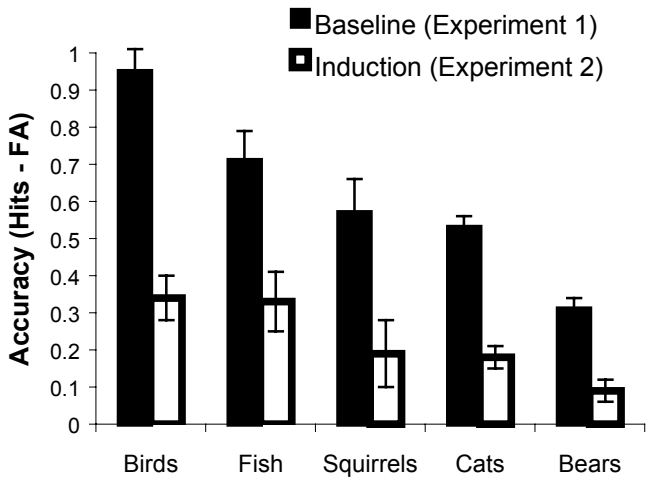


Figure 2: Mean accuracy (Hits – FA) across target categories in Experiment 1 and 2.

Method

Participants Participants were introductory psychology students at a large Midwestern university (N = 24, M age= 20.2 years, SD = 1.6 years; 9 women and 15 men) who received a partial credit for participation.

Design, Materials and Procedure Overall structure of the task was similar to Experiment 2, however only one target category (*Cats*) was tested, and participants were presented with a different question during the study phase. Participants were first presented with a picture of a cat, and told the animal was young. Then they were presented with 30 pictures of animals from 3 different categories (10 cats, 10 bears, and 10 birds), and asked to determine whether each animal was young or mature. Participants received random feedback, thus blocking any possible categorization. Similar to Experiment 2, participants were not warned about a subsequent memory test.

During the recognition phase participants were presented with 28 pictures: 14 previously studied pictures (7 cats and 7 bears), and 14 new pictures (7 novel cats and 7 squirrels, which served as control items). Participants were asked to determine whether each picture presented during the recognition phase was "old" (i.e., exactly the same as previously seen in the study phase) or "new".

All participants were tested individually, and had all instructions and stimuli presented to them on a computer screen in a self-paced manner.

Results and Discussion

Three participants did not reliably reject control items (i.e., at least 5 out of 7 correct), and their data were excluded from further analyses. The rest of the participants were very accurate in recognizing previously studied items (over 87% of correct recognitions versus 86% in Experiment 1), and in

rejecting items from novel categories (95% and 99% of correct rejections respectively). Levels of false recognitions were also comparable to the baseline in Experiment 1 (38% and 33% respectively). Results of Experiment 3 suggest that differential accuracy on a recognition memory test in Experiment 1 and 2 cannot by attributed to the difference in task demands or difference in instruction, since memory performance in Experiments 3 was very close to performance in Experiment 1, despite a task added to the study phase, and a lack of warning about a subsequent memory test. These findings indicate that the level of processing required by a task (item-specific versus category-specific) influences the level of false recognition on a recognition memory test.

General Discussion

The present study replicated earlier findings that DRM-type intrusions are possible with pictorially presented stimuli, and generalized these earlier findings to multiple categories. The study also demonstrated that processing manipulations (rather than differential task demands) influence the levels of false recognition. Specifically, tasks that focus participants on category-level properties result in category-level representations, and lead to decreased memory accuracy compared to the tasks that focus participants on the item-specific properties of stimuli. This decrease, however, is not proportional to the level of memory performance in the Baseline condition: as a result of performing induction, memory accuracy decreases drastically, and becomes more comparable for all types of Targets.

The reported findings, indicating that DRM-type memory intrusions persist even with pictures, seem to weaken the source monitoring explanation of memory intrusions. Even if source-monitoring errors play a role in memory intrusions with word-lists, these errors are highly unlikely to generate recognition errors when stimuli are presented pictorially. Therefore, assuming that the same mechanism underlies DRM-type intrusion with verbally and pictorially presented materials, it seems reasonable to conclude that monitoring errors are unlikely to be the only source of DRM-type intrusions with verbally presented materials. Given that memory intrusions with pictures are isomorphic to those with word-lists (i.e., both modes of presentation elicit high levels of false recognition and different intrusion rates across different lists), the assumption does not seem unreasonable. Therefore, the reported results seem to support the idea that that DRM-type intrusions stem from category-level or "gist" representations rather than from source monitoring errors.

The finding that performance on an induction task results in an increased level of memory intrusions is theoretically important for the study of inductive reasoning as well as memory. In particular, researchers debate whether or not induction is category-based at different points of development (see Sloutsky, 2003), and the study of effects of induction on memory accuracy may bring critical evidence to this debate.

In short, the reported research brings new evidence to research on memory and induction: category-based induction results in the formation of category-level or "gist" representations, which in turn increase false recognition of new items from studied categories.

Acknowledgments

This research is supported by a grant from the National Science Foundation (REC # 0208103) to Vladimir M. Sloutsky.

References

Alba, J. & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93*, 203-231.

Arndt, J. & Hirshman, E. (1988). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language, 39,* 371 – 391.

Brainerd, C., Reyna, V., & Mojardin, A. (1999). Conjoint recognition. *Psychological Review*, *106*, 160-179.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17 – 22.

Gallo, D. & Roediger, H. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language, 47,* 469 – 497.

Hintzman, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411 – 428.

Koutstaal, W., & Schacter, D. (1997). Gist-based false recognition of pictures in older and younger adults *Journal of Memory & Language*, *37*, 555-583.

Roediger, H., & McDermott, K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 803 – 814.

Roediger, H., Watson, J., McDermott, K., & Gallo, D. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*, 385-407.

Seamon**,** J. G. Luo, C. R., Schlegel, S. E., Greene, S. E., & Goldenberg, A. B. (2000). False memory for categorized pictures and words: The category associates procedure for studying memory errors in children and adults. *Journal of Memory and Language,* 42, 120 – 146.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*, 246-251.

Sloutsky, V. M., & Fisher, A. V. (In press). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science*.

Sulin, R., & Dooling D. (1974). Intrusion of a thematic idea in retention of prose. *Journal of Experimental Psychology,* 103, 255- 262.

The Development of Induction: From Similarity-Based to Category-Based

Anna V. Fisher (fisher.449@osu.edu)
Department of Psychology & Center for Cognitive Science
The Ohio State University
208B Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210 USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210 USA

Abstract

The ability to perform inductive generalizations has been demonstrated to develop very early in life. We argue that while adults use their conceptual knowledge when performing induction, young children perform induction by computing the similarity among presented entities. We further argue that this differential processing underlying children's and adults' induction results in different memory traces, and affects accuracy on a subsequent memory test. Experiment 1 demonstrates that while performing induction decreases memory accuracy of adults and 12-year-olds, it does not affect memory accuracy of 5-, and 7-year-olds. In Experiment 2, 5- and 7-year-olds were trained to perform category-based induction, which resulted in a decrease of their memory accuracy. In Experiment 3, a delayed transfer task was used to examine whether 5- and 7-year-olds could retain their learning over time. Overall, results of the reported experiments point to a developmental trend from similarity-based to category-based induction.

Introduction

The ability to make inductive generalizations is undoubtedly crucial for humans, for not only does it facilitate acquisition of new knowledge and skills, but also aids survival: "our knowledge that leopards can be dangerous leads us to keep a safe distance from jaguars" (Sloman, 1993, p.321).

It has been demonstrated that infants and very young children can perform simple induction tasks (Gelman & Markman, 1986; Sloutsky, Lo, & Fisher, 2001, Welder & Graham, 2001). The process underlying this basic ability is, however, still open to debate.

According to one view, children's inductive generalizations are driven by *a priori* conceptual assumptions (Keil, Smith, Simons, & Levin, 1998, Gelman & Hirschfeld, 1999). Under this view, which has traditionally been referred to as a *naïve theory* position, even early in development, induction is driven by the category assumption – a belief that entities are members of categories, and members of the same category have much in common. Thus, in the course of induction, children first identify presented entities as members of categories, and then perform inductive inferences on the basis of this categorization, because they presumably believe that members of the same categories share many unobservable properties.

According to another position, young children perform induction (as well as categorization) by detecting multiple correspondences, or similarities, among presented entities (e.g., see Jones & Smith, 2002; McClelland & Rogers, 2003; Sloutsky & Fisher, in press-a; Sloutsky, 2003). Because members of a category often happen to be more similar to each other than they are to nonmembers, young children are more likely to induce unobserved properties to members of the category. One such similarity-based model, SINC (abbreviated for Similarity, Induction, and Categorization) was proposed recently by Sloutsky and colleagues (Sloutsky et al., 2001; Sloutsky, 2003, Sloutsky & Fisher, in press-a). Under this view, conceptual knowledge (i.e., knowledge that that members of the same category share many unobservable properties) is a product of learning and development rather than an *a priori* assumption.

In short, under the former view, induction is category-based (i.e., it is a product of categorization), whereas under the latter view, induction is a product of computation of similarity. One of the goals of this research is to distinguish between these positions.

Traditionally, inductive inference in children has been studied directly, by asking participants to perform inductive generalizations and assessing their performance. However, this approach may not be an optimal way of examining representations underlying performance on induction tasks. An alternative framework has been recently suggested (Sloutsky & Fisher, in press-b). In this framework, representations underlying induction performance are studied by examining memory traces formed in the course of Induction. Participants are first presented with sets of pictures of familiar animals, and are asked to make inductive inferences about these animals. Later participants are given a surprise recognition memory test, in which they are presented with some old pictures (i.e., pictures they had previously reasoned about in the induction task), and some Critical Lures (i.e., "new" pictures that belong to the same category as "old" pictures). If participants perform induction in a similarity-based manner, they should form

item-specific representations, and exhibit high accuracy on a recognition test. At the same time, if participants form category-level representations (which might be the case if induction is category-based), they should poorly distinguish between Old Targets and Critical Lures.

Results reported by Sloutsky and Fisher (in press-b) indicate that young children exhibit high recognition accuracy for Critical Lures (thus pointing to similarity-based induction), whereas adults exhibit low recognition accuracy for Critical Lures (thus suggesting category-based induction). It has also been demonstrated that adults' category-based induction results in a decrease in memory accuracy compared to the Baseline memory tasks, while young children's similarity-based induction does not.

The goal of the series of experiments presented below is to compare the two theoretical positions by examining the pattern of development of inductive inference. The similarity-based position assumes a gradual transition from similarity-based to category-based induction in the course of learning and development, whereas no such transition is predicted by the naïve theory position. According to this position, even young children perform category-based induction. Another goal is to provide a learning account of the transition from similarity-based to category-based induction.

Experiment 1

Method

Participants Participants were 45 5 year-olds (19 girls, 26 boys, *M* age= 5.2 years, *SD* = .32 years), 35 7 year-olds (21 girls, 14 boys, *M* age= 7.9 years, *SD* = .54 years), 39 12 year-olds, (18 girls, 21 boys, *M* age = 12.1 years, *SD* = .48), and 30 introductory psychology students at a large Midwestern university (12 women and 18 men, *M* age= 19.5 years, *SD* = .99 years).

Materials, Design and Procedure Materials were 44 color photographs of animals presented against the white background. All animals were highly familiar to both children and adults, with familiarity established in a separate experiment (Sloutsky & Fisher, in press-b). Examples of the photographs used are presented in Figure 1. During the study phase, participants were presented with 30 pictures, one picture at a time, from three different categories (10 cats, 10 bears, and 10 birds). During the recognition phase, they were presented with 28 pictures, one picture at a time, and were asked whether they had seen each picture during the study phase. Half of the recognition pictures were previously presented during the study phase, and the other half were new pictures. The recognition pictures represented animals from three different categories: cats (7 of which were old and 7 were new), bears (all 7 of which were old), and squirrels (all 7 of which were new).

The experiment included two between-subject conditions: Baseline and Induction. The recognition phase was identical in both conditions, whereas the study phase differed across conditions.

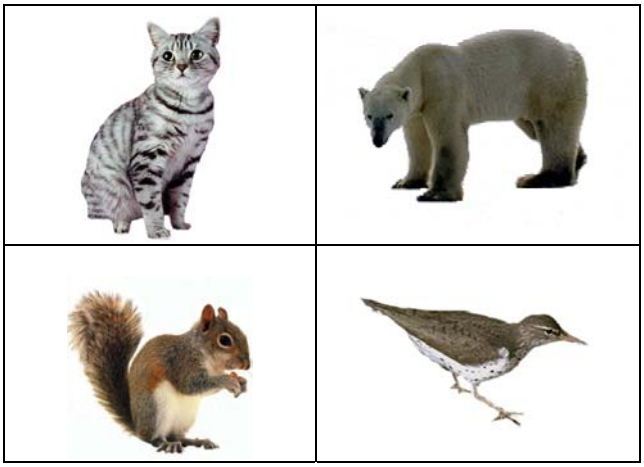


Figure 1: Examples of stimuli in Experiment 1.

In the study phase of the Baseline condition participants were presented with 30 pictures of animals, and their task was to remember these pictures for a subsequent recognition test. In the study phase of the Induction condition participants were first presented with a picture of a cat, and informed that it had "beta-cells inside its body". Participants were then presented with 30 pictures of animals (identical to those presented in the Baseline condition), and asked whether each of the animals also had beta-cells inside. After responding, participants were provided with "yes/no" feedback, indicating that only cats, but not bears or birds, had beta-cells. The recognition test was not mentioned in the study phase of this condition.

During the recognition phase, which immediately followed the study phase, participants were presented with 28 pictures and were asked to determine whether each was "old" (i.e., exactly the one presented during the study phase) or "new." No feedback was provided during the recognition phase

Children were tested individually in their day care centers by female hypothesis-blind experimenters. Undergraduate students were tested individually in a laboratory on campus. For all participants, stimuli were presented on a computer screen, and stimuli presentation was controlled by Super Lab Pro 2 software (Cedrus Corporation, 1999).

Results and Discussion

Although participants in every age group were very accurate in the study phase of the Induction condition, adults and 12 year-olds were somewhat more accurate (averaging 91% and 94% of correct inductions respectively) than 5- and 7-year-olds (74% and 84% of correct inductions respectively), $F(3, 78) = 5.9$, $p < .01$, post-hoc Tukey test, ps $< .05$ for all differences.

In the recognition phase of the experiment all participants were highly accurate in rejecting non-target distracters (i.e., squirrels), averaging over 91% of correct responses across conditions.

However, participants exhibited differential accuracy for the Targets (items previously presented during the study phase, i.e. old cats and bears) and Critical Lures (new items from the same category as the Targets, i.e., new cats) in the Induction and the Baseline. To examine the ability of participants to discriminate previously presented Targets from Critical Lures, memory sensitivity A-prime scores were computed. A-prime is a non-parametric analogue of the signal-detection statistics d-prime (Snodgrass & Corwin, 1988). If participants do not discriminate old Targets from Critical Lures, A-prime is at or below 0.5. The greater the discrimination accuracy, the closer A-prime scores are to 1. Proportions of hits (i.e., correct recognitions), false alarms on Critical Lures (FA), and A-prime scores by age group and condition are presented in Table 1.

Data in the table indicate that 5-, 7- and 12-year-olds well discriminated old items from Critical Lures in the Induction as well as the Baseline condition (A-primes > 0.5, one-sample *t*s > 2.8, *p*s < .01). At the same time, adults were accurate in the Baseline condition (A-primes > .5, one-sample t (14) 16.1, $p < .001$), whereas they were not accurate in the Induction condition: unlike children, adults' A-primes in this condition were not different from 0.5, one-sample $t < 1$, indicating no discrimination between old items and Critical Lures. Furthermore, adults' accuracy was lower than that of 5-year-olds or 7-year-olds, both independent sample *t*s > 2, *p*s <.05.

Table 1: Proportions of Hits, False Alarms (FA) and A-prime scores by age group and condition.

Age group	Baseline			Induction		
	Hits	FA	A-prime	Hits	FA	A-prime
5 year-olds	.82	.59	.66	.71	.56	.69
7 year-olds	.75	.40	.72	.77	.45	.74
12 year-olds	.79	.39	.78	.79	.59	.63
Adults	.88	.40	.84	.81	.74	.54

These findings are summarized in Figure 2, which presents a change in the A-prime scores in the Induction condition compared to the Baseline. Data in the figure indicate that in the Induction condition recognition memory was somewhat reduced in 12 year-olds and dramatically attenuated in adults, while Induction had virtually no effect on the recognition accuracy of 5- and 7-year-olds. The significant age by condition interaction was confirmed by the two-way (age by experimental condition) ANOVA performed on the A-prime scores, $F(3, 141) = 5.7$, $p < .001$.

We argue that high recognition accuracy of younger participants of Experiment 1 was due to the fact that they were engaged in item-specific processing regardless of the experimental condition. Adult participants, on the contrary, demonstrated high memory accuracy only in the task that forced them to perform item-based processing, the Baseline condition. In the Induction condition, however, adults demonstrated low memory accuracy, due to engagement in category-level processing. Results of Experiment 1 therefore point to a developmental trend from similarity-based to category-based induction: memory sensitivity of younger children does not decrease at all in the Induction compared to the Baseline, while sensitivity of 12 year-olds decreases somewhat, and sensitivity of adults reduces dramatically.

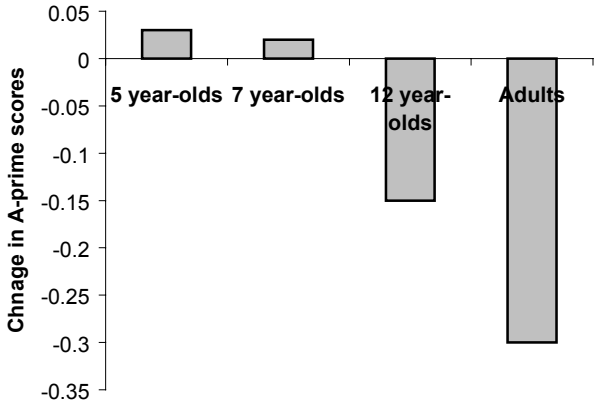


Figure 2: Change in the A-prime scores in the Induction condition compared to the Baseline across age groups.

Experiment 2 was designed to provide a learning account of the category-based induction found in adults, by training 5- and 7-year-olds to perform induction in the category-based manner. If training is successful, that is if memory accuracy of younger children can be reduced to the level of adults, this would further undermine the claim that reasoning in young children is *a priori* conceptually constrained.

Experiment 2

Method

Participants Participants were 27 5 year-olds (16 girls, 11 boys, *M* age= 5.2 years, *SD* = .26 years), and 15 7 year-olds (11 girls, 4 boys, *M* age= 7.6 years, *SD* = .44 years).

Materials, Design and Procedure Materials were identical to those of Experiment 1, however, participants were tested in the Induction condition only. The procedure of Experiment 2 was different from Experiment 1 in that prior to the study phase, participants were trained to perform category-based induction. Children were taught that animals that have the same names belong to the same category, and

that animals that belong to the same category share unobservable properties. Picture cards representing rabbits, dogs, and lions were used for the training procedure; none of these categories of animals were used in the experiment proper.

Upon completing the training, participants were presented with the experimental task, which was identical to the Induction condition of Experiment 1. Hypothesis-blind female experimenters tested children individually in their schools and child care centers.

Results and Discussion

Overall participants were highly accurate during the study phase of Experiment 2, averaging over 92% of correct inductions. Similar to Experiment 1, participants were also very accurate in rejecting non-target distracters (i.e. squirrels), giving on average 98% of correct responses.

However, unlike Experiment 1, memory sensitivity of participants in both age groups, as indicated by the A-primes scores, did not differ from chance, both one-sample *ts* < 1.7, *ps* > .1. Proportions of hits, false alarms on Critical Lures, and A-prime scores are presented in Table 2.

Results of Experiment 2 indicate that training to perform category-based induction significantly reduced memory accuracy of both 5- and 7-year-olds, bringing their recognition performance to chance and making it comparable to the performance of adult participants in Experiment 1.

Could it be that training had a non-specific effect on memory accuracy, such that, regardless of the experimental task, children participating in the training experiment exhibited reduced memory accuracy? This issue was addressed by Sloutsky and Fisher (in press-b), who demonstrated that training had no adverse effects on children's memory in the Baseline condition.

Table 2: Proportions of Hits, False Alarms (FA) and A-prime scores by age group in Experiment 2.

Age group	Training Condition		
	Hits	FA	A-prime
5 year-olds	.82	.65	.58
7 year-olds	.73	.59	.57

While demonstrating that both 5- and 7-year-olds were successful in learning to perform category-based induction, the experiments left an important question unanswered. In particular, it remained unclear whether effects of this training would be retained over time. Results of Experiment 2 together with the developmental trend found in Experiment 1 suggest that older children should be better able to retain what they learned during training over a time delay. Experiment 3 was designed to investigate the ability of 5- and 7-year-olds to retain this new knowledge over a time delay.

Experiment 3

Method

Participants Participants were 17 5 year-olds (11 girls, 7 boys, *M* age= 5.3 years, *SD* = .17 years), and 19 7 year-olds (4 girls, 15 boys, *M* age= 7.6 years, *SD* = .44 years).

Materials, Design and Procedure Materials and procedure were identical to those of Experiment 2 with one important difference: there was a delay between training to perform category-based induction and the experiment proper. The delay was on average 14.6 days (SD = 1.5 days, range 14 – 18 days). As in the previous experiments hypothesis-blind female experimenters tested children individually in their schools and child care centers.

Results and Discussion

Similar to previous experiments participants were highly accurate both in rejecting non-target distracters (averaging over 96% of correct rejections), and making correct inductions during the study phase (84% and 86 % of correct inductions in the groups of 5- and 7-year-olds respectively). However, in contrast to Experiment 2, participants demonstrated differential memory accuracy for Critical Lures. Proportions of hits, false alarms on Critical Lures, and A-prime scores are presented in Table 3. Memory accuracy of 7 year-olds indexed by the A-prime scores was close to chance (which was similar to their accuracy in Experiment 2), one-sample t (18) = 1.9, $p > .07$. On the other hand, recognition memory of 5 year-olds was clearly above chance, one-sample t (16) = 4.9, $p < .0001$.

Table 3: Proportions of Hits, False Alarms (FA) and A-prime scores by age group in Experiment 3.

Age group	Delayed Transfer Condition		
	Hits	FA	A-prime
5 year-olds	.82	.59	.67
7 year-olds	.91	.75	.59

Thus, results of Experiment 3 indicate that while 7 year-olds retained what they had learned during training over a two-week delay, 5 year-olds were unable to do so. Therefore, retaining of the learned ability to perform induction in a category-based manner seems to be a function of age.

Memory accuracy of 5- and 7-year-olds across three reported experiments is presented in Figure 3. Results presented in Figure 3 point to an interesting developmental pattern: while both 5- and 7-year-olds do not perform

category-based induction spontaneously (i.e., under the no-training condition of Experiment 1, their accuracy is high), children in both age groups can be successfully trained to perform category-based induction (as evidenced by their reduced accuracy in the training condition of Experiment 2). However, only 7 year-olds are able to retain the results of training over longer periods of time (i.e., after the delayed condition in Experiment 3, their accuracy remained low). At the same time, 5-year-olds reverted back to similarity-based induction (i.e., after the delayed condition in Experiment 3, their memory accuracy returned to the high pre-training level).

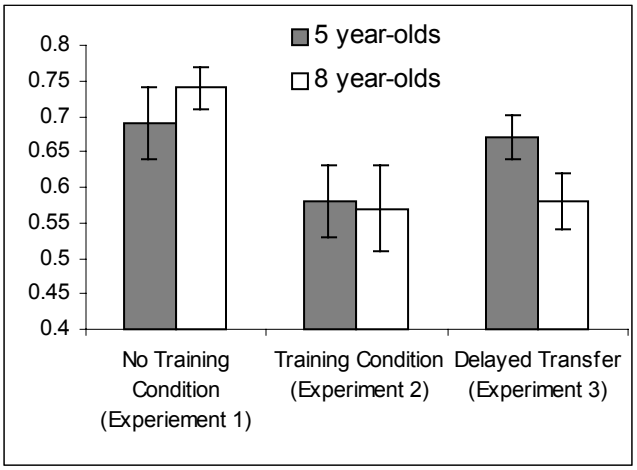


Figure 3: A-prime scores for 5- and 7-year-olds in the Induction task across Experiments 1 – 3.

General Discussion

Several important findings stem from the three reported experiments. First, there is a clear trend in the development of induction: induction task attenuates memory accuracy for individual items of 12-year-olds and adults, whereas 5- and 7-year-olds exhibit accurate memory for individual items. Furthermore, in the Induction (but not in the Baseline) condition younger participants exhibit greater memory accuracy than older participants or adults. Second, training to perform category-based induction leads to a decrease in memory accuracy of 5- and 7-year-olds to the level of adults. And third, 7-year-olds retain training over longer periods of time than 5-year-olds: 5-year-olds sooner than 7-year-olds exhibit high levels of memory accuracy for individual items, returning to their pre-training high accuracy.

These results indicate that: (1) while 12-year-olds and adults perform category-based induction (which results in mostly category-level representations), 5- and 7-year-olds perform similarity-based induction (which results in item-level representations); (2) there is a gradual developmental transition from similarity-based to category-based induction; and (3) category-based induction does not have to be *a priori*, it can be learned and retained over time (although the length of retention is a function of age). These results support predictions of the similarity-based account of induction, while presenting challenges to the naïve theory approach. In what follows, we consider theoretical implication of these results.

Induction and Memory Accuracy Across Development

The results support the contention of the similarity-based approach that early in development children perform induction by computing similarity among compared entities. As a result, these participants form item-specific representations, and accurately remember individual items encountered in the course of induction. At the same time, older children and adults perform category-based induction (i.e., they first categorize entities, and then generalize properties to members of the same category), and as a result they form category-level, but not item-specific representation, thus exhibiting poor memory for individual items encountered in the course of induction. Note that older children and adults have no difficulty remembering individual items, in the Baseline condition, in which they are not required to perform induction. Furthermore, there is additional evidence that category-based induction affects memory accuracy for individual items: when younger participants were trained to perform category-based induction, their memory accuracy in the Induction (but not in the Baseline) condition dropped to the level of adults.

Taken together, these findings do not support the contention of the naïve theory position that induction in young children is category-based, but they rather support the contention of SINC that induction in young children is similarity-based.

The Development and Learning of Category-Based Induction

The reported results also present developmental and learning accounts of category-based induction. First, category-based induction gradually emerges in the course of development: there is little evidence that 5- or 7-year-olds spontaneously perform category-based induction, whereas 12-year-olds are more likely to perform it than younger children, and adults are more likely to perform it than 12-year-olds. Second, people do not need *a priori* category assumption – young children can be trained to perform category-based induction. However, the retention of this training is a function of age – 7-year-olds are more likely to retain training over time than younger children. Therefore, it seems reasonable to conclude that (a) there is a transition from similarity-based to category-based induction, and this transition is gradual; (b) category-based induction can be successfully learned; and (c) the retention of learning is a function of age. These findings provide a learning account of category-based induction suggesting that it is unnecessary to posit that conceptual knowledge is *a priori*. Recall that in Experiment 2, participants were taught that (a)

similar things that have the same name belong to the same kind, (b) things that belong to the same kind share many non-observable properties, and (c) things that have the same name share many non-observable properties. It is possible that (a) and (b) are taught in school, whereas (c) is a direct consequence of (a) and (b). Therefore, results of Experiment 2 may explain the transition from the similarity-based induction exhibited by children to category-based induction exhibited by adults, suggesting that category-based induction and requisite conceptual knowledge could be a product of feedback-based learning. While presenting a learning account of category-based induction, these findings seriously challenge the contention of the naïve theory position that category-based induction has to be based on *a priori* assumptions.

Conclusion

Overall, results of the three reported experiments represent novel findings indicating that (a) early in development people spontaneously perform similarity-based rather than category-based induction; (b) there is a gradual transition from similarity-based to category-based induction; and (c) category-based induction is a product of learning. These results support the similarity-based account of young children's induction, while presenting challenges to the naïve theory approach.

Acknowledgments

This research is supported by grants from the National Science Foundation (BCS # 0078945 and REC # 0208103) to Vladimir M. Sloutsky.

References

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.

Gelman, S. A. & Hirschfeld, L. A. (1999). How biological is essentialism? In S. Atran & D. Medin (Eds.). *Folkbiology*, Cambridge, MA: MIT Press.

Jones, S. S., & Smith, L. B. (2002). How children know the relevant properties for generalizing object names. *Developmental Science*, *5*, 219-232.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, *65*, 103-135.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*, 310-322.

Sloman, S. A. (1993). Feature-based induction. *Cognitive psychology, 25,* 231 – 280.

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. *Child Development*, *72*, 1695-1709.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*, 246-251.

Sloutsky, V. M., & Fisher, A. V. (In press-a). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General.*

Sloutsky, V. M., & Fisher, A. V. (In press-b). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science.*

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34-50.

Welder, A. N., & Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, 72, 1653-1673.

Sensitivity to Confounding in Causal Inference: From Childhood to Adulthood

E. Christina Ford (christis@ucla.edu)
Department of Psychology, Box 951563
Los Angeles, CA 90095-1563 USA

Patricia W. Cheng (cheng@psych.ucla.edu)
Department of Psychology, Box 951563
Los Angeles, CA 90095-1563 USA

Abstract

A necessary condition for correctly assessing causality is the absence of confounding causes. This paper reports a pair of experiments that investigate whether people are sensitive to confounding when they infer causation. Two stories were constructed, one in which two candidate causes perfectly covaried with each other (*confounded*), and another in which the two candidate causes occurred independently of each other (*unconfounded*). In the confounded story, both causes covaried perfectly with an outcome; in the unconfounded story, only one of the two candidates covaried with the outcome. If people control for alternative causes while they evaluate a candidate cause, then subjects in the confounded condition should indicate that it is impossible to determine causality for either candidate alone, whereas those in the unconfounded condition should be able to judge that one of the candidates is causal and the other not. If people are not sensitive to confounding, however, subjects in the confounded condition should attribute causality to both candidates, and their judgments for these candidates should be the same as those for the target causal candidate in the unconfounded condition. Two experiments were conducted respectively with children and adults: Children received one or the other story, while adults received both. Both children and adults distinguished between confounded and unconfounded candidate causes when making attributions of causality. Our results show that children are able to state the indeterminacy of confounded candidate causes at an age much earlier than previously documented.

Introduction

One view of how children learn is that they approach the world as scientists and form theories about the world using information about variation and covariation to establish causal connections (e.g. Gelman, 1996; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Gopnik, Sobel, & Schulz, 2001). Further, they intervene upon the world in order to discover these relationships (Schulz, 2003). Although children may have misconceptions in their explanations, as when a child states that he thinks God made the sun out of gold and lit it with fire (Siegler, 1998), the presence of such misconceptions does not mean that children are unable to use the data present in the environment to form correct causal attributions. Given that adults have had many more experiences than children, we should not expect children's theories to be the same as adult's theories, especially for complex phenomena. What is important is whether the same process is utilized when determining causality. In particular, this paper seeks to examine whether both children and adults are sensitive to confounding when there are two candidate causes for a novel outcome.

In addition to the potential implications for improving science instruction, assessing children's sensitivity to confounding is also important for differentiating between two types of models of causal learning. The first type is instantiated in the *unconditional* ΔP model (Jenkins & Ward, 1965); the second type consists of models that tease apart the influence of a candidate cause from the influences of alternative causes (e.g., Cheng, 1997; Cheng & Novick, 1992; Glymour, 2001; Gopnik et al., 2004; Novick & Cheng, 2004; Pearl, 2000; Spirtes, Glymour, & Scheines, 2000; Tenenbaum & Griffiths, 2001).

Under the unconditional ΔP model, people contrast the frequency of e, an effect of interest, when c, a potential cause, is present, with the frequency of e when c is absent:

$$\Delta P = P(e|c) - P(e|\sim c)$$

If ΔP is equal or close to 0, then c is considered noncausal; if it is noticeably greater than 0, then c is thought to cause e, and if it is noticeably less than 0, c is thought to prevent e. The unconditional ΔP model implies that people ignore confounding and pool over all the information known about the candidate cause. Thus, if two candidate causes perfectly covary with each other and the effect, then both candidates will be judged as causal.

Under the alternative approach, a definite causal judgment can result from the above contrast only when alternative causes are controlled (i.e., they occur independently of the candidate cause). One simple variant of this approach is the *conditional* ΔP model: the same ΔP formula is applied to a focal set of events in which alternative causes occur independently of the candidate (Cheng & Novick, 1992). For example, if there is a situation in which there are two possible causes of an event, one way in which a person could determine the causality of the individual candidates would be to compare the frequency of e in the situation in which only one candidate is present to a situation in which no candidate is present, holding the other candidate constantly absent. If people utilize conditional ΔP, they would be unable to draw a definite causal conclusion when

there is confounding because no focal set of events could be formed.

Although unconditional ΔP has fallen into disfavor in the adult causal learning literature, previous studies of children suggest that they do not withhold judgments of causality in the presence of confounded variables (e.g., Kuhn, Amsel, & O'Loughlin, 1988). In other words, they seem to behave as predicted by the unconditional ΔP model. Our study focusses on young children. If the ability to reason causally is an unlearned fundamental human process, then it should be present at an age much earlier than indicated by prior research.

One study looking at third, sixth and ninth graders, as well as non-college young adults and undergraduate college students found that before the ninth grade, students were unlikely to state that there was insufficient evidence to determine causality when there is confounding (Kuhn, Amsel, & O'Loughlin, 1988). In fact, not a single subject suggested indeterminacy as the correct answer until the 9th grade. In one condition, 20 subjects in each age group was asked to determine the whether a feature of a ball (namely, texture) caused a ball to be bouncier in the presence of a perfectly confounded covariate (namely, color). None of the 3rd or 6th graders, one 9th grader, 2 non-college adults, and 5 college subjects proposed indeterminacy as the correct answer. But, these experiments involved causes for which the students were likely to have prior theories, and people interpret ambiguous data in ways that are consistent with their prior beliefs (Darley & Gross, 1983). Kuhn et al. (1988) do not indicate whether students who did not notice the indeterminacy were answering in a manner consistent with their prior theory. Also, because their studies focus on the coordination of theory and evidence, one of criteria used for assessing students' answers was their ability to justify their responses. But, if causal learning is an unconscious process, students might be sensitive to confounding, yet unable to justify their responses. In the present study, the task is made simpler, by presenting subjects with a novel effect, thereby reducing the relevance of prior causal beliefs, and by measuring subjects' causal attribution without asking for a justification.

Data from two pilot experiments are presented. Both experiments test whether people differentiate between confounded and unconfounded candidate causes. In one experiment, the subjects were undergraduates, while in the other the subjects were pre-school age children. In both experiments, participants were presented with two possible causes for a novel event, and were asked to determine the cause of that event. In one condition the two possible causes were independently occurring, while in the other condition the two candidate causes always occurred together. If people are sensitive to confounding they should be able to make a causal attribution in the first condition but not the second.

Methods

These experiments were designed to test whether people are sensitive to the independent occurrence of potential causes of an effect when making judgments of causality. The first experiment was conducted on adults. Even if adults are able to succeed in this task, however, their success might well be due to prior training. The second experiment was therefore conducted on children. Similar materials were used for both experiments.

The second experiment was actually conducted first. Because the data with the children was not very clean, in order to develop a better protocol the materials were piloted with adults. The small sample sizes in both experiments are due to the preliminary nature of the data. Also, the adult pilot was ended when the adult answers became consistent. Both experiments will be re-run with larger sample sizes using the final adult version of the stimuli**.** Below we describe the methods for both experiments before reporting the results.

Experiment 1

Participants 10 undergraduates at the University of California, Los Angles enrolled in an Introduction to Psychology Course participated in the study. Students received class credit for participating in the study and were recruited using an on-line bulletin board for this course.

Design This experiment had two conditions and utilized a within-subjects design. In one condition, the two possible causes of an unusual event were perfectly correlated (confounded). In the other condition, the same two possible causes occurred independently of one another (unconfounded). Subjects were asked about the causality of the candidate causes in turn. The ordering of the stories, as well as the order in which the subject was asked about each candidate cause, was counterbalanced across subjects.

Materials Two passages of approximately the same length were constructed (one story was 668 words and the other was 681 words). Both passages tell the story of bunny rabbits that went to two different parties.

In both stories, the parties occur at the same time and on the same day. On the day of the party, the bunnies are randomly assigned to a party via a coin toss. Half of the bunnies ate candy before going to the party. At one of the parties the bunnies ate cake, while at the other party they did not. In the confounded condition all the bunnies who ate candy also ate cake, whereas in the unconfounded condition half of the bunnies who ate candy also ate cake, and vice versa. All the bunnies at the cake party grew new pink wings; none of the bunnies at the "no cake" party did. To avoid confusion between the two stories, in one story, the bunnies ate green grass candy and yellow cheesecake; in the other story the bunnies ate blue berry candy and orangey orange cake.

At the end of the story, participants were asked about the causality of each of the causal candidates in the story:

1) Does Yellow Cheese Cake/ Blue Berry Cake all by itself make bunnies grow new pink wings? Yes, No, or Impossible to tell?
2) Does Green Grass Candy/ Orangey Orange Candy all by itself make bunnies grow new pink wings? Yes, No, or Impossible to tell?

The text of the story was accompanied by illustrations. An appropriately colored wedge in the bunnies' stomachs represented the cake, and a candy shaped object in the bunnies' stomach represented the candy.

Because we were attempting to revise the stimuli in order to make the directions clearer for the children, the stimuli underwent slight modification across the 10 subjects. The conditions remained the same, but there were slight changes in wording and pictorial presentation across groups. The most significant wording change was that the story narrative was condensed, leaving only a concise explanation of the meaning of the symbols that represented the outcomes as well as the candidate causes. The most significant pictorial modifications occurred in the confounded condition. In the original stimuli, the bunnies were in two groups (the cake group and the no cake group), with the bunnies who ate candy evenly distributed throughout each group. In the final stimuli, the bunnies were arranged into four groups of bunnies that underwent each treatment (i.e., one group had cake and candy, one group had only candy, one group had only cake, and one group had neither cake nor candy). Both of these changes served to make the experiment easier for the subjects to understand and interpret.

The stories were shown as a power point presentation. The power point presentation was presented on a 15" computer screen.

Procedure Participants were randomly assigned to conditions that differed on the ordering of the stories and assessment questions. Participants were then told that they were going to hear a story about bunny rabbits in two little bunny towns. They were told that something interesting was going to happen to these bunny rabbits, and that it was their job to try to figure out what happened.

Participants looked at the illustrations on the screen as the experimenter read the story aloud. At the end of the story, participants were asked about the causality of each of the causal candidates in the story. The experimenter wrote down their answers on an answer sheet as they progressed through the story.

Experiment 2

Participants Sixteen pre-school children from the Bellagio daycare center at the University of California, Los Angeles participated in the study. Nine male and seven female children between the ages of 4;5 and 5;7, with a mean age of 4;11 participated in the study. One child was excluded from the analysis for answering incorrectly factual questions about the stories presented. The rest of the children answered all of these questions correctly (as explained later).

Design This experiment had the same two conditions as Experiment 1 but utilized a between-subjects design. The order in which children were asked about each candidate cause was counterbalanced across conditions.

Materials The stories presented to the children had the same content as the stories presented to the adults, with three differences. First, the children's protocols did not undergo significant changes. The initial adult protocols (with the distributed confounding variable and long narrative) are the same as the child protocol. Second, in both conditions, children saw green grass candy and yellow cheesecake. (This was possible because subjects only saw one story, which ruled out the possibility of carryover between stories.) The children's assessment procedure also differed from that of the adults.

Children were first asked for their spontaneous attribution. "Do you think that it is possible to figure out why the bunnies grew new pink wings?" If the child answered yes then the following questions were asked:

1) Why do you think these bunnies [pointing to those who went to the cake party] grew new pink wings?
2) Why do you think these bunnies [pointing to those who went to the no cake party] did not grow new pink wings? The ordering of these two questions was counterbalanced across conditions.

Because children sometimes did not give a free response, did not address both of the causal candidates, or did not address the causal candidates in their responses (e.g., "Bunnies grew wings because they wanted to"), additional probes were added, asking about each of the candidate causes separately. Children were told about statements that other children had made while reading this story. Children were asked whether they thought these statements were "definitely right, definitely wrong, or impossible to tell." The statements they were asked to judge were

1) GREEN GRASS candy all by itself makes bunnies grow pink wings.
2) YELLOW CHEESE CAKE all by itself makes bunnies grow pink wings.
3) YELLOW CHEESE CAKE and GREEN GRASS candy together make bunnies grow pink wings.

If the child had previously indicated that the yellow cheesecake was causal, they were not asked about the yellow cheesecake again (and the same for the other candidates).

Procedure Children were randomly assigned to the conditions. They were video taped during the session. In order to accustom children to the camera, they were first introduced to the camera and allowed to see themselves on the LCD screen. As with the adults, children were told that they would hear a story about bunny rabbits, and that it was their job to figure out what happened.

The children looked at the illustrations on the screen as the experimenter read the story aloud. At the end of the story, the children were asked four factual questions to assess whether they understood and remembered the content of the story. The experimenter pointed to a picture of the bunnies with the candy in their tummies and asked, "What did these bunnies eat?", the correct answer being "candy" (or cake and candy in the confounded condition). The experimenter then pointed to the bunnies without candy in their tummies and asked, "Did these bunnies eat candy?", the correct answer being "no". The experimenter then pointed to a picture of the bunnies at the cake party and asked, "What did these bunnies eat at the party?" (this question was omitted in the confounded condition if children answered cake and candy to the first question above), the correct answer being "cake". The experimenter then pointed to a picture of the bunnies at the no-cake party and asked, "Did these bunnies eat cake?" the correct answer being "no". Children who did not correctly answer all factual questions were excluded from the study.

Results

Experiment 1

Adult subjects were sensitive to confounding when they made causal judgments. The data are presented in Tables 1 and 2 below, with the number who answered correctly in bold font for each condition.

Table 1: Cake attributions for adults

	Causal Attribution		
Condition	Yes	No	Can't Tell
Confounded	0	0	**10**
Unconfounded	**8**	1	1

Table 2: Candy attributions for adults

	Causal Attribution		
	Yes	No	Can't Tell
Confounded	0	0	**10**
Unconfounded	0	**6**	4

Using McNemar's test for 2-related samples of categorical data, we see that the pattern of responses differed across conditions for both of the causal candidates. Subjects were more likely to say the cake was causal in the unconfounded condition than in the confounded condition, and conversely, more likely to say it was impossible to assess causality in the confounded condition than in the unconfounded condition ($p < 0.05$, exact statistic, binomial distribution used). Likewise, subjects were more likely to say that the candy was not causal in the unconfounded condition than in the confounded condition, and more likely to say it was impossible to tell in the confounded condition than in the unconfounded condition ($p < 0.05$, exact statistic, binomial distribution used). In fact, as can be seen in Table 1, for both cake and candy in the confounded condition, all subjects said that it was impossible to tell if either candidate was causal. In contrast, in the unconfounded condition, most subjects said that the cake was causal, and none said that the candy was causal.

The order of presentation of the conditions had no effect: Subjects were just as likely to give a correct response whether they received the confounded or unconfounded condition first (n.s.). We were unable to test the effect of the ordering of the questions on the two candidate causes because some information on the ordering was lost. There was no effect of candidate in either condition: subjects were just as likely to make the correct causal judgment for the candy as they were for the cake (n.s.).

Experiment 2

Children were also sensitive to confounding when they made causal judgments, but this data show more variability than the adult data.

Children were first categorized into one of five causal attribution categories: the cake is causal, the candy is causal, both causal (jointly or independently), it is impossible to tell, and other causal attribution (Table 3). Because the focus of this paper is on children's ability to determine whether there is sufficient evidence to attribute causality to an individual cause in the case of confounding, the answers were collapsed into two groups, either assigning causality to individual candidates or not assigning causality to individual candidates (Table 4). If children made a spontaneous causal attribution, this was taken as the value for the measure. If a child did not give a spontaneous response or gave an ambiguous answer, the value for this measure was taken from the child's answers to follow-up questions about each individual candidate.

Table 3: Children's causal responses

	Condition	
Child responses	Confounded	Unconfounded
Cake	0	**4**
Candy	0	0
Both	3	2
Can't Tell	**4**	1
Other	0	1

Table 4: Children's attributions to individual candidates

	Condition	
Causal Category	Confounded	Unconfounded
Individual Attribution	2	**7**
No Individual Attribution	**5**	1

In the confounded condition, 3 children said both candidates were causal and 4 said it was impossible to tell.

Children in this condition gave no other responses. All of these statements were spontaneous attributions, and their answers in the follow-up questioning were consistent with their spontaneous answers. If children in the "Both" category had meant that the candy and the cake together made the pink wings, but that it was not possible to judge either candidate by itself, this would have been a correct response. Given these children's pattern of responding, for two of the children it is not clear whether they meant the two were independently or conjunctively causal. To score conservatively, against sensitivity to confounding, we counted these as attributions to the individual candidates. The third child indicated that she thought the conjunctive answer was the best answer despite the fact that she answered yes to the probes asking about the causality of each candidate individually. This child was recoded as "no individual attribution" in Table 4.

In the unconfounded condition, 4 children said cake was causal, 2 children said they both were causal, one child gave an alternate attribution, and another said that it was impossible to establish causality. Five of these statements were spontaneous attributions, only one of which was consistent with follow-up questioning. Two children who were coded as making a causal cake-attribution initially began by giving unclear spontaneous responses.

Children were more consistent in their responses across question formats in the confounded condition than in the unconfounded condition (Fisher's exact test, $p < 0.05$).

Even with our small samples, given the scoring as we just explained, individual attribution (see Table 3) was more likely in the unconfounded condition than in the confounded condition (Fisher's exact test, $p < 0.05$). Moreover, the correct response was the modal response in each condition (Table 4). Even so, given our small sample size, in both the confounded and unconfounded conditions children's ability to make a correct causal attribution was not significantly different from chance (Fisher's exact test, n.s.).

Discussion

Like Kuhn et al. (1988), it is likely that this study underestimates children's causal reasoning abilities. The correctness of the children's causal attributions, when they occurred, where not impressive in this experiment. These results may seem at odds with recent literature suggesting that children are often able to make correct causal attributions at an early age (e.g., Gopnik et al., 2001). It is likely that the poor performance in Experiment 2 is due to an overly confusing experimental protocol, particularly the visual presentation of the stimuli. Because the confounding cause was distributed equally across both groups, it was difficult to visually isolate the correct focal set.

Because the stimuli used were different between the child and adult studies, it is impossible to tell what, if anything, developed from childhood to adulthood. Our results do not speak to whether adults are better at causal reasoning than children. But, it is clear that both adults and children can differentiate situations in which it is possible to make a causal determination from those in which it is impossible to make a causal determination.

One puzzling result is that the adults responded unanimously to the confounded condition but less clearly to the unconfounded condition, and seemed to be more confused about the candy. This difference between conditions, however, is likely due to confusion in the protocol. After the assessment, when subjects were asked to explain their answers, the two subjects in the unconfounded condition who did not answer that the cake was causal cited reasons of experimental control. For example, one subject answered, "The bunnies at one party could have drank something that the bunnies at the other party did not." When it was made clear that the bunnies were exactly the same, except for whether they had eaten cake or candy, both subjects gave the correct response. Similarly, the subjects who said "can't tell" for the candy in the unconfounded condition occurred in the earlier portion of the experiment, before the materials were finalized.

One might think that the unconfounded condition could use a simpler, less confusing, design. But, in order to compare the unconditional ΔP model with the conditional ΔP model regarding confounding, the unconfounded condition must present information on all four possible combinations of the two candidate causes: both candidate causes are present, neither is present, and only one or the other of the candidate causes is present. To form the appropriate focal set for each candidate cause, the subject could compare the outcome in a single candidate "cell" to the outcome in the no-candidate-cause cell. Alternatively, the subject could compare the outcome in the combined candidate cell to outcomes in each of the single candidate cause cells.

Let us consider two potential two-cell designs. In the simpler design, there is a single candidate cause. In one cell both the candidate cause and the effect are present, in the other cell neither the candidate cause nor the effect is present. In this case, if children attribute causality to the candidate cause, it could be due purely to association. Furthermore, if fewer children attribute causality in the confounded condition, it could be argued that that condition is simply more difficult because it involves two candidate causes rather than one.

In a more complex two-cell design, there are two candidate causes: In one cell candidate cause A is paired with the effect, in the other cell candidate cause B occurs without the effect. This situation is, in fact, confounded. There is a perfect (inverse) correlation between the occurrence of candidates A and B. The only appropriate focal set would compare the occurrence of the effect before and after the introduction of the candidate causes, essentially adding a no-candidate-cause cell. Even if temporal information is used, this type of design does not rule out the possibility of a third unobserved candidate systematically affecting both candidates as well as the effect. There would be a number of correct responses depending on the extraneous

assumptions made by the subjects regarding the influence of the unobserved candidate.

Conclusion

Both children and adults distinguish between confounded and unconfounded candidate causes when making causal inferences. All adults said that it was impossible to tell whether the cake or the candy alone caused the wings in the confounded condition. Because the subjects in the adult experiment were UCLA undergraduates and not less well-educated adults, and because the conditions used in this experiment were consistent with examples used in scientific methodology classes, it is possible that the adults have had prior training that allowed them to say that causal attribution was not possible when the causes were confounded. The same was not true of the children.

The child data suggests that children are able to state, at a much earlier age than documented by previous studies, that it is impossible to infer causality when two candidate causes perfectly covary. Children as young as 5 years old, less than half as old as previously believed, made such judgments. Moreover, these judgments were consistent across response formats.

Despite the variability in the data, and despite the fact that the children were presented with a version of the task that was more difficult than necessary, more children indicated that it was impossible make a causal judgment in the confounded condition than in the unconfounded condition. As predicted by the conditional ΔP model, children (and adults) are sensitive to confounding.

Acknowledgments

The preparation of this paper was supported by an NSF fellowship to the first author and an NIH Grant MH64810 to the second author.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365-382.

Darley, J. M. & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20-33.

Gelman, S. A. (1996) Concepts and Theories. In T.K.F Au & R. Gelman (Eds), *Perceptual Development* (pp. 117-150). San Diego, CA: Academic Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical models in psychology*. Cambridge: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111 (1),* 3-32.

Gopnik, A., Sobel, D. M., Schulz, L. E. (2001) Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620-629.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1, Whole No. 594).

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988) *The development of scientific thinking skills*. New York, NY: Academic Press.

Novick, L.R., & Cheng, P.W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*, 455-485.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Schulz, L. E. (2003, April) The play's the thing: Intervention and causal inference. In L. E. Schulz (Chair) *Understanding children's causal knowledge: Exploring the origins of causal inference.* Symposium conducted at the meeting of the Society for Research in Cognitive Development, Tampa, FL.

Siegler, R. S. (1998) C*hildren's Thinking* (3rd ed.). Upper Saddle River, NJ : Prentice Hall.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search* (2nd edition). Boston, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 59-65). Cambridge, MA: MIT Press.

Measuring Card Sort Complexity

T. V. Fossum (fossum@cs.uwp.edu)
Computer Science Department, University of Wisconsin-Parkside
Kenosha, WI 53141 USA

S. M. Haller (haller@cs.uwp.edu)
Computer Science Department, University of Wisconsin-Parkside
Kenosha, WI 53141 USA

Abstract

Card sorts can be used to study the way human subjects organize conceptual knowledge. In this paper we define three measures of complexity of card sorts produced by human subjects. These measures are applied to a particular data set of subjects (students and experts) collected in a large, multi-institutional study where the concepts are taken from a first-year programming course. We show that certain of these measures are statistically significant in discriminating between students and experts and among students based on their performance levels.

Introduction

Card Sort [5] is a technique that seeks to elicit individual conceptual frameworks by giving a subject a collection of cards – each pre-printed, for example, with a word or phrase – and asking the subject to partition (*sort*) the cards into subsets based on the subject's own criteria. The subject is asked to repeat the process anew with different criteria until the subject can think of no additional sorts.

As part of a National Science Foundation (NSF) workshop (Grant DUE-0122560, awarded to the Institute of Technology at the University of Washington–Tacoma), workshop participants conducted a card sort study [4] on "first-competency" programmers: those who have typically completed a second-semester programming course for computer science majors. This study surveyed 243 student subjects and 33 "expert" subjects drawn from 22 institutions represented by the workshop participants. The "experts" were chosen for their proven programming maturity and experience. Both the student and expert subjects were given twenty-six cards used as stimuli. Each card was printed with a one-word programming-related term from the following table:

function	method	procedure
dependency	object	decomposition
abstraction	if-then-else	boolean
scope	list	recursion
choice	state	encapsulation
parameter	variable	constant
type	loop	expression
tree	thread	iteration
array	event	

The data set gathered by the workshop participants includes 1199 card sorts produced by 276 subjects.

A background questionnaire given to each card sort subject provided information on the subject's gender, age, and self-rating of familiarity with programming languages from a specified list (Java, C, C++, Ada, Scheme, Pascal, Visual Basic). At each participating institution, the institution's workshop participant ranked each of the institution's student card sort subjects on a scale from 1 to 5, with 1 representing low-performance and 5 representing high-performance.

The NSF workshop study addressed several questions, including the following:

- Do students and experts organize concepts differently?
- Are there differences between low- and high-performing students? similar to the differences between students and experts?
- Are there differences between male and female students? Between male and female experts?

In this paper, we describe how to compute numerical values that measure the "complexity" of the collection of all the sorts produced by a given card sort subject. These measures are based solely on how the subject partitions the cards into different categories and not on the names the subject gives to the categories nor on the criterion the subject used to sort the cards. We show that, for the NSF workshop study, certain of these measures can be used to distinguish (in a statistically significant sense) experts from novices and high-performing students from low-performing ones. We show that there are no significant differences between males and females.

The Measures

Following [1], we use the "edit distance" metric to define the distance between two card sorts. Specifically, if S_1 and S_2 are two partitions of the set of 26 cards into subsets, we define the *edit distance* $d(S_1, S_2)$ to be the minimum number of edit operations on S_1 (moving a card from one partition subset to another, possibly creating an empty subset and moving a card into it) that will transform it into the same set of subsets of S_2.

As shown in [1], the value of $d(S_1, S_2)$ can be computed using an algorithm for finding the maximum matching weight sum in a bipartite graph [3]: each subset in the partitions S_1 and S_2 is a node in the graph, and an edge between a subset in partition S_1 and a subset in partition S_2 has weight equal to the number of cards in

the intersection of the two subsets. If w is the maximum matching weight sum in this graph, the edit distance $d(S_1, S_2)$ is equal to $26 - w$.

The edit distance function d satisfies the usual properties of a metric except that two sorts S_1 and S_2 can have a zero edit distance without being the "same" sort from the point of view of the subject (the subject may have used different criteria to carry out the two sorts even though the resulting partitions are identical):

$d(S_1, S_2) >= 0$ for all S_1, S_2
$d(S, S) = 0$ for all S
$d(S_1, S_2) = d(S_2, S_1)$ for all S_1, S_2
$d(S_1, S_2) \leq d(S_1, S_3) + d(S_3, S_2)$ for all S_1, S_2, S_3

Definition. Suppose a card sort subject has produced k sorts represented by the sequence $X = (S_1, S_2, \quad S_k)$. Construct the complete graph where each node is a sort from the list X and where the edit distance metric d gives the weight of the edge between any two nodes (sorts) in the graph. The *minimum spanning tree (MST) measure* $\mu(X)$ of X is the sum of the weights of a minimum spanning tree of this graph.

This MST measure is based on all the sorts for a given subject. If a subject produces sorts that are largely similar, the pairwise edit distances will be relatively small. If one sort produced by the subject is quite distant from all the others, for example, this distance will appear only once in the minimum spanning tree. We regard the MST measure as evaluating the *complexity* of the collection of sorts produced by a subject. If a subject has produced a number of sorts that represent sort criteria having a high degree of "pairwise orthogonality", we would expect that two sorts produced by this subject would exhibit relatively small overlaps and thus a large edit distance. Consequently the MST measure of the sorts produced by the subject would be large. Observe that if a subject identifies two different sort criteria but produces sorts that are identical (the same number of piles and the same cards in the piles, up to permutation), the edit distance between these two sorts will be zero, which will contribute a zero term in the sum of the weights in the minimum spanning tree.

As an alternative to the MST measure, we considered the sum of *all* the pairwise edit distances between the sorts of a given subject. However, adding all the pairwise edit distances can result in a large value even if there is just one sort that is distant from several of the others. For example, consider two graphs representing two possible configurations of edit distances between four sorts, illustrated in Figure 1.

The sum of the edit distances in both of these graphs is 30. However, the MST measure of the left-hand graph is 10 while the MST measure of the right-hand graph is 14. (Minimum spanning trees for these graphs are illustrated with heavy lines for the edges.) We consider the right-hand graph in Figure 1 as having more "complexity" than the left-hand graph, and the MST measure captures this better than summing all the edit distances.

Definition. Another measure of card sort complexity is the number of sorts produced by a subject. Specifically,

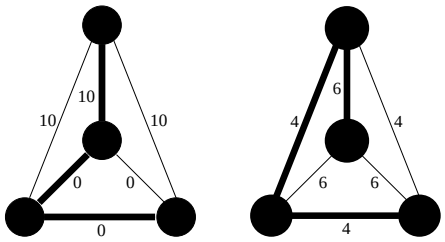


Figure 1: Example graphs

if the sorts produced by a subject are represented by the sequence $X = (S_1, S_2, \quad S_k)$, the *number of sorts (NSORT) measure* is k, and we define $n(X) = k$.

Since each new sort produced by a subject adds one edge to the minimum spanning tree used to compute the MST measure, the MST measure will normally increase as the number of sorts increases. We regard the NSORT measure as significantly less informative than the MST measure in taking into account card sort complexity for a given subject.

Note that if a subject produces just one sort, the underlying graph will have just one node and its spanning tree will have no edges, so the MST measure will be zero.

Definition. As $n(X)$ increases, we have remarked that $\mu(X)$ generally increases as well (see section). We define a measure that "factors out" the number of sorts as a contribution to the complexity of the set of sorts produced by a subject. The *normalized minimum spanning tree (NMST) measure* is

$$\nu(X) = \mu(X)/n(X)$$

.

Example

We illustrate our MST measure with two subjects in the study, both of whom produced three sorts. We will call them A and B (not their real names).

The sorts A_1, A_2, and A_3 produced by subject A have the following partitions:

Sort	Partitions
A_1	1,3,4,5,6,7,13,14,15,22,23
	2,8,9,10,11,12,16,17,18,19,20,21,24,25,26
A_2	1,3,4,6,7,10,13,14,15,18,23,26
	2,5,8,9,11,12,16,17,19,20,21,22,24,25
A_3	1,2,5,8,9,11,12,16,17,18,19,20,21,22,24,25
	3,4,6,7,10,13,14,15,23,26

For subject B, the sorts B_1, B_2, and B_3 are

Sort	Partitions
B_1	9,17,18,25
	1,3,5,8,12,13,20,26
	2,6,7,10,11,15,19,22
	4,14,16,21,23
	24
B_2	12,20
	18,25
	3,13,19
	5,10,15
	1,17
	8,9
	2,4,6,7,11,14,16,21,22,23,24,26
B_3	1,3,5,6,7,9,10,11,12,14,15,16,18,19,20,21,25,26
	2,4,8,13,17,22,23,24

The edit distances between the sorts for each of these subjects are given in the following tables:

Subject A

x	y	$d(x,y)$
A_1	A_2	5
A_1	A_3	5
A_2	A_3	2

Subject B

x	y	$d(x,y)$
B_1	B_2	15
B_1	B_3	18
B_2	B_3	18

The above tables give the pairwise edit distances between sorts for each subject, from which it is easy to determine the MST measures μ:

X	$\mu(X)$
A	7
B	33

Observe that the MST measure of subject B is much larger than that of subject A. Visually comparing the sorts produced by these two subjects, this is not surprising.

Comparison with number of sorts

The two subjects, A and B, were chosen because their MST measures were the smallest and largest among all subjects who produced exactly three sorts. Recalling that $n(X)$ represents the number of sorts produced by subject X, the following table gives the range of values of $\mu(X)$ for all values of $n(X)$ in the study. This table also includes the frequencies of subjects with the given number of sorts.

$n(X)$	**freq.**	**min** $\mu(X)$	**max** $\mu(X)$
1	2	0	0
2	30	3	19
3	63	7	33
4	67	13	46
5	40	20	58
6	33	23	69
7	17	30	66
8	12	32	77
9	5	45	72
10	3	35	72
11	2	64	98
13	1	50	50
14	1	79	79

This table shows that the MST measures $\mu(X)$ generally increase as the NSORT measures $n(X)$ increase and that for a given $n(X)$, there can be large variations among the values of $\mu(X)$.

Comparing Students to Experts

We expect that the greater experience and knowledge base of an expert will show up by experts producing larger NSORT values and that their sorts exhibit greater "orthogonality" (*i.e.*, larger pairwise edit distances) compared to non-experts, resulting in larger MST measures.

We use the nonparametric Wilcoxon two-sample test (also called the Wilcoxon-Mann-Whitney test) [2] to compare observations from two populations, with the null hypothesis being that the observations come from the same distribution. We use a one-sided test when a one-sided alternative hypothesis is supported by the observed data.

We will examine three measures to compare observations for students and experts: NSORT, MST, NMST.

NSORT measure

We first examine the numbers of sorts $n(X)$ produced by students and experts, with 243 student observations and 33 expert observations. The Wilcoxon two-sample test using numbers of sorts yields a z score of 1.33. Since the observed data suggests a one-sided alternative (the observed number of sorts produced experts is generally larger than the number of sorts produced by students), we reject the null hypothesis at the 10% level.

MST measure

We next examine the MST measures $\mu(X)$ produced by students and experts. The Wilcoxon two-sample test using MST measure observations yields a z score of 2.48, and we reject the null hypothesis with a one-sided alternative at the 1% level.

NMST measure

Since the MST measure more strongly distinguishes students from experts compared to the NSORT measure, we conjecture that the NMST measure $\nu(X)$ will factor out the less powerful number of sorts and produce a stronger difference between students and experts. Indeed, using the NMST measure observations yields a z score of 2.61, and we reject the null hypothesis at the 0.5% level.

Summary

The Wilcoxon-Mann-Whitney z scores for the three measures are given in the following table:

Students & Experts

measure	z	reject at
NSORT	1.33	10%
MST	2.48	1%
NMST	2.61	0.5%

Student Performance Levels

The researchers in this study were asked to give each student participant a performance ranking from 1 to 5, where 1 represents an assessment of low performance and 5 an assessment of high performance. Only 217 of the 243 students appear in the study ranking data, so we restricted our observations to this subset.

We used the nonparametric Mann-Kendall test (also called the *S-Test* in [2]) to test for randomness against a monotone trend. The "trend" is represented by an increase in student performance levels. Consider the MST measure, for example. If there is no trend (*i.e.*, a random trend) relating MST measure and performance level, then an MST measure for a lower-performing student would be just as likely to be larger than the MST measure for a higher-performing student as to be smaller. If there is a monotone, positive-going trend, then an MST measure for a lower-performing student would be less likely to be larger than the MST measure for a higher-performing student than to be smaller.

As with our comparison of students and experts, we again examined three measures: NSORT, MST, and NMST.

NSORT measure

Using NSORT measures to test for a monotone trend, the Mann-Kendall test produced a z score of 3.40, and we reject the null hypothesis at the 0.03% level.

MST measure

Using the MST measure μ, the Mann-Kendall test produced a z score of 3.21, and we reject the null hypothesis at the 0.06% level.

NMST Measure

Using the NMST measure ν, the Mann-Kendall test produced a z score of 0.37, which does *not* support rejecting the null hypothesis.

Summary and Analysis

The Mann-Kendall z scores for the three measures are given in the following table:

Monotone Trend

measure	z	reject at
NSORT	3.40	0.03%
MST	3.21	0.06%
NMST	0.37	not significant

These results show that differences among performance levels seem to correlate best with the numbers of sorts produced by the student subjects (as measured by $n(X)$) and not by the complexity of their sorts (as measured by $\nu(X)$). We suggest that better students are characterized as being more likely to take risks by producing a larger number of sorts, but that they do not have the depth of knowledge of an expert that would be necessary for their sorts to be highly orthogonal to one another.

Comparing high performing students to Experts

If high-performing students tend to have a larger number of sorts but with relatively little complexity, we predict that high-performing students will be similar to experts in numbers of sorts but that experts will have sorts that are measurably more complex. The Wilcoxon-Mann-Whitney z scores for the three measures are given in the following table:

High Performers & Experts

measure	z	reject at
NSORT	0.35	not significant
MST	0.66	not significant
NMST	1.83	5%

As predicted, the number of sorts for high performing students and experts are similar, but the NMST measure – which captures best the complexity of the sorts – statistically distinguishes experts from high performing students.

Comparing Females and Males

We conjecture that there are no differences between female (58) and male (184) subjects in our student populations. The Wilcoxon-Mann-Whitney z scores for the three measures are given in the following table:

Females & Males

measure	z	reject at
NSORT	0.34	not significant
MST	0.06	not significant
NMST	0.87	not significant

While the statistical results are not conclusive, the actual observations show that females in the study produced fewer sorts than their male counterparts, but the sorts they did produce were more complex. This is an area of interest for further study.

Since the data available to us did not include gender information on experts, we were unable to compare females and males among the experts.

Random sorts

We can expect a human to put cards into categories based on a sort criterion in a way that is very different from just throwing cards randomly into piles. On average, we would expect much larger edit distances between two random sorts than between two sorts produced by human subjects. In this section, we show that our MST and NMST measures distinguish human subjects from "random" subjects in a statistically significant way.

We first built a table giving the frequency distribution of the number of sorts for our human subjects. The fre-

quency data has already been given in a table in section .

For a given sort, we defined the sort *footprint* as the list of the sizes of each of the piles in the sort, in increasing order. (In mathematical terminology, this list is called a *partition*.) We then built a frequency distribution of all the sort footprints for the human subjects. The following table illustrates the distribution, displaying only part of the entire footprint data:

footprint	freq.
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3	1
1 1 1 1 1 1 1 1 1 5 12	1
1 1 1 1 1 1 1 1 1 5 13	1
1 1 1 1 1 1 1 2 2 2 2 2 4 5	1
1 1 2 2 3 3 3 3 3 5	2
1 1 2 2 3 3 3 3 5 6	1
1 1 2 2 3 3 3 4 4 6	2
8 8 10	10
8 9 9	7
9 17	26

Using these frequency distributions, we generated 100 "subjects". For each generated subject, we randomly generated the number of sorts for the subject using the frequency distribution in section from our sample population. To generate a random sort for this subject, we randomly selected a sort footprint using the frequency distribution given in this section. Once we identified a randomly generated footprint, we randomly assigned cards to piles so that the sizes of the piles matched the selected footprint. In this way, the "subjects" we generated have a distribution of number of sorts similar to that of the human population, and each "subject" has a distribution of sort footprints similar to that of the human population.

Because the sorts for the generated "subjects" are randomly generated, we hypothesize that they will exhibit significant "complexity", in the sense that their pairwise edit distances will be large. Consequently, we expect that the randomly generated "subjects" will produce MST measures (both un-normalized and normalized) that are significantly larger than the human subjects.

After generating one such set of "subjects", we compared them to human subjects using the NSORT, MST, and NMST measures. The following table summarizes the results:

Human Subjects & Random Subjects

measure	z	reject at
NSORT	0.62	not significant
MST	11.07	off the charts
NMST	14.79	off the charts

Observe that the random population of subjects was generated in such a way that their distribution of number of sorts should be similar to the distribution of number of sorts from the human population. Consequently, we expect that the NSORT measures of these two populations should not be significantly different. Our statistics show that this is the case. However, the MST and NMST measures of human subjects are clearly different from those of the random subjects.

Conclusions

We conclude that our MST measure μ appropriately quantifies the combination of number of sorts and the complexity of these sorts for a subject. Subjects with larger MST values exhibit greater numbers of sorts with greater complexity. Since the MST measure, to some extent, includes a measure of the number of sorts of a subject, the NMST measure ν factors out the number of sorts, giving a value that distills the complexity of the sorts.

Comparing experts to students in the NSF workshop population, all the measures (NSORT, MST, and NMST) are significantly larger for experts compared to students. The NSORT measure is the least significant measure, and the NMST measure is the most significant.

Comparing students based on performance level, both the NSORT and MST measures have significant positive-going trends compared to performance level, but the NMST measure is not significant. This suggests that higher-performing students produce more sorts, but that because of their lack of experience (in this study, the student subjects were at the CS2 level), their sorts did not exhibit greater complexity. High-performing students and experts had similar NSORT measures, but the experts had significantly larger NMST measures, which gives further evidence to our observations about higher-performing students in the student population.

Comparing females to males in the student population, the three measures (NSORT, MST, and NMST) showed no significant differences. Examining the data, however, shows that the NSORT values appear to be smaller for females than males, but that the NMST values appear to be larger. This is an area that warrants further study.

We have shown that the MST and NMST measures of card sort complexity are statistically significant in discriminating between experts and non-experts. These measures may have broader applicability to other experiments requiring complexity measures.

References

[1] R. Anderson, R. Anderson, and K. Deibel. Analyzing concept groupings of introductory computer programming students. In submission.

[2] G. Noether. *Introduction to Statistics: A Nonparametric Approach*. Houghton Mifflin Company, 2nd edition, 1976.

[3] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, 1982.

[4] M. Petre, S. Fincher, J. Tenenberg, et al. "My criterion is: Is it a Boolean?": A card sort elicitation of students' knowledge of programming constructs. Technical report, University of Kent, 2003.

[5] G. Rugg and P. McGeorge. The sorting techniques: A tutorial paper on card sorts, picture sorts, and item sorts. *Expert Systems*, 14(2):80–93, 1997.

Simulating the temporal reference of Dutch and English Root Infinitives

Daniel Freudenthal (DF@Psychology.Nottingham.Ac.Uk)
School of Psychology, University Park
Nottingham, NG7 2RD UK

Julian Pine (JP@Psychology.Nottingham.Ac.Uk)
School of Psychology, University Park
Nottingham, NG7 2RD UK

Fernand Gobet (Fernand.Gobet@Brunel.Ac.UK)
Department of Human Sciences, Brunel University
Uxbridge, Middlesex, UB8 3PH UK

Abstract

Hoekstra & Hyams (1998) claim that the overwhelming majority of Dutch children's Root Infinitives (RIs) are used to refer to modal (not realised) events, whereas in English speaking children, the temporal reference of RIs is free. Hoekstra & Hyams attribute this difference to qualitative differences in how temporal reference is carried by the Dutch infinitive and the English bare form. Ingram & Thompson (1996) advocate an input-driven account of this difference and suggest that the modal reading of German (and Dutch) RIs is caused by the fact that infinitive forms are predominantly used in modal contexts. This paper investigates whether an input-driven account can explain the differential reading of RIs in Dutch and English. To this end, corpora of English and Dutch Child Directed Speech were fed through MOSAIC, a computational model that has already been used to simulate the basic Optional Infinitive phenomenon. Infinitive forms in the input were tagged for modal or non-modal reference based on the sentential context in which they appeared. The output of the model was compared to the results of corpus studies and recent experimental data which call into question the strict distinction between Dutch and English advocated by Hoekstra & Hyams.

Root Infinitives in Child Language

A striking feature of the speech of children who are acquiring their native language, is that, in many languages, children go through a stage where they produce *Root Infinitives* (non-finite verb forms in contexts that require a finite verb form). Thus, English-speaking children may produce utterances such as (1), and Dutch children may produce utterances such as (2).

(1) Daddy drink coffee
(2) Papa koffie drinken (Daddy coffee drink-inf)

This phenomenon has been subject to considerable linguistic theorizing, as the fact that it occurs in several languages (including English, Dutch, Swedish, German and French) suggests the operation of invariant principles. A particularly influential Nativist theory has been provided by Wexler (1994). According to Wexler's *Optional Infinitive (OI) Hypothesis*, by the time children begin to produce multi-word speech, they have already correctly set all the basic inflectional and clause structure parameters of their language. They thus have adult-like knowledge of the word order and inflectional properties of the language they are learning. However, there is a stage of development (the Optional Infinitive stage), during which the abstract features of Tense (TNS) and Agreement (AGR) can be absent from the underlying representation of the sentence. This results in children initially using both finite and non-finite verb forms in contexts in which a finite form would be obligatory in the adult language. The great strength of the Optional Infinitive Hypothesis is that it explains the data from a wide variety of languages, as well as the relative sparseness of other errors. However, the theory also has some important weaknesses.

Firstly, the theory assumes a large amount of innate knowledge and ignores the possibility that the Optional Infinitive phenomenon may be understood as the result of an input-driven learning process without the need to assume large amounts of innate knowledge. Simulations with the MOSAIC model have already shown that a simple learning mechanism which is sensitive to the distributional characteristics of the input can give a close quantitative fit to the prevalence of the Root Infinitives in English and Dutch over a range of MLUs (Freudenthal, Pine & Gobet (2002a, 2003, in preparation).

Secondly, while the Optional Infinitive phenomenon occurs in several languages, cross-linguistic differences exist in the finer detail of the phenomenon that are problematic for Wexler's theory. One obvious way of explaining these differences is in terms of differences in the distributional characteristics of the language being learned. In this paper we assess the viability of such an explanation by simulating cross-linguistic differences in the fine detail of the OI-phenomenon in Dutch and English using MOSAIC. The central aim of this paper is therefore to investigate whether the same mechanism that captures one of the key similarities in the speech of children learning different languages can also capture differences in the way

that this phenomenon patterns as a function of differences in the languages being learned, and hence can provide a unified account of patterns of cross-linguistic similarity and difference in children's early multi-word speech.

The Modal Reference of Root Infinitives

The majority of Root Infinitives that Dutch children produce carry a modal meaning: they tend to express desires and wishes, or relate to unrealized events. Hoekstra & Hyams (1998) have dubbed this the *Modal Reference Effect*. The type of verbs that occur as Root Infinitives also differs from inflected verbs. Dutch speaking children appear to use Root Infinitives when referring to actions rather than static situations. This has been called the *Eventivity Constraint*. Wijnen (1996), analysed the speech of four Dutch children, and found that 95% of the children's Root Infinitives contained eventive verbs, and 85% of the Root Infinitives had a modal reference, thus confirming the Modal Reference Effect and Eventivity Constraint. According to Hoekstra & Hyams, the Modal Reference Effect and Eventivity Constraint do not hold for English. They present data based on an (unpublished) paper by Ud Deen (1997), who found that only 13% of English Root Infinitives carry a modal meaning. Ud Deen also found that, while the majority of English Root Infinitives are eventive in nature, this effect is less pronounced than it is in Dutch, with 75% of English RIs containing eventive verbs. Hoekstra & Hyams explain this cross-linguistic difference by referring to differences between the English and Dutch infinitive form. The English infinitive, they claim, is not a true infinitive, but a 'bare form'. Dutch has a true infinitive as it has an infinitival morpheme. This infinitival morpheme is thought to carry an *irrealis* feature which is responsible for the modal reference. This, they argue, is evident from the analysis of the following utterances:

3. I see John cross the street*
4. I saw John cross the street
5. I see John crossing the street

Utterance (3) is ungrammatical in English, because the English bare form denotes 'not only the processual part of the event, but includes the completion of that event' (Hoekstra & Hyams 1998, p. 105). A correct description of an ongoing event in English would therefore require the use of the past tense as in (4), or the progressive as in (5). Sentence 6 makes it clear that this constraint does not operate in Dutch: an ongoing event may be described using a present tense construction. Apparently, the Dutch infinitive does not signal completion of the event.

6. Ik zie/zag Jan de straat oversteken
   I see/saw John the street cross-INF
   I see/saw John cross the street.

This difference between the English and Dutch infinitival form also explains the difference with respect to the eventivity of Root Infinitives, as, according to Hoekstra & Hyams it is the modal reading of Dutch Root Infinitives that forces the selection of an eventive verb. Since English Root Infinitives are not exclusively modal, they can occur with stative as well as eventive verbs.

Problems with Hoekstra & Hyams' Account

While the Hoekstra & Hyams' account explains the differential reading of Dutch and English Root Infinitives, it predicts that the proportion of modal readings of RIs in Dutch and English is radically different. Theoretically, all RIs in languages with an infinitival morpheme should be modal, while the reference of English RIs is free. The proportion of modal RIs in Dutch and German appears to be considerably lower than 1.00 however. Wijnen (1996) reports a proportion of .85 averaged over 4 children, and Ingram & Thompson (1996) report a proportion of .55 using a strict criterion and .79 using a lenient criterion.

Ingram & Thompson also suggest that the modal reading of RIs in German (and Dutch), is caused by the fact that infinitive forms in adult German and Dutch are typically used in conjunction with a modal, as in (7) and (8). Since Dutch-speaking children predominantly hear infinitive forms in modal contexts in the input, they come to associate these forms with the modal reading and use them predominantly to express desires.

7. Ik ga morgen werken (I go-FIN Tomorrow work-INF)
8. Wil je spelen? (Want-FIN you play-INF)

The proportion of modal Root Infinitives in English may also be considerably higher than the .13 that was found in a corpus study by Ud Deen. An inherent weakness of corpus studies is that the modal/nonmodal reading of an utterance is assigned on the basis of the context in which it is produced. However, since the corpora are transcripts of spontaneous speech, the information required to discriminate between modal and non-modal readings is often lacking. For this reason, Blom, Krikhaar & Wijnen (2001) conducted an experiment in which children produced descriptions of modal and non-modal events. In the experiment, the majority of Dutch children's Root Infinitives (68%) were used to describe modal events. For the English children this was 44%. While this difference was significant and in the expected direction, this finding is problematic for Hoekstra & Hyams, as it suggests that the difference between Dutch and English is not a qualitative difference, but a graded, quantitative one which may well be related to the distributional characteristics of the language rather than differences in infinitival morphology.

In this paper, MOSAIC will be used to investigate the source of the differential reading of Dutch and English Root Infinitives. MOSAIC has a number of characteristics that make it a suitable candidate for such an investigation.

Firstly, the model has already been shown to successfully simulate the developmental change in the prevalence of Root Infinitives in Dutch and English (Freudenthal, Pine & Gobet 2002a, 2003, in preparation), as well as phenomena related to Subject Omission in English (Freudenthal, Pine & Gobet 2002b). The model's success in simulating the finer detail of the OI phenomenon therefore provides a strong test

of an input-driven account of the OI phenomenon. Secondly, the model learns off Child Directed Speech. The use of Child Directed Speech ensures a realistic frequency distribution, so that differences in the surface characteristics of a language are reflected in the input in a quantitatively realistic way. This is of particular importance as the practice of using artificially created input sets (which is common in simulations of phenomena in child speech) may lead the researcher to misrepresent the distributional characteristics responsible for the phenomenon under investigation.

Thirdly, MOSAIC uses no built-in linguistic knowledge. Whatever representations it builds up during learning are a result of the interaction between its learning mechanism and the distribution of the input it sees. This last characteristic is important because Hoekstra & Hyams' explanation of the differential reading of RIs is dependent on the assumption that the child knows that the infinitival morpheme implies a modal interpretation (rather than learning the association through exposure to the input). MOSAIC will be described below, followed by the details of the simulation.

Simulating Language Acquisition in MOSAIC

Whilst the version used for the simulations discussed here has changed from the earlier simulations, the main theoretical underpinning of the model remains the same. The basic tenet of the model is that the learning of language is a performance-limited process which is heavily weighted towards the most recent elements in the speech stream (i.e., which has an utterance final bias). Several authors have argued that children are better at learning material that occurs towards the end of the utterance (Naigles & Hoff-Ginsberg, 1998; Shady & Gerken, 1999; Wijnen et al. 2001).

MOSAIC learns from orthographically coded input, with whole words being the unit of analysis. The model is a simple discrimination net (an n-ary tree) which is headed by a root node. At the start of learning the discrimination net consists of just the root node. More nodes (encoding words or phrases) are added as the model is shown more utterances. An important requirement for nodes to be added is that whatever follows the word to be encoded in the input, must already have been encoded in the model. That is, the model will only learn a new word, when it has already encoded the rest of the utterance. This results in the model building up its representation of the utterances it is shown by starting at the end of the utterance, and slowly working its way to the beginning.

If the model were to see the utterance *I go home* three times, it would on its first pass encode the fact it has seen the word *home* at the end of an utterance. On the second pass, it would encode the sequence *go home*. After a third pass, it would have encoded the whole utterance. Figure one gives a graphical representation of the model at this stage.

The fact that MOSAIC builds its representation of an utterance by starting at the end of the utterance is the major mechanism responsible for its simulation of the development of Root Infinitives in Dutch. Early in Dutch children's development, 80-90% percent of their utterances containing verbs are Root Infinitives. This drops to 10-20% later in development (Wijnen et al, 2001). Early in training, the model encodes many utterance final phrases. Since the infinitive takes sentence-final position (as can be seen in examples 7 and 8), the model produces many utterances with only non-finite verb forms. As the model encodes longer and longer phrases, these Root Infinitives are slowly replaced by auxiliary/modal plus infinitive constructions.

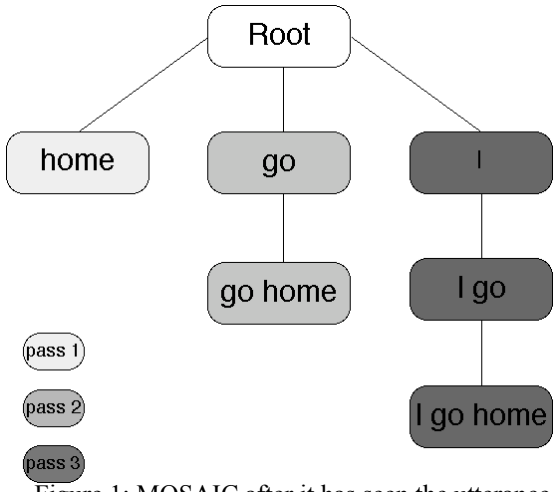


Figure 1: MOSAIC after it has seen the utterance *I go home* three times.

In the example illustrated in Figure 1, a (sentence final) word is encoded after one exposure. In fact, MOSAIC actually learns much more slowly than this, and the input corpus is fed through the model several times, so output of increasing average length can be generated after consecutive exposures to the input corpus. The probability of creating a node in MOSAIC is given by the following formula:

$$NCP = \left(\frac{1}{1 + e^{m - u/c}} \right)^{\sqrt{d}}$$

where: NCP = Node Creation Probability

m = a constant, set to 20 for these simulations.
c = corpus size.
u = total number of utterances seen.
d = distance to the end of the utterance.

The formula results in a basic sigmoid curve. The formula contains the size of the corpus and total number of utterances seen. The size of the corpus is included because the size of the available input corpora differs considerably (13,000 to 30,000 utterances for the corpora used in these simulations). The use of the term (m – u/c) ensures that after n presentations of the complete input corpus the Node Creation Probability is identical for corpora of different sizes. The 'distance to the end of the utterance' in the exponent causes material that occurs near the beginning of the utterance to have a lower likelihood of being encoded than material that occurs near the end. This effect decreases as the model sees more input.

Production of Novel Utterances

Utterance production in MOSAIC involves outputting all the utterances the model has encoded. However, the output that MOSAIC produces consists of more than the input it has seen. MOSAIC has a mechanism for linking words or phrases that have occurred in similar contexts. When the overlap between two words is sufficiently high (more than 10% of both the words that preceded and followed the target words are the same), the two words get linked. Two words that are linked can be substituted for each other when the model produces output. This mechanism allows MOSAIC to produce utterances that were not present in the input. Redington, Chater & Finch (1998), and Mintz (2003), have shown that similar mechanisms based on co-occurrence statistics can be quite effective in grouping words that are of the same syntactic category.

The model also includes a chunking mechanism, which results in frequent multi-word phrases being treated as one unit. Since the chunking mechanism does not play an important role in these simulations, it is not discussed any further in this paper.

The Simulations

In order to distinguish modal from non-modal Root Infinitives in the output of the model, infinitive verb forms in the input were tagged to reflect the context in which they occurred. In order to do this, all utterances in the input were (automatically) searched for words denoting a modal or not-realised context. These words constituted the standard modals, as well as some other words denoting a not-realised context, including *want* and *go*. The utterances containing a word denoting a modal context were then searched for words that matched an infinitive form. If such a word was found, it was tagged for having occurred in a modal context (by adding –MOD to the verb).

For both languages, two input corpora (available through the CHILDES data base (MacWhinney 2000)) were selected. For Dutch, these were the corpora of Matthijs (Simulation1) and Peter (Simulation2). For English the corpora of Anne (Simulation1) and Becky (Simulation2) were selected. The size of the corpora was approximately 13,000 utterances for Dutch, and 30,000 utterances for the English input. Since the input corpora consist of Child Directed Speech, the distributional characteristics are representative of the language that children hear. Each input corpus was fed through the model iteratively until the output reached a Mean Length of Utterance (MLU) of approximately 2.5. At this stage, all the utterances the model could produce were generated. The full output consisted of approximately 7,500 utterances for the Dutch corpora, and 15,000 for the English corpora. Next, all utterances in which all verbs matched the infinitive form were selected. The proportion of these utterances that had the infinitive form tagged for a modal context (i.e. had been learnt off a modal context) was then calculated. Table 1 gives the results for the English and Dutch simulations. For both Dutch simulations the proportion of modal infinitives is larger than it is for English. For all four possible comparisons of Dutch against English simulations, the difference was statistically significant ($\chi 2(1) > 12.00$ $p < .001$).

Table 1: Proportion of modal infinitives and total number of infinitives for Dutch and English simulations.

	Dutch	English
Simulation1	.70 (1447)	.47 (1577)
Simulation2	.54 (1474)	.48 (2581)

While these values are very close to those reported by Blom et al., this analysis ignores the complication that (especially in English) it is difficult to unambiguously identify the infinitive. In the English present tense, only the third singular differs from the infinitive. A true Root Infinitive can therefore only be identified in a third singular context (since many of the forms resembling the infinitive may actually be correctly inflected finites).

A second analysis was therefore performed on the subset of utterances that had a third singular subject (e.g. *He go/Hij gaan*). While the Dutch data could be restricted to all singulars rather than just third singular, it was considered preferable to use the same restriction for both languages. The results of these analyses are shown in Table 2.

Table 2: Proportion of modal infinitives and total number of infinitives for Dutch and English simulations, third singular context only.

	Dutch	English
Simulation1	.66 (104)	.43 (89)
Simulation2	.58 (102)	.42 (118)

Again, the proportion of modal RIs is larger in the Dutch simulations than it is in the English simulations (For all four possible comparisons, $\chi 2(1) > 4.30$, $p < .05$). Thus, MOSAIC clearly captures the difference between the two languages, suggesting that the differential reading of Root Infinitives in English and Dutch is related to the surface characteristics of the languages.

Having established that MOSAIC simulates the differential reading of RIs, we can now assess whether MOSAIC simulates the difference in verb types that occur in Root Infinitives. Hoekstra & Hyams cite a paper by Wijnen (1996) who found that 95% of Dutch Root Infinitives contained eventive verbs. In English, Ud Deen (1997), found that only 75% of Root Infinitives contained eventive verbs. While a direct comparison between these numbers is difficult as Wijnen and Ud Deen used a different set of verbs, stative verbs do appear to be used more often in English than in Dutch Root Infinitives. In order to perform a more controlled analysis, all Root Infinitives in Table 2 were coded for whether the main verb denoted an event or not. As can be seen in Table 3, MOSAIC does simulate the effect, though only three out of four differences are

statistically significant ($\chi2(1) > 6.20$, $p < .02$). The value of .92 for Dutch (simulation2) was not significantly different from .87 (English, simulation2). Inspection of the non-eventive verbs in simulation2 for Dutch revealed that 5 out of the 8 instances consisted of the verb *zien* (see), which was linked to the verb *kijken* (look), and was thus substituted in production. It thus appears that the generativity mechanism may have inflated the proportion of non-eventive Root Infinitives for this simulation.

Table 3: Proportion of Root Infinitives (third singular context only) that have an eventive main verb.

	Dutch	English
Simulation1	.98	.80
Simulation2	.92	.87

What causes the Modal Reference Effect?

The fact that MOSAIC simulates the difference between Dutch and English for both the modal reading of Root Infinitives and the eventivity of the main verb, suggests that these effects are related to differences in the surface characteristics of the two languages. In order to understand what these relevant surface characteristics might be, it is useful to examine more closely in what contexts third singular subjects are followed by a form resembling the infinitive. In both Dutch and English, the majority of the contexts are likely to be questions. English and Dutch differ however, in the way questions are formed. Whereas Dutch uses inversion to transform a declarative into a question, in English, the auxiliary *do* is inserted, resulting in phrases such as *Does he go*. While the auxiliary *do* patterns like a modal, it does not carry a modal or not-realised meaning. In English, a third singular subject followed by an infinitive form can therefore occur in both a non-modal (*Does he go*), and a modal context (*Can he go*). In Dutch, a third singular followed by an infinitive form can occur in double verb constructions such as *Kan hij fietsen* (Can he cycle-inf), or *Ik zie Jan lopen* (I see John walk-inf). The first of these is modal, but the second is not. Root Infinitives in the input are a further source of third singular plus infinitive contexts. In some situations (for example elliptical answers to questions) Root Infinitives are allowed in Dutch. In the two languages, the third singular is thus followed by the infinitive in different contexts. If a child learns Root Infinitives off the input, it is likely to produce them in the context in which they are most frequently encountered. One way of directly testing such an input-driven account of Root Infinitives is to search the input for occasions where a third singular is followed by an infinitive form, and noting whether the context is modal or not. Table 4 presents the results of such an analysis (using the most frequent third singular subjects that were present in the Root Infinitives of the model's output). Table 4 shows that in Dutch third singular plus infinitive constructions occur in a modal context more often than they do in English (again all four differences are statistically significant $\chi2(1) > 6.2$, $p < .02$).

Table 4: Proportion of modal contexts for third singulars followed by infinitive in English and Dutch input corpora.

	Dutch	English
Input1	.77 (113)	.37 (212)
Input2	.56 (140)	.41 (153)

Also of interest now is the question of what the non-modal contexts are in which these constructions occur. In English a large majority (90%) of these non-modal contexts do turn out to be 'do-questions'. Thus, while third-singular plus infinitive constructions can occur in modal as well as non-modal contexts, the dummy modal *do* alone makes up close to 40% of these contexts. In the Dutch input, non-modal contexts are limited in number, and largely confined to Root Infinitives (though a few double verb constructions do occur). The majority of non-modal Root Infinitives therefore appear to be learned off double-verb constructions in English and off Root Infinitives in Dutch.

Turning to the eventive-stative distinction it now also becomes apparent why Root Infinitives are more likely to contain statives in English. In English, stative verbs such as *want* or *need* frequently occur in utterances like *Does he want it*. In Dutch such an utterance does not carry the inflection on the dummy modal, but on the inverted main verb (*Wil hij dat; Wants he that*). As a result, statives in Dutch only occur as infinitives in very infrequent double verb constructions such as *Dat zou hij willen (That would he want*) The higher proportion of statives in English Root Infinitives can therefore also be explained in terms of the use of the dummy modal *do* in the input.

Conclusions

While Wexler's Optional Infinitive Hypothesis can explain the cross-linguistic occurrence of Root Infinitives, additional assumptions are required the explain the cross-linguistic differences in the fine detail of the phenomenon. Hoekstra & Hyams (1998), explain some of this fine detail (the differential temporal reference of Dutch and English children's Root Infinitives) by referring to qualitative differences between the languages. Specifically, they argue that the English infinitive is a bare form rather than a true infinitive. They argue that the English bare form does not carry the *irrealis* feature, which signals the modal reading of the infinitive. They therefore postulate qualitative differences in the fine detail of infinitival morphology between the two languages. Since children are thought to know the full grammar, English-speaking children will use Root Infinitives both in modal and non-modal contexts, and Dutch children should use Root Infinitives overwhelmingly in modal contexts. Experimental work by Blom et al. has shown however, that the difference is not as large as predicted by Hoekstra & Hyams, suggesting that it is a graded quantitative, rather than a qualitative difference.

In this paper MOSAIC was used to investigate whether the difference between Dutch and English could be related

to the surface characteristics of the two languages. The fact that MOSAIC has already been used to simulate the developmental patterning in the prevalence of Root Infinitives (relative to simple and compound finites) in English and Dutch lends credence to the basic mechanism for producing Root Infinitives that MOSAIC employs. While MOSAIC is clearly insufficient as a full model of the language acquisition process, it is a valuable tool for exploring how an utterance final learning bias can interact with the distributional properties of the input to produce the phenomena apparent in children.

MOSAIC clearly simulates the difference between English and Dutch in terms of the modal reading as well as the eventive/stative nature of the verbs in Root Infinitives. The fact that both phenomena can be explained by the use of the dummy modal 'do' (which patterns like a modal without ascribing modal meaning) shows that subtle differences in the distributional characteristics of languages can have quite profound effects on the patterning of child data. As these effects can be quite difficult to predict a priori, computational modelling can be an invaluable tool for investigating the complex relations between input characteristics and child speech.

The analyses reported here also underscore the importance of using realistic input when simulating child speech, as some effects are only likely to be simulated when realistic input is used. While dummy *do* is only one of many modals in English, it accounts for 40% of the third singular + infinitive contexts. However, *do* is also responsible for the occurrence of eventive RIs. Since these occur at a maximum rate of 20% in these simulations, an underestimation of the incidence of *do* in artificial input would likely result in failure of the model to simulate the effect. This is even more apparent when one realises that third singular plus infinitive contexts are largely restricted to questions; it seems unlikely that without prior knowledge of the importance of questions a researcher constructing artificial input would include 'do-questions' at a sufficient rate to simulate these effects.

The fact that the interaction between the learning mechanism and the distributional characteristics of the input can produce the different levels of modal Root Infinitives in the two languages strongly suggests that the observed phenomenon is related to surface characteristics (the contexts in which infinitives occur), rather than to differences in the infinitival morphology between the languages. Since MOSAIC does not have any knowledge of infinitival morphology (in fact does not employ any built in linguistic knowledge), these results clearly show that the Modal Reference Effect can be explained without assuming that children know the full grammar of their language, or the fine detail of infinitival morphology.

Acknowledgements

This research was funded by the ESRC under grant number RES000230211

References

Blom, E., Krikhaar, E. & Wijnen, F. (2001). Nonfinite clauses in Dutch and English child language: An experimental approach. In: A. H-J. Do, L. Domínquez & A. Johansen (Eds.): *Proceedings of the 25th annual Boston University Conference on Language Development*. pp. 133-144. Cascadilla Press.

Freudenthal, D., Pine, J. & Gobet, F. (2002a). Modelling the development of Dutch Optional Infinitives in MOSAIC. In: W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* pp. *322-327* Mahwah, NJ: LEA.

Freudenthal, D., Pine, J. & Gobet, F. (2002b). Subject omission in children's language; The case for performance limitations in learning. In: W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, pp. *328-333*. Mahwah, NJ: LEA.

Freudenthal, D., Pine, J. & Gobet, F. (2003). The role of input size and generativity in simulating language acquisition. In: F. Schmalhofer, R. Young, & G. Katz (Eds.): Proceedings of EuroCogSci 03. The European Cognitive Science Conference 2003. Mahwah NJ: LEA.

Freudenthal, D., Pine, J. & Gobet, F. (in preparation). Modelling the development of children's use of optional infinitives in English and Dutch using MOSAIC.

Ingram, D & Thompson, W. (1996). Early syntactic acquisition in German: evidence for the modal hypothesis. *Language*, *72*, *97-120*.

Hoekstra, T. & Hyams, N. (1998). Aspects of root infinitives. *Lingua*, *106*, *81-112*.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mintz, T.H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, *91-117*.

Naigles, L. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs. Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, 25, *95-120*.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science 22*, 425-469.

Shady, M. & Gerken, L. (1999). Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language*, 26, *163-176*.

Ud Deen, K. (1997). The interpretation of root infinitives in English: Is eventivity a factor? Term paper, UCLA.

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivation in child grammar. In N. Hornstein & D. Lightfoot (Eds.), *Verb Movement*. Cambridge: Cambridge University Press.

Wijnen, F. (1996). Temporal reference and eventivity in root infinitives. *MIT occasional Papers in Linguistics*, *12*, *1-25*.

Wijnen, F., Kempen, M. & Gillis, S. (2001). Root infinitives in Dutch early child language. *Journal of Child Language*, *28*, 629-660.

Extending the Computational Abilities of the Procedural Learning Mechanism in ACT-R

Wai-Tat Fu (wfu@cmu.edu)
John R. Anderson (ja+@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

The existing procedural learning mechanism in ACT-R (Anderson & Lebiere, 1998) has been successful in explaining a wide range of adaptive choice behavior. However, the existing mechanism is inherently limited to learning from binary feedback (i.e. whether a reward is received or not). It is thus difficult to capture choice behavior that is sensitive to both the probabilities of receiving a reward and the reward magnitudes. By modifying the temporal difference learning algorithm (Sutton & Barto, 1998), a new procedural learning mechanism is implemented that generalizes and extends the computational abilities of the current mechanism. Models using the new mechanism were fit to three sets of human data collected from experiments of probability learning and decision making tasks. The new procedural learning mechanism fit the data at least as well as the existing mechanism, and is able to fit data that are problematic for the existing mechanism. This paper also shows how the principle of reinforcement learning can be implemented in a production system like ACT-R.

Introduction

Human choice behavior is often studied under various probability learning situations. In a typical probability learning situation, participants are asked to select one of the many options available, and feedback on whether the choice is correct or not is given after the selection. There are usually two main manipulations in a probability learning task: (1) the probabilities for each of the options being correct, and (2) the magnitudes of reward (usually monetary) received when the correct option is selected. One robust result is that people tend to choose the options a proportion of time equal to their probabilities of being correct – a phenomenon often called "probability matching" (e.g. Friedman et al., 1964). However, when the reward magnitudes are varied, the observed choice probabilities are sometimes larger or smaller than the outcome probabilities (e.g. Myers, Fort, Katz, & Suydam, 1963). These studies show consistently that people are sensitive to both outcome probabilities and reward magnitudes in making choices.

One limitation of the current ACT-R procedural learning mechanism (Lovett, 1998) is that it requires a pre-specification of correct and incorrect responses. Besides, feedback received is limited to a binary function (i.e. whether a reward is received or not). Apparently, a simple binary function may not be sufficient to represent the feedback from the environment. For example, imagine a situation in which there are several possible treatments for a particular disease and a physician has to choose a treatment that has the highest expected effectiveness. One may have to evaluate the effectiveness of each treatment through case-by-case feedback. For example, consider the case where the probabilities of effectiveness of three treatments 1, 2, and 3 are as shown in Figure 1. Since the effectiveness of each treatment follows a continuous distribution, a simple binary feedback function is obviously insufficient to represent the information received from the feedback.

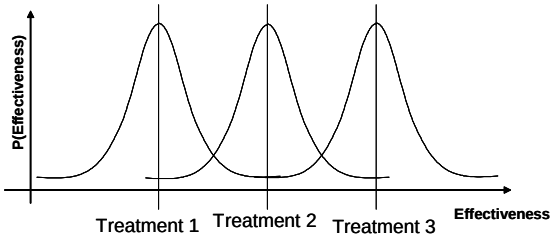


Figure 1. Probability of effectiveness of three treatments.

Another motivation for extending the current mechanism comes from recent findings of the functional role of dopaminergic signals in basal ganglia during procedural learning. Research shows that learning is driven by the deviation between the expected and actual reward (Schultz et al., 1995; Schultz, Dayan, & Montague, 1997). In other words, the reward magnitude is often processed as a scalar quantity – depending on whether the magnitude of the actual reward is higher or lower than expected, a positive or negative reinforcement signal is generated respectively. The pre-specification of correct and incorrect responses is therefore inconsistent with the current understanding of the procedural learning mechanism in basal ganglia.

The ACT-R 5.0 architecture

Figure 2 shows the basic architecture of the ACT-R 5.0 system. The core of the system is a set of production rules that represents procedural memory. Production rules coordinate actions in each of the separate modules. The modules communicate to each other through its buffer, which holds information necessary for the interaction between the system and the external world. Anderson, Qin, Sohn, Stenger, and Carter (2003) showed that the activity in these buffers match well to the activities in certain cortical areas (see Figure 2). The basal ganglia are hypothesized to implement production rules in ACT-R, which match and act on patterns of activity in the buffers. This is consistent with a typical ACT-R cycle in which production rules are matched to the pattern of activity in the buffers, a production is selected and fired, and the contents in the buffers updated. In ACT-R, when there is more than one production matching the pattern of buffer activity, the system selects a production based on a conflict resolution mechanism. The basis of the conflict resolution mechanism is the computation of expected utility, which captures the

effectiveness and efficiency of the production in accomplishing the goals of the system.

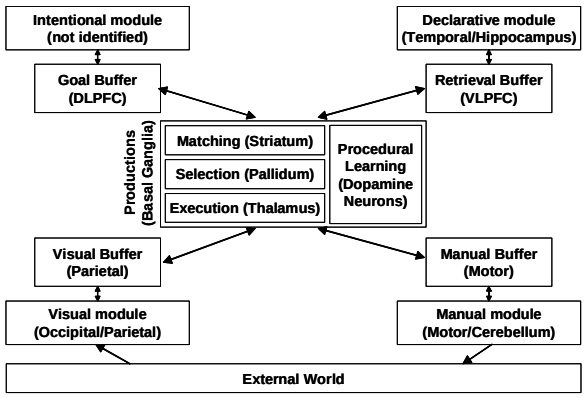


Figure 2. The ACT-R 5.0 architecture.

To adapt to the environment, the system must learn from the consequences of its actions so that when the same conditions are met in the future, a better choice of productions can be made. Procedural learning updates the expected utility of a production from the consequences of firing the production, and the dopamine systems in basal ganglia are believed to be involved in the learning process. Specifically, procedural learning appears to be coded by dopaminergic signals from the ventral tegmental area (VTA) and substatntia nigra to the striatum in basal ganglia (Schultz, et al., 1995; Schultz, et al., 1997), and different patterns are either reinforced or penalized according to the dopaminergic signals. Previous studies (Ljungberg, Apicella, & Schultz, 1992; Mirenowicz, Schultz, 1994) show that the activation of dopamine neurons depends entirely on the difference between the predicted and actual rewards. Once an unpredicted reward is perceived, response in dopamine neurons is transferred to the earlier reward-predicting stimulus. Inversely, when a predicted reward fails to occur, dopamine neurons are depressed in their activity at exactly the time when the reward would have occurred (Schultz, Apicella, Ljungberg, 1993). It therefore appears that dopamine signals do not simply report the occurrence of rewards. Rather, outputs from dopamine neurons appear to code for a deviation or error between the actual reward received and predictions or expectations of the reward. In other words, dopamine neurons seem to be feature detectors of the "goodness" of environmental events relative to the learned expectations about those events.

The current procedural learning mechanism

During each cycle of ACT-R, productions that match the contents of the buffer will be put into a conflict set. The productions in the conflict set are ordered in terms of their expected utility and ACT-R considers them according to that ordering. The expected utility of a production is defined as E = PG-C, where P is the estimated probability that the goal will be achieved if that production is chosen, G is the value of the goal, and C is the estimated cost of achieving the goal if that production is chosen (see Table 1). Procedural learning updates the value of P and C according to the following equations:

$$P = \frac{successes}{successes + failures} \qquad C = \frac{efforts}{successes + failures},$$ where

successes and *failures* are the number of times the production has succeeded or failed to accomplish the current goal respectively (i.e. a reward or penalty), and *efforts* is the total amount of time taken over all past uses of the production rule, successful or failed. These quantities start out with initial values that are updated with experience. For example, if for production n the initial *successes* equals 1, *failures* equals 1, and *efforts* equals 0.5, when a pre-specified *success* is encountered 0.1 second after k has fired, P will change from 0.5 to 0.67 (=2/(2+1)), C will change from 0.25 to 0.2 (=(0.5+0.1)/(2+1)). If G equals 20, then the expected utility (E=PG-C) will increase from 9.75 (=0.5*20-0.25) to 13.13 (=0.67*20-0.2). The successful experience has thus acted as a reward and reinforced production n by increasing its expected utility, and as a consequence, n will be more likely to be selected in the future..

Table 1. A list of free parameters and their definitions.

Parameters	Definition (Old mechanism)
G	Value of the goal (measured in seconds)
successes/ failures	Initial number of times the production has led to a success/failure state before the model starts
efforts	Total amount of time taken over all past uses of the production, successful or failed.
Parameters	**Definition (New mechanism)**
r_n	The actual reward received
K	The discount factor ($0 < K \leq 1$). Future rewards are discounted by 1/(1+KD), where D is the time between the firing of the current and the next production.
a	The learning rate.
D_{n+1}	The time between the consecutive firing of production n and n+1

Although the existing mechanism was able to match to human choice behavior, there are aspects in which the mechanism can be improved. First, in the existing mechanism, learning of P requires pre-specification of successful or failure states and the expected utility will increase or decrease respectively when the state is reached. The use of success and failure states may not be sufficient in situations where a continuous feedback function is required. From a practical perspective, pre-specification of success and failure states could be difficult especially in complex tasks, in which some states are often considered "more successful" than others. One way to improve the current mechanism is to learn from a scalar reward value. Being able to assign a scalar reward value to a production therefore allows more flexible pre-specification of the reward structure of the environment and allows the model to adapt to the environment accordingly. Second, the existing

procedural learning mechanism will change the expected utilities of productions only when the actual outcome is experienced, which requires keeping track of the whole sequence of previous productions that leads to the outcome. This could be computationally expensive especially when the number of productions is large. It is therefore desirable to have a mechanism that learns from local information before the outcomes are known.

The new procedural learning mechanism

In the artificial intelligence community, algorithms have been developed to allow agents to learn in different environments (Sutton & Barto, 1998). One established algorithm is the Temporal Difference (TD) algorithm, which was originally inspired by behavioral data on how animals learn prediction (Sutton & Barto, 1981). Research showed that the TD algorithm is well suited to explain the functional role of dopaminergic signals (e.g. Houk, et al., 1995; Holroyd & Coles, 2002, O'Reilly, 2003). The TD algorithm is designed to learn to estimate future rewards based on experience, and has a built-in credit assignment mechanism that reinforces the predicting stimuli.

In its simplest form, the new mechanism can be represented as $U'(n) = U(n) + aTD(n)$, where $U'(n)$ is the updated value of the expected utility $U(n)$ of production n after an ACT-R cycle, a is the learning rate, and $TD(n)$ is the temporal difference error. $TD(n)$ calculates the difference between the actual and expected rewards, i.e. $TD(n) = R(n) - U(n)$. The basic learning mechanism is therefore similar to the learning rule of Rescola and Wagner (1972) (e.g. see Sutton & Barto, 1981). The measure of future rewards has to take into account long-term as well as short-term consequences. It is plausible to weigh immediate primary reinforcement more strongly than delayed primary reinforcement. We chose to use the hyperbolic function to discount delayed reinforcement (the justification of using the hyperbolic function is beyond the scope of this paper, but see Lowenstein & Prelec, 1991; Mazur, 2001). A good estimate of the total future rewards is therefore $R(n) \approx r_n + U(n+1)/(1+KD_{n+1})$, where r_n is the immediate reward received for production n, $U(n+1)$ is the expected utility of the production that fires after production n, K is the discount parameter, and D_{n+1} is the time lag between the times when production n and production n+1 fire. To implement the mechanism in ACT-R, the basic algorithm has to be modified to take both the reward and cost into account and translate them into a single dimension[1] – i.e. the reinforcement will be the difference between the reward and cost (i.e. the net reward). In other words, the estimate becomes $R(n) \approx r_n - C_n + U(n+1)/(1+KD_{n+1})$, where C_n is the cost of firing production n. Putting the estimate of R(n) back to the equation for $U'(n)$, we have:

$$U'(n) = U(n) + a[r_n - C_n + U(n+1)/(1+KD_{n+1}) - U(n)]$$

One can see that when the estimate is perfectly accurate, $TD(n) = 0$, or $U(n) = r_n - C_n + U(n+1)/(1+KD_{n+1})$ and learning will stop. The value of TD(n) can therefore be considered the prediction error (as encoded by dopaminergic signals), and the mechanism learns by reducing this prediction error. It can easily be seen that once a primary reward is received, the expected utility of the productions that lead to the reward will be credited with a discounted reward, and discounting is heavier the farther away the production is from the reward.

The new mechanism updates the expected utility based on the difference between the predicted and actual net reward. There are two main differences between the new and existing mechanisms. In the new mechanism, the reward is a scalar quantity, and the amount of change is determined by the difference between the predicted and actual reward, which is consistent with the functional role of dopaminergic signals. This characteristic allows the new mechanism to extend its learning capabilities beyond a binary function as in the existing mechanism. Second, in the existing mechanism, learning requires keeping track of a long sequence of productions that lead to the reward. However, in the new mechanism, only the expected utility of the next production is required. The reinforcement signal will eventually propagate back to the productions that lead to the reward.

Testing the new mechanism

The goal of this paper is to show the limitations of the existing mechanism and how the new mechanism is able to extend the learning capabilities of ACT-R. However, owing to space limitation, we are unable to show all properties of mechanism. For example, none of the data sets in this paper was sensitive to the discount parameter K, so we fixed it at 1.0 and just varied the value of r_n to fit the data[2]. The learning rate a was also fixed at 0.1. We first used the new mechanism to fit two data sets from the probability learning tasks by Friedman et al. (1964) and Myers et al. (1963). Since these two sets of data were also modeled well by the existing mechanism (Lovett, 1998), we were able to compare the results of the two mechanisms and show that the use of TD error to drive the learning process is at least as effective as the existing mechanism. Finally, we used the new mechanism to fit the data from a decision making task studied by Busemeyer and Myung (1992), which we believe were problematic for the existing mechanism.

Probability matching behavior

In Friedman et al., participants completed more than 1,000 choice trials over the course of three days. For each trial, a signal light was illuminated, participants pressed one of the two buttons, and then one of the two outcome lights was

[1] ACT-R takes the agnostic economist's position of simply assuming these map onto some internal values without deeply inquiring why.

[2] Since the delay D is a constant for all data sets, it can be shown that the parameter K is absorbed into the value or r_n.

illuminated. Task instructions encouraged participants to try to guess the correct outcome for each trial. The study extended the standard probability learning paradigm by changing the two buttons' success probabilities across 48-trial blocks during the experiment. Specifically, for the odd-numbered blocks 1-17, the probabilities of success of the buttons (p and 1-p) were 0.5. For the even-numbered blocks 2-16, p took on the values from 0.1, to 0.9 in a random order. We focus on the analysis of the even-numbered blocks, as they show how people adapted to the outcomes with experience.

Table 2. Observed and predicted choice proportions from the experiment by Friedman et al. (1964). Predicted scores are in parentheses. Each block has 12 trials.

	Probabilities			
P	Block 1	Block 2	Block 3	Block 4
0.1	0.34 (0.37)	0.23 (0.24)	0.18 (0.17)	0.15 (0.13)
0.2	0.37 (0.41)	0.26 (0.26)	0.29 (0.23)	0.31 (0.23)
0.3	0.49 (0.49)	0.41 (0.41)	0.44 (0.34)	0.35 (0.33)
0.4	0.46 (0.53)	0.44 (0.50)	0.38 (0.43)	0.38 (0.38)
0.6	0.56 (0.59)	0.51 (0.59)	0.52 (0.55)	0.52 (0.57)
0.7	0.50 (0.56)	0.53 (0.64)	0.58 (0.72)	0.62 (0.75)
0.8	0.50 (0.51)	0.76 (0.71)	0.74 (0.77)	0.73 (0.78)
0.9	0.66 (0.62)	0.78 (0.79)	0.78 (0.81)	0.79 (0.81)

Table 2 shows the observed and predicted proportion of choices in the experiment by Friedman et al. Participants in general exhibited probability matching behavior. Across the four 12-trial subblock, participants chose the correct buttons in roughly 50% of the trials in the first block and approached the corresponding p values in each block. The predicted proportions were generated by the model, which had two critical productions, *Choose-Right-Button* and *Choose-Left-Button*, and the expected utilities of these productions were learned according to the new mechanism. The exact sequence of outcomes as reported in Friedman et al. was presented to the model. A reward of 3 is obtained when the correct button was chosen (i.e. r_n=3). The initial expected utilities of the two productions were set to 0. The fit was good, R^2 = 0.97, MSE = 0.003, which was similar to the model based on existing procedural learning mechanism. We conclude that the new mechanism can represent the learning mechanism at least as well as the existing mechanism with the same number of free parameters.

Overmatching behavior

Myers et al. performed another probability learning experiment, but they also varied the amount of monetary reward that participants would receive for each correct response. Participants would either receive no reward or penalty, ±1¢, or ±10¢ for each correct and incorrect responses. The probabilities that one of the alternatives was correct were p=0.6, p=0.7, and p=0.8. Table 3 shows the choice proportions for the participants in each of the conditions. When there was no reward, participants seemed to be exhibiting probability matching behavior. However, when there was a monetary reward, participants seemed to be "overmatching". From the data, it also appears that the higher the reward, the more the choice proportion exceeds the matching probability.

Table 3. Observed and predicted choice proportions from the experiment by Myers et al. (1963). Predicted scores are in parentheses.

Reward	Probabilities		
(cents)	p = 0.6	p = 0.7	p = 0.8
0	0.624 (0.612)	0.753 (0.750)	0.869 (0.829)
1	0.653 (0.676)	0.871 (0.834)	0.925 (0.938)
10	0.714 (0.711)	0.866 (0.836)	0.951 (0.944)

Since the task is basically the same as in Friedman et al., we used the same model to fit the data. We used the same set of parameters to fit the data in the no reward conditions (i.e. reward = 3). We chose the reward parameters (reward=±4.97 and ±5.7 for the ±1¢ and ±10¢ conditions respectively[3]) in the reward conditions to maximize the fit, and obtained R^2 of 0.98 and MSE of 0.0008, which is similar to the fit obtained by the existing procedural learning mechanism. However, we had only two free parameters in this model, compared to three free parameters in the model reported in Lovett (1998). In addition, the new mechanism provides a more natural interpretation of the overmatching behavior – when the reward was large, learning increases the expected utilities of the successful productions to higher values (since the deviation was larger). As a consequence, the model exhibited overmatching behavior. On the other hand, Lovett (1998) manipulated the architectural parameter G to fit the data, which seems awkward, as G is not supposed to be directly under strategic control.

Learning from normally distributed rewards

Busemeyer and Myung (1992) conducted an experiment in which participants were told to select one of the three treatment strategies for patients suffering from a common set of symptom patterns. Feedback on the effectiveness produced by the treatment was given after each selection. For the sake of convenience, the treatment with the highest expected effectiveness is called Treatment 3, and the next less effective treatment is called Treatment 2, and so on (see Figure 1). The effectiveness produced by each treatment was normally distributed with equal standard deviation, but the mean payoffs are different (as explained below). Participants had to evaluate each treatment based on trial-by-trial feedback. Participants were told to maximize the sum of the treatment effects over training and they were paid 4¢ per point. The means of the normal distributions are m-d, m and m+d for Treatment 1, 2, and 3 respectively. The two independent variables were mean difference (d) (i.e. the

[3] Since the reward values used in the model reflect subjective values, they do not necessarily follow a linear relationship with the external reward values.

separation of the distributions in Figure 1) and standard deviation (s) (which affects the amount of overlap in Figure 1). The exact values of d and s are shown in Table 4. Each participant was given 9 blocks (50 trials per block) of training in each condition. The model received the same amount of training as the participants.

From Table 4, we can see that as the mean difference increased, the observed choice proportions of the optimal treatment increased. As the standard deviation increased, the observed choice proportions of the best treatment decreased except when the mean difference was 3.0. The results showed that participants adapted their choice by learning the expected effectiveness of treatments. The results also showed that the more distinguishable the distributions were (larger mean difference or smaller standard deviation), the more likely the participants would choose the best treatment.

Table 4. Observed and predicted choice proportions of the optimal treatment from the experiment by Busemeyer & Myung (1992). Predicted scores are in parentheses.

Standard deviation (s)	Mean difference (d)		
	2.0	2.5	3.0
3.0	0.69 (0.74)	0.84 (0.79)	0.85 (0.84)
4.5	0.69 (0.72)	0.72 (0.76)	0.84 (0.80)
6.0	0.65 (0.68)	0.63 (0.69)	0.86 (0.83)

To model the data, we built three productions that chose each of the treatments. The initial expected utility of each production was set to 0. For each trial, the rewards obtained by the model were simulated by drawing a sample from the normal distribution that represents the effectiveness of the treatment chosen by the model. The value of r was chosen to be 1.76 to best fit the data. We obtained a fit of R^2=0.94, RMSE=0.007. The good fit to the data show that the new learning mechanism was able to build up the expected effectiveness of the treatments from trial-by-trial feedback, and was able to exhibit similar sensitivity to the differences of the distributions as participants. Since the effectiveness was sampled from a normal distribution, it is difficult to pre-specify which treatment was successful. It is therefore difficult to use the existing learning mechanism to model these data. In the new mechanism, however, whenever the actual reward was higher than the expected utility of the production, the production will be reinforced; otherwise the production will be penalized. With the same amount of experience (50 trials), the expected utilities of the production were able to reflect the actual expected effectiveness of the treatments.

Summary

We have fit a new procedural learning mechanism of ACT-R to three separate sets of data with all parameters held constant except the reward magnitudes the models received after each trial. In the first two cases, the new mechanism did at least as well as the existing mechanism in capturing the observed choice proportions in different settings. In the last case, we showed that the new mechanism fits data that are problematic for the existing mechanism. The new mechanism learned to probability match the true probabilities of outcomes by reducing the difference between the expected and actual reward. As the difference diminished, the change in the prediction decreased. When the reward was large, learning increases the expected utilities of the successful productions to higher values (since the deviation was larger). As a consequence, the chance of selecting the option that had the higher probability of being correct increased – i.e. the model exhibited overmatching behavior.

Although the first two sets of data can be modeled by the existing learning mechanism, the new mechanism provided a more natural explanation to the results. In the final set of data, we showed how the new mechanism generalizes and extends the computational abilities of the existing mechanism. The mechanism was able to learn the expected effectiveness of each treatment based on trial-by-trial feedback, without the need to pre-specify whether the productions had led to successful or failure states.

Discussion

We have presented a new procedural learning mechanism in ACT-R. The use of the deviation between the expected and actual reward values in the new learning mechanism is consistent with the current understanding of the functional role of VTA dopamine neurons in basal ganglia. We showed that the new mechanism generalizes and extends the computational abilities of the existing procedural learning mechanism. Specifically, the new mechanism is not limited to learning from binary feedback functions. Rather, the new mechanism is able learn from continuous reward functions with similar sensitivity to the variations in the reward distributions. The current paper also showed how the reinforcement learning mechanism observed in basal ganglia can be implemented in production systems such as ACT-R.

In practice, the current mechanism allows the use of a scalar reward parameter without the need to pre-specify success or failure states in a task. This pre-specification could be difficult especially in complex tasks in which a state could sometimes be good or bad depending on one's experience with the task, as experience may change one's expectation of different states. In addition, although the existing mechanism can adapt to different magnitudes of reward, the change of the architectural parameter G (in E=PG-C) to fit the data may not be easy in complex tasks that has many subgoals, especially when some subgoals may be considered "more successful" than the others.

Owing to space limitations, we are not able to show all properties of the mechanism. In fact, we have only tested the mechanism in single-choice tasks, which did not depend critically on the credit assignment mechanism. The discounting of future rewards therefore did not affect performance of the models in all three tasks that we have presented. However, we believe the discounting mechanism

is more plausible than the existing mechanism, in which immediate and future rewards are weighted equally.

In all three data sets, the model had the same amount of experiences as the participants and reached the same level of asymptotic performance. In the first data set, we also showed that the performance of the model in each of the four subblocks matched the participants well, suggesting that the learning rate of the mechanism is comparable to that of the participants. However, it is possible that the reinforcement learning mechanism could be slow for more complex tasks. It could be problematic, for example, when a primary reward is received after a long sequence production firings. Since only one production is updated during each ACT-R cycle, the primary reward may take several cycles to propagate back to the production where the critical decision is made. It is not clear how people learn in such situations. It is possible that they rely on direct instruction to point out such contingencies rather than counting on a automatic learning mechanism. It does not seem that the mechanisms behind the dopamine reward system are capable of spanning unbounded lengths of time in a way that would lead to rapid convergence.

Acknowledgments

The current work is supported by a grant from the office of naval research (N00014-99-1-0097). We thank Niels Taatgen, Pat Langley, and two anonymous reviewers for the useful comments on an earlier version of this paper.

References

Anderson, J. R. (1990). *Rules of the mind.* Mahwah, NJ:Erlbaum.

Anderson, J. R. & Lebiere, C. (1998). *Atomic components of thought.* Mahwah, NJ:Erlbaum.

Anderson, J. R., Qin, Y., Sohn, M.-H., Stenger, V. A., Carter, C. S. (2003). An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin & Review, 10* (2), 241-261.

Busemeyer, J. R. & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, *121* (2), 177-194.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747-749.

Friedman, M. P., Burke, C. J., Cole, M., Keller, L., Millward, R. B., & Estes, W. K. (1964). Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 250-316). Stanford, CA: Stanford University Press.

Holroyd, C. B. & Coles, M. G. H. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109* (4), 679-709.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233-248). Cambridge, MA: MIT Press.

Ljungberg, T., Apicella, P., Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of neurophysiology, 67*, 145-163.

Loewenstein, G. & Prelec, D. (1991). Negative time preference. *The American Economic Review, 81* (2), 347-352.

Lovett, M. C. (1998). Choice. In C. Lebiere and J. R. Anderson, *Atomic components of thought* (chapter 8, pp. 255-296).

Mazur, J. E. (2001). Hyperbolic value addition and general models of animal choice. *Psychological review, 108* (1), 96-112.

Mirenowicz, J., Schultz, W. (1994). Importance of unpredictedness for reward responses in primate dopamine neurons. *Journal of neurophysiology, 72*, 1024-1027.

Myers, J. L., Fort, J. G., Katz, L., & Suydam, M. M. (1963). Differential monetary gains and losses and event probability in a two-choice situation. *Journal of Experimental Psychology, 66*, 521-522.

O'Reilly, R. C. (2003). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Institute of Cognitive Science, University of Colorado, Boulder, Technical Report 03-03.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: variation in the effectiveness of reinforcement and non reinforcement. In Black, A. H. and Prokasy, W. F. (Eds.). *Classical Conditioning II: Current Research and Theory*. New York, Appleton Century Crofts.

Schultz, W., Apicella, P., Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of neuroscience, 13*, 900-913.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward., *Science, 275*, 1593.

Schultz, W., Romo, R., Ljungberg, T., Mirenowica, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233-248). Cambridge, MA:MIT Press.

Sutton, R. S. & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review, 88* (2), 135-170.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Qualitative and Quantitative Effects of Surprise: (Mis)estimates, Rationales, and Feedback-Induced Preference Changes While Considering Abortion

Jennifer Garcia de Osuna (jmgdo@berkeley.edu)[1]
Michael Ranney (ranney@cogsci.berkeley.edu)
Janek Nelson (jamin@socrates.berkeley.edu)
University of California, *Graduate School of Education, 4533 Tolman Hall, Berkeley, CA 94720

Abstract

The Numerically Driven Inferencing (NDI) paradigm, and one of its methods, EPIC (Estimate, Prefer, Incorporate, and Change), are used to study both one's estimates and the effects of numeric feedback on one's personal policies (herein, about abortion). Both quantitatively and qualitatively, 92 undergraduates offered estimates and preferences for the legal U.S. abortion rate, explaining and justifying them. After receiving the (usually, quite surprising) true rate as feedback, they provided another (typically changed) preference-and-rationale. Results show that people vastly underestimated the abortion rate, and largely advocated decreases in it—both pre- and post-feedback. Feedback caused most of those who initially wanted no change in the abortion rate both to abandon the status quo and change preference-justifications; after feedback, two thirds of these students preferred a rate decrease, while the rest preferred an *increase*. Although many researchers hold that belief revision and conceptual change are quite difficult to elicit, these and other results show dramatic effects of simple base rate feedback on policy evaluation. Our findings highlight the importance of having and using data when reasoning about society-engaging topics such as abortion rates. This experiment represents a new way to study numerically-based reasoning that includes the subjective natures of our personal beliefs and social lives.

Please answer this question: "As a percentage of the current U.S. population, what is its *legal* immigration rate?" Does a typical response of 10% (Ranney, Cheng, Garcia de Osuna & Nelson, 2001) sound right? The true value is about *thirty-fold* less—only 0.3%. Does (or ought) this datum alter your immigration preference—your personal policy—some? Common sense may suggest that beliefs, decisions, and rationales will (or should) change with new information, but literatures from science learning to attitude change (e.g., from evolution or inertia to executions or diversity; Ranney et al., 2001), suggest that people are often unmoved by new data. Classical economics even suggests that preferences are exogenous (e.g., from estimates; Lurie & Ranney, 2003).

The Theory of Explanatory Coherence and its models (e.g., ECHO) describe a set of principles that guide belief evaluation and revision. Two such principles are that we (a) weigh evidence more strongly than conjecture, and (b) accept propositions explained more parsimoniously (Ranney & Thagard, 1988; Read & Marcus-Newhall, 1993; Schank & Ranney, 1991; Thagard, 1989). True base rates, then, would seem to represent parsimonious evidence (relative to a host of instances or anecdotes) and thus be (1) weighted heavily in one's reasoning about an issue and (2) evaluated as quite acceptable. The present paper explores aspects of this general hypothesis about (especially surprising) minimalist interventions—for instance, that a single, germane, critical number may foster conceptual change.

Some studies have noted that learning related base rate values (seeds) affects one's estimates (e.g., about spatial judgments or populations; Brown & Siegler, 1996; Brown, 2002, etc.). While intuitions about real world quantities are often incorrect (Brown, 2002), exposure to base rates increases the accuracy of one's estimates on closely related topics, and the benefits of such exposure can have lasting effects even months later (Brown & Siegler, 1996). Little is known, though, about the effects of base rate queries and feedback on preference/policy formation and change, so we suggest three "ifs." 1) If intuitions about real world numbers are often flawed, then they are likely being used to create anomalous or skewed personal policies among people. 2) If feedback can correct these intuitions, such feedback might affect individuals' policies. 3) If people are generally biased toward evidence (and they are; e.g., Schank & Ranney, 1991), then giving them factual, numeric feedback—say, the U.S. abortion rate, our main example—should affect conceptions and interpretations about the abortion rate, and thus affect both personal policies on abortion and the explanations supplied when justifying their policies. Among other questions, we seek to answer the following: Can supplying factual, numeric information about abortion markedly change one's abortion policy? Does receiving the actual rate as feedback affect the *Points of View* (POVs) by which people reason about abortion? Most such POVs (see below), involve moral or ideological reasoning aspects; religion plays a role, as well. (Space constraints prohibit reviewing the vast abortion literature here, e.g., Bernas & Stein, 2001, and we seek to focus on more paradigm-relevant aspects, in any case.)

We explore and measure phenomena of these sorts using a novel paradigm, Numerically Driven Inferencing (NDI; Ranney et al, 2001), and one of NDI's central empirical methods: EPIC (Estimate, Preference, Incorporate, and Change; cf. Lurie & Ranney, 2003, who introduced PEIC and IC as complementary methods). Such analytic frames allow us to study both estimates of, and dynamically changing preferences about, base rates (e.g., Munnich, Ranney, Nelson, Garcia de Osuna & Brazil, 2003). NDI also represents an emerging coherentist framework in which numbers are the "tips of the iceberg" of a person's thinking about a network of (magnitude-relevant) evidential and hypothetical propositions. A prime aspect of NDI's paradigmatic novelty is in its elicitation of what people *prefer* a quantity to be; it is further unique in its analysis of

[1] The order of the first two authors is alphabetical. *For J. Nelson's address, substitute in "Department of Psychology, 3510 Tolman Hall."

how such individuals' *policies* (as base-rate relative preferences) evolve in the face of numeric information. Using EPIC to study abortion numeracy, we asked each individual to *Estimate* the current legal abortion rate—per one million live births—and then offer a *Preference* (and thus a *policy,* relative to one's estimate) for what each would want the current rate to be (had one the power to change the rate), and give reasons for both numbers. They then received feedback (the actual abortion rate), which they *Incorporated* into their knowledge of the abortion issue. Students then provided a second preference and explanation; by contrasting their former and new preferences, estimates, and reasons, we can note *Changes* in preferences and policies that resulted from the feedback. Essentially, then, EPIC has four queries, about a rate X that has a value Y: (1) What is X's value? (2) What *should* X's value be? (3) X's value is actually Y, so (4) *Now*, what should X's value be?

Method

Participants, Design, Materials, and Procedure

Psychology pool undergraduates (N = 92) participated, as part of their course requirement. (The "N" will often be somewhat fewer in our Results, due to occasional missing data points.) The experiment used a pre- and post-feedback repeated measure, within a 2X2 factorial between-group design, although the two independent variables (Ranney et al., 2001) are tangential to the present issues and so omitted here, due to space constraints. Responses included numeric (continuous) estimates and preferences, written explanations of the estimates and preferences, and Likert ratings about: (a) general preference about a rate change, (b) familiarity with the topic, and (c) how much one cared about the topic.

This paper examines only one topic from a set of 16 (usually less emotion-laden) randomized topics: the U.S. legal abortion rate (which we defined for students—and represents the vast majority of abortions). Each person was first asked to estimate the current rate—per one million live births—and to explain the bases of that estimate. Next, each was asked how low and high the true abortion rate would have to be to be surprising, and to rate the confidence that the rate would fall in one's "non-surprise interval." Students were then each asked to give a numeric preference for the abortion rate, had one the power to change it, and to explain the preference. Then, they rated, on a 5-point Likert scale, how familiar they were with the topic of abortion rates, and how much they cared about the topic (with both ratings in contrast to the average American). As another measure of rate preference, on a 1-5 scale, students were also asked whether they generally preferred (1) a big decrease, (2) a decrease, (3) neither an increase nor a decrease, (4) an increase, or (5) a big increase. After this, feedback was provided—the then-current abortion rate—an often-shocking 335,000 per million live births (gleaned from independent federal agencies, e.g., the CDC & NIH; nb. the rate has since dropped some). Each was then asked to consider the feedback and again give a numeric preference, and to explain that final preference. Finally, the students were again asked both to rate how much they cared about the topic and to offer a 5-point general preference rating. (Our lab has since replicated this abortion item's results, and has noted the effects both of varying how the rate is framed––e.g., with respect to a million *fertile women*, Munnich et al., 2003—and of omitting a numeric referent.)

Coding Scheme for Written Justifications

As part of NDI's methodology, the written justifications for abortion preferences (before and after feedback) were coded qualitatively using verbal analysis methods. We developed a 14-category coding scheme by extracting major patterns from elicited explanations, and then coded *all* justifications with the scheme. A student's explanation could fit up to three of the 14 coding categories, with many requiring more than one code. Inter-rater reliability for coding the reasons was 90% among three coders. For ease of discussion, the 14 categories were grouped into six broader categories called *Points of View* (POVs) by which people justified their personal abortion rate preferences. These POVs are: (1) preferring a utopian world in which abortions are essentially unnecessary or moot; (2) that the abortion rate should reflect the greater good for society; (3) that abortions should always be allowed/legal, regardless of circumstance; (4) that abortions should only be allowed in some circumstances; (5) that abortions should never be allowed (e.g., illegal) under any circumstances; and (6) other/no explanation.[2]

Results

Findings are reported following the EPIC procedure: Estimates and Preference (EP), then Incorporation (feedback) and Change (IC). We also focus analysis on (a) written, post-preference, justifications, (b) correlations between preferences and some Likert ratings, and (c), two especially interesting subsets of participants: those who wanted a rate of zero abortions, and those who notably changed the direction of their preferences after feedback. (Space limits do not permit us to report all, or even all statistically significant, findings; cf. Ranney et al., 2001.)

[2] POVs (1-6) map onto the *original* 14 (a-n) code categories as follows. POV-1: (a) "perfect world" in which *all* pregnancies are desired or yield loving adoptions, (b) "birth control," with perfect contraception preventing *all* unwanted pregnancies, and (c) "responsibility" by *full* abstinence or pregnancy-completion. (I.e., a-c respondents wish abortions need never be considered.) POV-2: (d) abortions needed to optimize social benefit (e.g., economics, improved life-quality for all, and crime reduction). (POVs 3-5 concern availability or legality.) POV-3: (e) "better for mother or unborn child," as some wish to reserve abortion as perhaps better for the mother (e.g., her health) and/or fetus's predicted life, (f) "basic 'pro-choice' position," as some didn't expand on being pro-choice, (g) "women's right," with which they may choose abortion, and (h) "status quo," such that the abortion rate ought not change. (In categories e-h above, rationales include notions that abortions ought *always* be available.) POV-4: (i) "not for contraception," by which some decried abortions-as-birth-control, and (j) "emergency only," with abortions only allowed in extreme cases (e.g., after rape or to save a mother's life). POV-5: (k) "murder/loss of life," in which the fetus's loss of life is deplored, and/or abortion is seen as murder, and (l) the "basic con position," as some didn't expand on being anti-abortion. (Codes k-l reflect being *fully* anti-abortion.) POV-6: (m) a comment/rationale that was not captured by prior codes (e.g., "Why are you asking me this?") or (n) no explanation.

Estimates and Initial Preference/Policy

Participants greatly underestimated the abortion rate. The median *estimate* was 5,000 abortions per million live births (M=50,479; S.D. = 148,469), much less than the true—and often evocative—rate of 335,000. (Due to high variance, we focus more on median estimates, as they usually inform more than do means.) In general, students' initial numeric *preferences* differed from what they thought the true rate to be (i.e., there was, overall, a significant difference between one's estimate and initial preference; t(88)=-2.62, p=.01). The median initial preference was only 100 abortions per million live births (M = 19,381; S.D. = 110,372), or 4,900 less than the median *estimate*—that is, a policy advocating a 98% (or fifty-fold) decrease. Thus, most people (62.2%) preferred rate decreases, relative to their estimates. Counter-intuitively, ratings of familiarity and caring about the topic did *not* significantly correlate with estimate accuracies, indicative of rather modest metacognition.

Recall that we also elicited initial *general* preference ratings (from 1-5) for the abortion rate. These ratings tended to favor decreasing the abortion rate (M= 2.07), and were negatively correlated with initial caring ratings (r(90)=-.32, p=.002). The initial numeric preferences of those who wanted a *general* decrease (a "1" or a "2" rating) differed significantly (F(1,86) = 4.33; p = .04) from preferences of those who did not. The median initial preference was zero for those initially wanting a (Likert scale) decrease in abortions (n=56, M= 1,337.32; S.D.= 3,577). The median initial preference was 2,000 abortions per million live births for those who initially preferred either (a) an increase, or (b) neither an increase or a decrease on the Likert scale (n=34, M= 50,001; S.D.=178,723). Before feedback, the majority (62%) chose either "prefer big decrease" or "prefer decrease" for the abortion rate, prior to feedback, mirroring the numeric preferences. Only two participants wanted an "increase," and none preferred a "big increase." The rest (35.6%) preferred "neither an increase nor a decrease." As expected, given the great tendency to underestimate, initial preferences were negatively correlated with later being surprised by the feedback (r(89)= -.25, p=.02).

Of the 67 explanations by those who initially preferred a decrease in abortion (relative to their estimate), the bulk of them fell into three POVs: 26 explanations held that abortions should never be allowed, 24 referred to a utopian world in which abortions are unnecessary or moot, and 10 asserted that abortions should only be allowed in some circumstances. For example, one decrease-preferring participant (with Estimate: 100,000; Initial Preference: 0) wrote, "It would be great if every baby was cherished enough to be allowed to live." Another person preferring a decrease (Est: 1,000; Init. Pref: 10) stated, "Because I don't believe women should end a child's life unless it affected their own physical health." People preferring a rate equal to their estimates most often indicated that abortions should always be allowed (17 of 29 explanations); the remaining justifications were from all other POVs except that for the "Other/no explanation" (see Table 1). For example, one of these status-quo students (Est: 800; Init. Pref: 800) stated, "[I prefer] as many as necessary to not have unwanted children. I believe people have the right to have an abortion if they cannot have the child for personal reasons." All three people whose preferences exceeded their estimates indicated that abortions ought always be allowed. For example, one (Est: 20,000; Init. Pref: 1,000,000) stated, "People should choose whether or not they can bring a kid to the world."

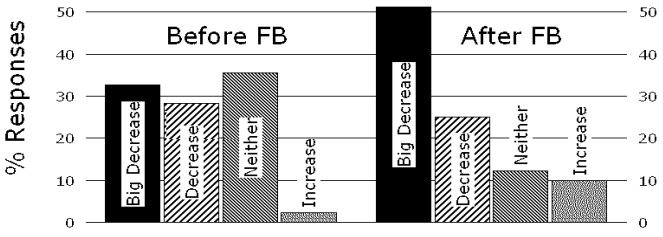


Figure 1: Distributions of general preference ratings (Likert; 1-4 out of 1-5), before and after feedback.

Incorporation (of Feedback) & Preference Change

After feedback, the median numeric preference *increased* to 1,000 abortions per one million live births (M = 108,178; S.D.= 174,688). However, that median is a larger 99.7% decrease-policy from the true abortion rate of 335,000. (Recall that the median initial preference called for a 98% decrease-policy in the rate, relative to their estimates.) Final numeric preferences still significantly differed from the feedback value (t(88)=-12.25, p<.001), and represented a non-proportionate shift in policy (p < .001 via a Wilcoxon Signed Rank test). Mirroring this policy shift toward a greater decrease (percentage-wise) in abortions, the mean Likert rating for general preference dropped from 2.1 (out of 5) to 1.8 after feedback (t(89)=3.39, p=.001). After feedback, 51.1% preferred a "big decrease," dramatically up from 32.6% (and 25.0% preferred a "decrease," down slightly from 28.3%). This movement is evident in Figure 1, which contrasts the distribution of initial Likert ratings for general preferences before and after feedback. Those who initially chose "neither an increase nor a decrease" most notably changed Likert ratings for general preferences; after feedback, 66% of them (21 of 32) "moved off the status-quo fence," with five coming to prefer a "big decrease," nine a "decrease," and the other seven dramatically diverging to prefer an "increase." In Figure 1, this scattering is seen in the shrinking of the "Neither" bar and the growth of both the "Big Decrease" bar and—more surprisingly—the "Increase" bar. (Note that no one in this study ever preferred a "Big Increase.") This striking bifurcation of most of the (initially) status-quo group is qualitatively analyzed below.

Mean care ratings (from 1 to 5: "not at all" to "much more than average"), increased significantly after feedback, from 3.29 to 3.51 (t(90)=-2.89, p=.005). Numeric and (Likert) general preference measures concurred, because after feedback, numeric preferences continued to differ between those wanting either a "decrease" or "big decrease" and those who did not (M= 1,337 and 50,001, respectively F(1,87) = 445.8; p<.001). Interestingly, both before and after feedback, "care" ratings were negatively correlated with general-preference Likert ratings (respectively, r(90)= -.32, p=.002; r(90)= -.21, p=.047). That is, those "caring" more about abortion preferred more of a decrease in its rate.

Table 1 shows the effect of feedback (i.e., pre- vs. post-) on the percentage of each POV mentioned, within the four

Table 1: Percent usage of POVs, from pre-feedback to post-feedback, by general/Likert preference rating.

	Big Decrease (1)		*Decrease (2)*		*Neither (3)*		*Increase (4)*	
Point of View (POV) Justification	Pre-	Post-	Pre-	Post-	Pre-	Post-	Pre-	Post-
Abortions should never be allowed	48.6	40.4	28.1	21.4	3.5	0	0	0
Utopian world: abortion as non- issue	40	31.6	31.3	50*	6.9	0	0	0
Allow abortions only in some circumstances	2.9	24.6*	28.1	25	3.5	0	0	0
(Other / No Explanation)	0	1.8	0	3.6	17.2	18.2	0	12.5
Abortion rate ought depend on the greater good	2.9	0	3.1	0	10.3	0	0	37.5*
Abortions should always be allowed	5.7	1.8	9.4	0	58.6	81.8*	100	50*
Totals (i.e., each column sums to 100%)	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*

levels of general preference Likert ratings that people used. Five notable distributional changes are marked by asterisks (*s). One such change (from 2.9% to 24.6*%) represents the finding that, before feedback, only *one* of 35 "big decrease" justifications was coded as "Abortions should be allowed only in some circumstances," but *14* of 57 "big decrease" justifications were so coded after feedback.

As a rather orthogonal analysis from those above, the following set of subsections examine three sets of participants and how they were differentially affected by feedback: those preferring zero abortions (both before and after feedback), those changing their basic position on the abortion rate, and the remaining participants.

Zero Preference Participants Of all students, 34.8% initially wanted zero abortions. After feedback, 84.2% of the 34.8% still preferred zero abortions. Initially, such people typically used a utopian world POV (50%), or a POV that abortions should never be allowed (41%). Feedback spurred only a non-significant drop in the use of a utopian world POV (42%), concomitant with a non-significant rise in the view that abortions should never be allowed (44%). Preferring zero abortions before feedback was correlated with initial caring ratings ($r(88)=.24$, $p=.025$); there was *no* correlation, though, between preferring zero abortions *after* feedback and *final* caring ratings ($r(88)=.06$; $p=.60$).

Participants Who Changed Away from "Status Quo" Of 89 respondents, 24% changed their policy direction—and *all* of these 21 were those who initially preferred neither an increase nor a decrease in abortions on the Likert scale, yet preferred either an increase or a decrease after feedback. We refer to these as "semi-flips," as no one *fully* flipped sides (e.g., from "increase" to "decrease" or vice versa). Of the "semi-flippers," 14 shifted to preferring a "decrease" (11 of whom were "technically surprised" by the feedback, in that 335,000 fell outside of their non-surprise intervals). Remarkably, given that almost everyone underestimated the rate (mostly by vast amounts), the other seven changed to wanting an *increase* in the abortion rate (e.g., by concluding that numerically unwarranted taboos due to media-skewed rate perceptions may be *inhibiting* abortions); indeed, six of the seven were technically surprised by the high feedback. For all semi-flippers, as per above, the initial (Likert) general preference was "3" (i.e., "neither an increase nor a decrease"). The post-feedback mean general preference for the semi-flippers dropped to 2.43, but this aggregates a drop to 1.64 for "decreasers-come-lately" (DCLs) and a rise to 4.00 for "*increasers*-come-lately" (ICLs).

Table 2 shows the POV distributions for the 21 semi-flippers, pre- and post-feedback, which significantly differed ($\chi^2(5,N=105)=11.88$, $p=.036$). Pre-feedback, most of their rationales indicated that abortions should *always* be allowed (12 out of the 21 people, with the other nine being grounded in all other POVs). Post-feedback, the POVs these students used depended *entirely* on whether one had semi-flipped to prefer an abortion-rate increase or a decrease. (Note: there were 22 instances, post-feedback, as one person used two POVs.) This feedback-driven bifurcation was total, as shown in Table 2's last two columns; note the complete lack of overlap, post-feedback, between the first two POVs (rows) and the next three (i.e., excluding "other/no explanation"). After feedback, four of the seven ICLs justified their preferences with "abortions should always be allowed," while the other three justified with "abortions should reflect the greater societal good." (One used both.) However, DCLs didn't use either of these two POVs to justify preferences: Instead, the greatest number of

Table 2: POV changes for semi-flip participants/POV-instances (pre- vs. post-feedback; 21 people vs. 22 POV instances), by general preference/Likert ratings (ICL = "increaser-come-lately;" DCL = "decreaser-come-lately").

	Pre-Feedback (out of 21 semi-flippers)	Post-Feedback (of 22 POVs)	
Point of View (POV) Justification (Note: the order has changed from Table 1)	Neither an Increase nor Decrease	Increase	Decrease or Big Decrease
Abortions should always be allowed	12 *(8 became DCLs; 4 became ICLs)*	**5**	0
Abortion rate ought depend on the greater good	2 *(both became ICLs)*	**3**	0
Utopian world where aborting a non-issue	2 *(both became DCLs)*	0	**8**
Abortions should never be allowed	1 *(who became a DCL)*	0	**3**
Allow abortions only in some circumstances	1 *(who became a DCL)*	0	**3**
(Other / No Explanation)	3 *(2 became DCLs; 1 became ICL)*	0	0

their post-feedback instances involved the utopian POV (eight of 14 instances); the remainder of these participants' explanations included the POVs that either abortions should never be allowed (three of 12 instances) or should be allowed only in some circumstances (three of 12 instances).

Thus, most semi-flippers initially wrote that abortions ought always be allowed, yet not one DCL wrote that belief after feedback. Further, semi-flippers were so polarized after feedback that there was no overlap at all between the POVs of the ICLS and the DCLs. For example, one initially-status-quo participant who first both estimated and preferred a rate of 20,000 (per million live births) wrote, "I think it is a good number." After feedback, the same person changed views, preferring an increase in the abortion rate to 500,000, stating: "I think there are too many kids being [born] into this country, especially since a lot...are being raised by teen/bad/druggie parents." In contrast, a semi-flipper who changed to prefer a *decrease* in the abortion rate post-feedback, at first both estimated and preferred a rate of 800, stating, "[I prefer] as many as necessary to not have unwanted children. I believe people have the right to have an abortion if they cannot have the child for personal reasons." Post-datum, this person wanted a rate of 200,000 (again, a decrease from the feedback's 335,000), stating, "A lot of these probably happen because women/men aren't taking the right precautions and with education or birth control. I think this number could start to decrease. But I do believe women have the right to abortions, but not the right to use abortions as a birth control method."

Remaining ("Non-Zero, Non-Semi-Flip") Participants
The distributional shift in POVs was also significant for the rest of the people (χ^2 (5, N=80)=139.24, p<.001)—those who preferred abortion rates above zero, but did not change the direction of their general preference (Likert) rating after feedback. Such people were neither semi-flippers nor those wanting zero abortions, and so represent a less extreme subgroup than those discussed above. It is instructive to note how the relative POV use changed for these intermediate "non-semi-flip/non-zero-preference" participants: Before feedback, the top three abortion POVs for these non-zero, non-semi-flip students were equally split (with 23.5% of instances apiece) among "allowed in some circumstances," "never allowed," and "always allowed." After feedback, though, the POV that abortions should be allowed in some circumstances represented 33.9% and the two absolute POVs that abortions should either *always* or *never* be allowed—made up 16.9% and 18.6% of the responses, respectively. The other POVs' percentages changed rather less (utopian world: 13.8→18.6; other/no explanation: 9.8→10.3; should depend on greater good: 5.9→1.7).

Discussion

Respondents largely estimated the legal U.S. abortion rate to be far lower than the true rate—seven times lower in mean, and 67 times lower in median. In fact, 79% of students were "technically surprised" by the feedback, such that the true abortion rate was not contained in their elicited non-surprise intervals. Thus, only 21% of our students captured the true value (335,000 legal abortions per million live births) in their non-surprise intervals—even though their mean confidence of doing so, just after offering their intervals, was 74%. Thus, participants were roughly 3.5 times less likely to capture the true rate than they anticipated.

The effect of the feedback on one's preference was likely due, in part, to its shocking magnitude. Results show that learning the true abortion rate clearly changed reasoning about abortions—with regard both to peoples' preferences and the points of view (POVs) by which they justified their policies. While the median person went from preferring 100 abortions per million live births to preferring 1,000, since the median estimate was 5,000 and the feedback was 335,000, students' new preferences represented a much more stringent relative abortion policy (from -98% to -99.7%). Indeed, the failure to capture the feedback value in one's non-surprise interval was significantly correlated with exhibiting a dramatic (i.e., non-proportionate) change in abortion policy (r(54)=.4; p<.01).

Overall relative policy preferences also became more fervently abortion-reducing (constrictive) on other metrics post-feedback. There was a significant overall drop in general Likert preference ratings for the abortion rate, after feedback. Indeed, almost 20% more of the students preferred a "big decrease" after learning the true rate (see Figure 1). The justifications also changed: Before feedback, 25.2% of the full set of explanations provided that abortions should always be allowed, while only 11.1% stated that abortions should be allowed in only some circumstances. After feedback, these frequencies were essentially reversed (13.5% vs. 20.2%). The only people who showed little shift in either their numeric preference or written justifications were those who preferred zero abortions—before and/or after feedback. Considering that these zero-preferring people (a strange-bedfellows group seeming to include utopian liberals and abolitionist conservatives; see Results) essentially held absolutist policies (for zero abortions), it is not surprising that the true rate did little to change that wish.

In both more quantum and qualitative senses, one of the most dramatic of the above results is that those who initially adopted a "status quo" policy usually changed their positions after seeing the generally surprising base rate feedback. Of the 32 respondents who first preferred neither an increase nor a decrease in the abortion rate, feedback caused 21 of them (66%) to take a directional position, thus becoming "semi-flip" participants. Base rate feedback also changed the POVs by which semi-flip people justified their preferences. While before feedback, most such students claimed that abortions should always be allowed, regardless of circumstances, these justifications shifted dramatically after feedback. The participants who changed to preferring decreasing the abortion rate no longer claimed that abortions should always be allowed, and instead largely justified their new preferences by preferring a utopian world in which abortions need never be considered, because either (a) all pregnancies would be wanted (or all unwanted pregnancies prevented), or (b) all unwanted babies would be readily adopted by loving homes. (This is consistent with "Wow! That's too many!" reactions.) Conversely, those semi-flippers who changed to prefer an *increase* after feedback used none of the justifications eventually used by those who

changed to prefer a decrease, and vice versa. Instead, these increasers-come-lately largely continued to claim that abortions should always be allowed, regardless of circumstances, or that the abortion rate should reflect the greater good for society. So, while semi-flip participants seemed rather homogenously like-minded before feedback, when it caused them to bifurcate into divergent positions (to prefer an increase or a decrease) there was no overlap in the types of justifications used. This is a dramatic effect, considering that the intervention is a single, albeit highly reliable, piece of information (i.e., 335,000:1,000,000).

Some Implications and Extensions

Our results show some of the effects of numerical feedback on personal preferences and policies regarding topics such as abortion—that is, topics for which the feedback is often surprising and quite far from individuals' estimates. More recent work from our laboratory observes this phenomenon in many realms, involving dozens of items and their rates—about incomes, inflation, executions, home ownership, etc.—and even SAT percentile use in college admissions. Lurie and Ranney (2003) are extending this work even further, into the arena of health-care research funding, and have proposed a general model that relates numeric estimates, preferences, feedback, and seeds.

One implication of this work is the need to improve citizens' thinking about critical base rates. For many of the topics our Reasoning With Numbers group employs, many people are clearly quite unaware of crucial numbers related to an issue (Ranney, et al., 2001), and they even have low metacognitive knowledge-awareness (e.g., no significant correlation, for abortion, between familiarity and accuracy). Therefore, our lab has carried out a variety of promising classroom-based experiments, from grades 5-12, to foster such metacognition (e.g., Munnich, Ranney, & Appel, 2004). Among other goals, our curricula are meant to improve students' abilities to (a) estimate (e.g., by disconfirming sub-par, early, estimate-hypotheses and bringing more knowledge and accountability to bear), (b) prefer (or justify what they prefer, e.g., by reflecting on more dimensions influencing one's wishes), (c) utilize feedback (e.g., by "letting go" of one's estimate), and (d) triangulate or "N-angulate" (e.g., by seeking relevant external information sources). Curricular assessments show broad numerical reasoning gains over control students (e.g., in estimation; Munnich, Ranney, & Appel, 2004).

Most people use information other than statistics (e.g., ethics, pragmatics, and conventions) to form their positions, and even use *misconceptions* of true statistics, at times. Still, the actual numbers, especially when surprising, can significantly affect how they think about the issue. Thus, NDI results contrast with views of scholars in diverse fields who suggest that learning, transfer, belief revision, and attitudinal (or conceptual) change are quite difficult to foster (Munnich, Ranney, & Appel, 2004; Munnich et al, 2003; Ranney et al., 2001). Incorporating numeric feedback—even just a single, critical, number—can often dramatically change how one views an issue and reasons about one's positions. Further, such changes clearly transcend the domain of numbers, in that there is much non-numeric reasoning that lies beneath the iceberg-tips of estimation and quantitative preference. NDI's methods (e.g., EPIC) represent new tools with which science may better probe the submerged prominences and embedded fissures of thought.

Acknowledgments

We thank Laura Germine, Geoff Saxe, Franz Cheng, Ed Munnich, Nick Lurie, Patti Schank, Christine Diehl, Steve Adams, Michelle Million, Dan Appel, Luke Rinne, Sujata Ganpule, David Levine, Barbara Ditman, Tracy Craig, Tae Kim, Danny Kahneman, and our UC-Berkeley Reasoning Group, among others, for comments.

References

Bernas, R., & Stein, N. (2001). Changing stances on abortion during case-based reasoning tasks: Who changes and under what conditions. *Discourse Processes, 32*, 177-190.

Brown, N.R. (2002). Real-world estimation: Estimation modes and seeding effects. In B.H. Ross (Ed.) *Psychology of Learning and Motivation*, *41*, New York: Academic.

Brown, N.R. & Siegler, R.S. (1996). Long-term benefits of seeding the knowledge-base. *Psychonomic Bulletin & Review, 3*, 385-388.

Lurie, N., & Ranney, M. (2003, November). *Estimates, Preferences, and Preference Change: Biasing, Debiasing, and Seeding Effects in Thinking About Base Rates*. Paper presented at the annual meeting of the Society for Judgment and Decision Making, Vancouver, Canada.

Munnich, E.L., Ranney, M.A., Appel, D.M. (2004). Numerically-Driven Inferencing in instruction: The relatively broad transfer of estimation skills. To appear in *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Munnich, E.L., Ranney, M.A., Nelson, J.M., Garcia de Osuna, J.M., & Brazil, N.B. (2003). Policy shift through Numerically-Driven Inferencing: An EPIC experiment about when base rates matter. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society* (pp. 834-839). Mahwah, NJ: Erlbaum.

Ranney, M., Cheng, F., Garcia de Osuna, J., & Nelson, J. (2001, November). *Numerically Driven Inferencing: A new paradigm for examining judgments, decisions, and policies involving base rates*. Paper presented at the annual meeting of the Society for Judgment and Decision Making, Orlando, FL.

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426-432). Hillsdale, NJ: Erlbaum.

Read, S.J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65*, 429-447.

Schank, P. & Ranney, M. (1991). The psychological fidelity of ECHO: Modeling an experimental study of explanatory coherence. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp.892-897). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thagard, P.(1989). Explanatory coherence. *Behavioral & Brain Sciences, 12*(3), 435-502.

Does the Viewpoint Deviation Effect Diminish if Canonical Viewpoints are used for the Presentation of Dynamic Sequences?

Bärbel Garsoffky (b.garsoffky@iwm-kmrc.de)
Knowledge Media Research Center (IWM-KMRC)
Tübingen, Germany

Stephan Schwan (Stephan.Schwan@jku.at)
Johannes Keppler University
Linz, Austria

Friedrich W. Hesse (f.hesse@iwm-kmrc.de)
Knowledge Media Research Center (IWM-KMRC)
Tübingen, Germany

Abstract

Two studies examine the visual presentation of dynamic sequences. Experiment 1 tests if there are canonical viewpoints, that are especially appropriate for presentation. Participants agreed that viewpoints with 90 degree deviation between axis of sight and axis of main movement in the sequence are better than other viewpoints. Experiment 2 examines if these canonical viewpoints weaken the perspective deviation effect in a recognition task according to their postulated information richness. A perspective deviation effect was found both for canonical and less canonical views, even if it was weaker for the canonical views.

Viewpoint Deviation and Canonicality

This paper deals with questions concerning the cognitive representation of visually presented dynamic sequences, specially the role of viewpoint. A first experimental series (Garsoffky, Schwan & Hesse, 2002) showed that the viewpoint from which one sees a dynamic sequence becomes part of the cognitive representation of that sequence and therefore influences later memory retrieval processes. This viewpoint deviation effect appeared in three experiments examining recognition memory for visually presented dynamic sequences (Garsoffky et al., 2002) and comprises the stable result, that cuttings from sequences are best recognized if they are presented in a viewpoint most similar to the viewpoint from which participants before saw the whole sequence. This means the cognitive representation of dynamic sequences is not uncoupled from the viewpoint from which one primarily saw the specific sequence and therefore influences later memory retrieval processes. The question now is, if this viewpoint deviation effect holds for all kinds of viewpoints or if the use of special viewpoints may reduce this effect. The following studies therefore ask, if various viewpoints differ in their qualification to present a sequence – i.e. if there exist so called canonical viewpoints that could by now only be shown for static objects (e.g. Palmer, Rosch & Chase, 1981) (Experiment 1), and further it will be investigated, if these canonical viewpoints have an influence on the viewpoint deviation effect found by Garsoffky et al. (2002) (Experiment 2).

The concept of canonicality in connection with visual viewpoints was firstly empirically investigated and defined by Palmer et al. (1981). They discuss the idea of canonical viewpoints from an information-processing approach, a categorization perspective, in terms of phenomenology, and with regard to the concept of affordances (Gibson, 1982) and they conclude, that canonical viewpoints compared to other viewpoints contain more information as well as information of high salience, are the most typical viewpoints of an object, are those viewpoints from which an object is most perceivable, and are especially qualified to present the affordance structure of an object.

Empirically canonical viewpoints are defined e.g. by asking participants to imagine an object and then to describe the viewpoint from which the imagination took place, or participants were asked from which viewpoint they would make a photo of an object, or participants had to choose between photos with varying viewpoints which photo in their opinion presented the object best (Blanz, Tarr & Bülthoff, 1999; Palmer et al., 1981). Evidence for canonical viewpoints is stated if there is high inter- and intraindividual agreement.

At least for static objects some conclusions about the nature of canonicality can be made that do not mutual exclude each other. (i) Functionality and familiarity: Especially objects of everyday life we often see from a specific viewpoint that corresponds with the functionality of that object, i.e. when interacting with that object we see the object from a specific, i.e. canonical viewpoint that allows optimal interaction (Blanz et al., 1999). (ii) Information richness: In some studies canonical viewpoints were discovered even for abstract or nonsense objects – a fact that can not be explained by familiarity or functionality (Cutzu & Edelman, 1994; Edelman & Bülthoff, 1992; Perrett &

Harries, 1988). It was concluded that canonical viewpoints present more information and especially more salient information of an object than other, less canonical viewpoints. They present a high number of visible surfaces of an object, important parts of an object are not covered, and they are stable against small variations of the viewpoint, i.e. the informational advances of that viewpoint remain the same even if the viewpoint is changed slightly (for a comprehensive list see Blanz et al., 1999). (iii) Discriminability: Cutzu and Edelman (1994) concluded from their findings using abstract objects, that because of limited cognitive capacity for every specific object only the diagnostically valuable attributes are stored, that help to distinguish this object from other objects. This means that it varies with changing contexts or tasks, which viewpoint is more canonical than other viewpoints.

This paper investigates if the usage of such high informative viewpoints, that allow optimal discrimination, leads to a more viewpoint independent cognitive representation of visually presented dynamic sequences. I.e. the question is if so called canonical viewpoints help to recognize sequences better even if they are presented from new viewpoints.

Experiment 1

In the first study it has to be determined if there exist canonical viewpoints not only for static objects but also for dynamic sequences and how they can be defined. The study picks up one classical way to examine the canonicality of different visual viewpoints (Palmer et al., 1981) – namely rating measures, i.e. participants judge the goodness of various viewpoints for presenting the dynamic sequence.

For dynamic sequences, there often is one main direction of movement, and it is supposed, that viewpoints are the more appropriate to present the sequence the more they allow the observer to understand this movement. Moreover it is assumed, that viewpoints deliver the more information according to this movement the more orthogonal they are to the main movement direction of the dynamic sequence. This is argued because viewpoints with the axis of sight parallel to the main movement direction cause perspective shortenings. So it is hypothesized, that viewers prefer viewpoints that are as much as possible orthogonal to the direction of the main movement and that viewers rate viewpoints worse if these viewpoints are more parallel to the main movement direction.

Method

Participants Six male and ten female students, from the University of Tübingen participated in this experiment. They were paid for their participation.

Apparatus Experimental procedures were controlled by a Microsoft computer and realized by a html-program. Film clips were presented on a black background in the left and the right half of a color monitor.

Stimulus materials and design Eleven dynamic sequences were programmed using xyZET (Härtel & Lüdke, 2000), a simulation program to teach physics in school. Each dynamic sequence consisted of four spheres (balls) with different colors, sizes, starting positions and velocities. All balls moved on parallel laps towards a kind of blue goal at one end of the rectangular space. So the sequences were similar to a kind of race, with the exception that the balls did not start at the same line and that not all of them reached the goal within the duration of the sequence. Each sequence was filmed from 5 different viewpoints: All viewpoints had the same height but differed according to the horizontal amount of deviation between the axis of sight realized by the camera perspective and the axis of movement direction of the balls; this amount of deviation could be either 0˚ (i.e. parallelism), 22.5˚, 45˚, 67.5˚ or 90˚ (i.e. orthogonality); see Figure 1. This resulted in 55 film clips (11 sequences by 5 viewpoints); three of these sequences, i.e. 15 film clips, were used for training, eight sequences, i.e. 40 film clips, in the experimental test. This variation resulted in a design with the variable "canonicality of presentation" (0˚ / 22.5˚ / 45˚ / 67.5˚ / 90˚; within-subjects).

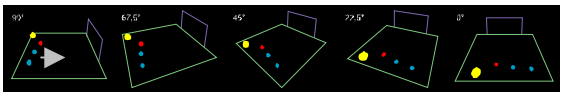


Figure 1: Visual presentations of the ball races used in experiment 1 with 0˚/22.5˚/45˚/67.5˚/90˚ amount of deviation between axis of ball movement (indicated by the little grey arrow) and axis of sight.

Procedure Participants were tested individually. The written instructions to the main part of the experiment – namely the rating of the film clip presentations at the computer - told them that they would see various dynamic sequences and that their task would be to rate the goodness of the viewpoint to present the dynamic sequence. It was explained, that the sequences were similar to races and that the relative positions of the objects to one another therefore is important when presenting the sequence. Further it was explained, that they always would see two film clips in succession presenting the same sequence but with different viewpoints and that they afterwards always should rate, which of the two film clips was the better presentation of that dynamic sequence by clicking with the mouse on one of two buttons (a button underneath the window of the film clip presented in the left half of the screen if they preferred this viewpoint or a button underneath the window of the film clip presented in the right half of the screen if they preferred that viewpoint). If participants had no more questions they were seated in front of the computer and a training phase started, which introduced some examples of the pair comparison task and trained to use the buttons. Data of the training phase were not analyzed.

When the film clip on the left side of the screen ended the last picture stayed in the window; after a short delay of one second the second film clip started on the right

side of the screen. At the end of the film clip also the last picture stayed on the screen. Now the participant had to made his or her rating by clicking one of two buttons. Then a short text note followed on the screen, inviting participants to click another button when they wanted to start the presentation of the next two film clips.

The order of the dynamic sequences was randomized as well as the order of viewpoint combinations. For each dynamic sequence there were 5 different viewpoints and each viewpoint was paired two times with each other viewpoint of this sequence: one time presented on the left half of the screen (i.e. the film clip they saw firstly) and one time presented on the right half of the screen (i.e. the film clip they saw secondly). So subjects in the experimental phase saw eight sequences, each in form of 20 pairs of film clips. All together subjects had to make 160 pair decisions; each specific viewpoint could reach a maximum of 64 preferences.

Results

Ratings For each participant it was counted, how often he or she favored a certain viewpoint compared to another viewpoint (number of "preferences"). An ANOVA with repeated measurement was performed, including the variable "canonicality of presentation" (0˚, 22.5˚, 45˚, 67.5˚ or 90˚; within-subjects). There was a significant main effect for this variable, F (4, 60) = 8.457, MSE = 227.802, $p < .01$ and also a significant linear contrast, F (1, 15) = 17.512, MSE = 397.233, $p < .05$. The viewpoints were rated better, the more the axis of camera sight deviated from the axis of sequence movement (0˚: 21.375 preferences; 22.5˚: 25.688 preferences; 45˚: 27.688 preferences; 67.5˚: 36.125 preferences; 90˚: 49.125 preferences). Single comparisons according to Scheffé revealed significant differences between 90˚ on the one hand and 0˚, 22.5˚ or 45˚ on the other hand.

Discussion

The results allow three statements: (i) In presentations of dynamic sequences specific viewpoints are preferred against other viewpoints. That means there are "canonical viewpoints" not only for static object presentations but also for sequences presenting dynamic movement. (ii) More than that it can be defined, which viewpoints are preferred, namely as predicted those viewpoints whose axes of sight are orthogonal to the axis of main movement in the sequences. (iii) Further it could be shown that viewpoints are rated worse the more they differ from this best, i.e. canonical viewpoint. The significant linear trend is a hint that canonicality – measured by preference judgements - is not an all-or-none concept.

Experiment 2

Experiment 2 now examines if there is a relation between the canonicality of viewpoints found in experiment 1 and the viewpoint deviation effect (Garsoffky et al., 2002). A recognition task is used and it is investigated if a thoughtful choice of viewpoint when presenting a dynamic sequence the first time (i.e. canonical viewpoints in the learning phase) can lower the effect of viewpoint deviation during recognizing cutouts of dynamic sequences (in the test phase). The rational for this question is the idea, that canonical viewpoints are information richer (Blanz et al., 1999; Cutzu & Edelman, 1994; Palmer et al., 1981), and that therefore it should be easier to recognize cutouts from a sequence even from deviating viewpoints, because one has more information about the sequence.

Method

Participants Eight male and twelve female, from the University of Linz, Austria participated in this experiment. Because the task was very difficult we tried to motivate the participants by informing them that the best three participants receive a gift coupon for a local cinema.

Apparatus The experimental procedures were controlled by an Apple computer (Power Macintosh 8100/80AV) and programmed using PsyScope (Cohen, MacWhinney, Flatt & Provost, 1993). Film clips and video stills were presented on a black background in the middle of a color monitor. Reaction times were measured by the computer internal clock, thereby resulting in an unsystematic measurement inaccuracy of 17 msec.

Stimulus materials and design Sixteen dynamic sequences were programmed, now using 3D canvas (amabilis.com) because this software offered more different colors than the simulation software used in experiment 1. Each of the sequences consisted of four balls with different colors, that moved on a rectangular plane in a linear parallel manner. The balls either moved towards a kind of goal or away from this goal and had different starting points. Further the balls moved with different and individually varying speed, i.e. they accelerated and decelerated – so again some kind of "races" resulted. Acceleration and deceleration was necessary to prevent that viewers could predict the end of the race after seeing only the first parts of the sequences by simply extrapolating the starting speed and position of each ball. Each sequence was filmed with a desktop camera from two different viewpoints (all camera viewpoints were 20˚ above the horizontal plane) – with 90˚ deviation between axis of ball movement and axis of camera sight (hypothesized to be the more canonical viewpoint) and with 0˚ deviation between the two axes (hypothesized to be the less canonical viewpoint). In the 90˚ condition the balls moved in 50% of the cases from the left to the right side and in 50% of the cases from the right to the left side to preclude that the 90˚ viewpoint simply is better, because it realizes the familiar reading direction. Accordingly in the 0˚ condition, the balls moved in 50% of the cases towards the observer and in 50% of the cases away from the observer (see Figures 2 and 3).

For each sequence 5 points of time that were evenly distributed throughout the sequences were defined, to get enough measurement possibilities. For each of these

points of time video stills for the recognition test phase were produced. These video stills had varying viewpoints on the sequence: All viewpoints had the same camera height (20˚ above the horizontal plane) but differed according to their horizontal deviation compared to the viewpoints in the film clips. This deviation could be 0˚, 45˚ or 135˚. Keep in mind that 0˚ does mean two different things for the film clips (learning phase) and the video stills (test phase): We speak of film clips with a 0˚ viewpoint, if there is no deviation between axis of ball movement and axis of sight. In contrast a video still with 0˚ is a video still the viewpoint of which does not deviate from the formerly presented viewpoint in the film clip – may this be a viewpoint with 0˚ or 90˚ deviation between axis of sight and axis of ball movement.

These variations resulted in a design with the variables "canonicality" (high / low; within-subjects), and "viewpoint deviation" (0˚ / 45˚ / 135˚ deviation between the viewpoints in the learning and the test phase; within-subjects).

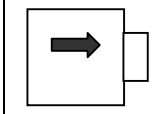
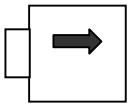
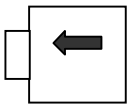
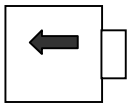
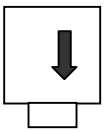
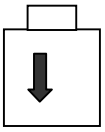
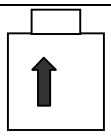
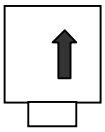
	balls moving towards the goal	balls moving away from the goal
90˚; from left to right		
90˚; from right to left		
0˚; towards observer		
0˚; away from observer		

Figure 2: Movement directions of the balls in experiment 2 towards / away from the goal, from the left to the right / from the right to the left, towards / away from the viewer, and with 0˚/ 90˚ deviation between axis of ball movement and axis of sight.

Procedure Again all participants were tested individually and received written instructions to the main part of the experiment – namely a description of the kind of dynamic sequences and their recognition task. First they passed through a training phase, the data of which were not analyzed. The experimental phase encompassed 8 races, i.e. 8 blocks. Each block consisted of an initial learning phase followed by a test phase. In the learning phase participants saw a dynamic sequence twice from either a canonical (90˚) or a less canonical (0˚) viewpoint, i.e. they saw the same film clip two times in succession from the same viewpoint. One second later, they successively saw 15 video stills (five points of time of the sequence each presented from three different viewpoints) as well as 15 distractor video stills which used the same viewpoints but presented other sequences, i.e. the sequences showed the same balls (same colors) but the video stills stemmed from other races with the balls moving with other speeds. So to perform the recognition task participants had to decide, if a video still showed a moment of the race seen before in the film or another race by checking the relative positions of the balls to each other. The order of the video stills was randomized. Each video still stayed on the screen until the participant pressed one of two reaction keys (one marked with "j" for the german word "ja" which means "yes", and one marked with "n" for the german word "nein" which means "no"). After the participant had reacted to a video still there always was a short delay of one second before the next video still was presented. The order of blocks (i.e. the different sequences) was randomized and each sequence was presented in the learning phase to half of the participants from a canonical viewpoint and to the other half of participants from a less canonical viewpoint.

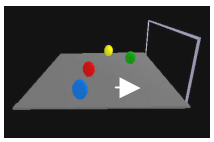
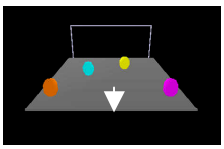
a)  b) 

Figure 3: Example pictures of the film clips used in experiment 2, arrows indicating direction of ball movement. a) shows balls moving towards the goal from the left to the right with 90˚ deviation between axis of ball movement and axis of sight. b) shows balls moving away from the goal towards the observer with 0˚ deviation between the two axes.

Results

Recognition accuracy For each participant his or her number of "hits" (the number of video stills correctly recognized as showing a moment from the ball sequence which he or she had previously seen) was determined. Across all participants and conditions a mean of 67.3 hits% resulted. Then an ANOVA with repeated measurement was performed, including the variables "canonicality" (high vs. low; within subjects), and "viewpoint deviation" (0˚, 45˚ or 135˚; within subjects). A significant main effect for "viewpoint deviation" was found ($F(2, 38) = 32.646$, MSE = 0,006571, $p < .01$) with 72.9 hits% at 0˚ viewpoint deviation between learning and test phase, 69.9 hits% at 45˚ viewpoint deviation and 59 hits% at 135˚ viewpoint deviation. Single comparisons according to Scheffé revealed significant differences between 0˚ and 135˚ viewpoint deviation as well as between 45˚ and 135˚ viewpoint deviation ($p < .01$). Accordingly, there was a significant linear effect of viewpoint deviation ($F(1, 19) = 52.432$,

MSE = 0.007386, p < .01), also indicating that recognition accuracy becomes worse the more the viewpoint used in the test phase differs from the initially presented viewpoint in the learning phase. In addition, the interaction (see Figure 4) between "viewpoint deviation" and "canonicality" became significant (F (2, 38) = 9.364, MSE = 0.004185, p < .01). Single comparisons revealed significant differences between high and low canonical viewpoints in the learning phase only if there was 0˚ viewpoint deviation (p < .01); further if the viewpoint in the learning phase was low canonical, there were significant differences between 0˚ (77.8 hits%) and 45˚ (71.5 hits%) viewpoint deviation (p < .05) as well as between 45˚ and 135˚ (57.7 hits%) (p < .01); but if the viewpoint in the learning phase was high canonical, there was no significant difference between 0˚ (68 hits%) and 45˚ (68.3 hits%) deviation, but only between 45˚ and 135˚ (60.3 hits%) deviation (p < .01). There were no more significant effects in this analysis of variance.

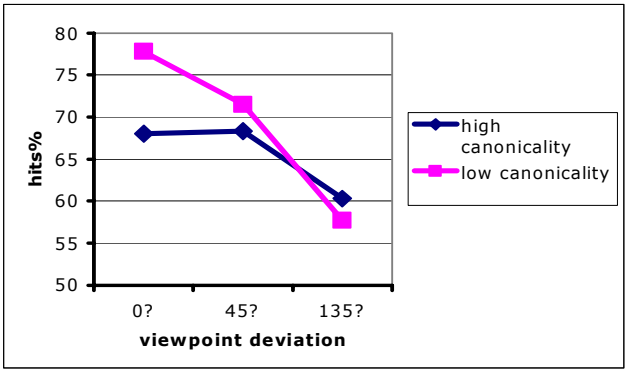


Figure 4: In experiment 2 for recognition accuracy there was a significant interaction between „viewpoint deviation“ and „canonicality“.

Recognition speed As a second dependent variable, reaction time was measured, i.e. the lapse of time from the beginning of each video still presentation until the participant pressed either the "j"- or the "n"-button. The following analysis only accounted for reaction times (RTs) to "hits" (i.e. correct "j"-reactions). Extreme RTs above 10 sec (i.e. more than 3 standard deviations above the overall mean) were excluded. This resulted in an exclusion of 1,16% of all RTs. To exclude outliers from analysis is a common method when dealing with reaction times (e.g. Cameron & Frieske, 1994; Diwadkar & McNamara, 1997; Eley, 1982; Hamm & McMullen, 1998; Lawson & Humphreys, 1996) because extremely slow responses indicate lapses of a participant´s attention on a particular trial. As the distribution of RTs was positively distorted, data were transformed by using natural logarithm, and analyzed in an ANOVA with the variables "canonicality" (high vs. low; within subjects), and "viewpoint deviation" (0°, 45° or 135°; within-subjects). For better vividness, the means reported in the text and figures are nontransformed RTs, despite the fact that the analysis of variance as well as the single comparisons were conducted using ln-transformed data. A significant effect was found for "viewpoint deviation" (F (2,38) = 12.194, MSE = 0.01284, p < .01). Applying single comparisons there were significant differences between 0° (2685 ms) and 45° (2944 ms) deviation (p < .01) as well as between 0° and 135° (3041 ms) deviation (p < .01). Accordingly also the linear trend for viewpoint deviation became significant (F (1, 19) = 16.387, MSE = 0.0179, p < .05). There were no more significant effects in this analysis of variance.

Discussion

In first line the results show once more (Garsoffky et al., 2002) a clear effect of viewpoint deviation: Recognition becomes worse the more the viewpoint from which one sees a cutout differs from the viewpoint from which one initially saw the sequence. This holds for recognition accuracy as well as for speed of recognition (see the two significant linear effects of viewpoint deviation). But the hypothesis that high canonical views in the learning phase weaken this viewpoint deviation effect receives only little support: On the one hand there is a significant interaction in recognition accuracy between canonicality and viewpoint deviation which shows that at least between 0° and 45° viewpoint deviation recognition accuracy does not become worse if in the learning phase a high canonical viewpoint is used. But on the other hand the use of a high canonical viewpoint in the learning phase does not weaken the viewpoint deviation effect between 45° and 135°. And at last there is no significant influence of canonicality on the viewpoint deviation effect for speed of recognition.

General Discussion

The results in the first place again support the stability and robustness of the viewpoint deviation effect for dynamic sequences (experiment 2): We used viewpoints that before (experiment 1) were rated as especially qualified to present critical aspects of visual dynamic sequences, namely the relative positions of the various balls to each other. I.e. these viewpoints were rated as being especially informative for this kind of dynamic event and therefore should allow to store in memory a maximum of discriminative information about the event. We hypothesized that if observers see dynamic sequences initially from these information rich "canonical" viewpoints, then the cognitive representations of the event should encompass more information and should therefore be more flexible if one has to rethink the event into other viewpoints – as demanded in the recognition task of experiment 2. But results show that the recognition performance still declines if the viewpoint presented during a later memory task differs from the viewpoint used in the initial learning phase, even if the observer initially saw the event from a canonical, information rich viewpoint. This means that the cognitive representation of a dynamic sequence still is viewpoint dependent, even if this viewpoint is especially information rich, i.e. delivers information about all or most important aspects of an event. This once more shows, that findings for static objects found by Biederman (Biederman, 1987; Biederman & Gerhardstein, 1993)

cannot simply be assigned to dynamic sequences: According to the geon structural description theory (Biederman, 1987) the cognitive representations of static objects are viewpoint independent, as long as these objects are shown from viewpoints that encompass the discriminative details of these objects, i.e. the so called "geons", and their relative positions to each other. Our findings contradict the applicability of this idea for dynamic sequences: Even using high discriminative viewpoints does not lead to a viewpoint independent cognitive representation; in fact the rethinking in other viewpoints still is critical for recognition performance. So our present findings rather point out, that findings with static objects or static arrangements of objects e.g. from Diwadkar and McNamara (1997), Shepard and Metzler (1971) and Tarr (1995) are more appropriate to predict memory processes of observers watching dynamic sequences, namely the formation of a viewpoint dependent cognitive representation (see the two significant effects of viewpoint deviation in experiment 2 for recognition accuracy and speed of recognition) and the occurrence of mental rotation processes (see the significant linear trend of viewpoint deviation in experiment 2) if new viewpoints are brought into play.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115 - 147.

Biederman, I. & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 1162 - 1182.

Blanz, V., Tarr, M. J. & Bülthoff, H. H. (1999). What object attributes determine canonical views? *Perception*, 28, 575 – 599.

Cameron, G. T. & Frieske, D. A. (1994). The time needed to answer: Measurement of memory response latency. In A. Lang (Ed.), *Measuring psychological responses to media*. Hillsdale, NJ: Lawrence Erlbaum, pp. 149 – 164.

Cohen, J. D., MacWhinney, B., Flatt, M. & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments & Computers*, 25, 257 - 271.

Cutzu, F. & Edelman, S. (1994). Canonical views in object representation and recognition. *Vision Research*, 34 (22), 3037 – 3056.

Diwadkar, V. A. & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science*, 8 (4), 302 - 307.

Edelman, S. & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32, 2385 – 2400.

Eley, M. G. (1982). Identifying rotated letter-like symbols. *Memory & Cognition*, 10 (1), 25 - 32.

Garsoffky, B., Schwan, S. & Hesse, F. W. (2002). Viewpoint dependency in the recognition of dynamic scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28 (6), 1035 – 1050.

Gibson, J. J. (1982). *Wahrnehmung und Umwelt: Der ökologische Ansatz in der visuellen Wahrnehmung* (transl. ed.). München, Wien, Baltimore: Urban und Schwarzenberg.

Härtel, H. & Lüdke, M. (2000). *XyZET: Ein Simulationsprogramm zur Physik*. Berlin, Heidelberg, New York: Springer.

Hamm, J. P. & McMullen, P. A. (1998). Effects of orientation on the identification of rotated objects depend on the level of identity. *Journal of Experimental Psychology: Human Perception and Performance*, 24 (2), 413 – 426.

Lawson, R. & Humphreys, G. W. (1996). View specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance*, 22 (2), 395 – 416.

Palmer, S., Rosch, E. & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & Baddeley, A. (Eds.), *Attention and Performance IX* (pp. 135 - 151). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Perrett, D. I. & Harries, M. H. (1988). Characteristic views and the visual inspection of simple faceted and smooth objects: Tetrahedra and potatoes. *Perception*, 17, 703 – 720.

Shepard, R. N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701 - 703.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, 2 (1), 55 - 82.

Acknowledgments

The work on these experiments was supported by grant from the DFG (Deutsche Forschungsgemeinschaft). The authors thank Andreas Wacker and Markus Huff for their help doing the programming of the stimulus material.

The Origins of Arbitrariness in Language

Michael Gasser (gasser@indiana.edu)
Computer Science Department; Lindley Hall 215
Indiana University; Bloomington, IN 47405 USA

Abstract

Human language exhibits mainly arbitrary relationships between the forms and meanings of words. Why would this be so? In this paper I argue that arbitrariness becomes necessary as the number of words increases. I also discuss the effectiveness of competitive learning for acquiring lexicons that are arbitrary in this sense. Finally, I consider some implications of this perspective for arbitrariness and iconicity in language acquisition.

A Language Design Task

Imagine you are inventing a language. It should associate signals ("forms") that can be produced and perceived by the users of the language with perceptual or motor categories ("meanings"). Assume that both forms and meanings are patterns of values across sets of dimensions and that you have been given the form and meaning dimensions. Assume further that the specific design task includes a set of meaning categories that need to get reliably conveyed. That is, given a particular pattern across the meaning dimensions, if it belongs to one of the given set of categories, a user who knows the language should be able to assign a form to it, that is, an appropriate pattern across the set of form dimensions. Similarly, given a pattern across the form dimensions, if it belongs to one of the set of form categories that you have built into your language, a user who knows the language should be able to assign a meaning to it. Furthermore, the form assigned to an input meaning should be the "right" form; that is, the form that gets output should pass the comprehension test in the reverse direction. Providing this form to a user who knows the language should result in an output meaning that is at least closer to the original meaning than to any of the other meaning categories. In the same fashion, the meaning assigned to an input form should pass the production test in the reverse direction.[1]

Your language is not hard-wired into a user; it must be learned through a series of presentations. A presentation consists of a pairing of a form and a meaning selected randomly from the set of possible form-meaning pairs that are built into the language, with a small amount of noise added to both the form and the meaning. Two constraints that you need to consider in your design are ease of learning and ease of storage. Each user has finite resources for learning and storage, and there is an advantage to languages that are learned with fewer presentations.

The main issue of concern in this paper is how the solution to a language design task of this type is constrained by the number of distinct meanings that are to be conveyed by the language. I will argue that there are advantages to languages with systematic relationships between forms and meanings and advantages to languages without such systematicity. I will then discuss how competitive learning fares at learning both types of languages. Finally I will discuss the implications for acquisition and evolution of human language.

Iconicity and Arbitrariness

How Iconicity Can Help

Learning the association between forms and meanings can be facilitated if there is a systematic relationship between the patterns. A simple example of such a relationship is a correlation between the values on a form dimension and a meaning dimension. There are two possibilities for where such a correlation might come from. One is for it to be based on a natural relationship between the two dimensions, for example, if they are the same dimension at a more abstract level. Such relationships are familiar in human language from onomatopoeia, in which form imitates meaning on one or more acoustic/auditory dimensions, for example, pitch. Examples of this type are more common in sign languages, where a movement of the hand in signing space may represent a physical movement of some object in meaning space.

A further possibility is for the relationship between the correlating dimensions to be completely arbitrary, or at least opaque to the users. In some sign languages, for example, American Sign Language and Japanese Sign Language, movement towards or away from the head represents the gain or loss of knowledge: learning, remembering, forgetting. But the motivation for the association between the form and meaning dimensions in this case would require that the user know that knowledge is in some sense in the head. Thus the relationship between the form and meaning dimensions in this case could be viewed as arbitrary by a particular learner, though the learner might still notice the systematicity of the rela-

[1] Note that in this sense, these simple languages deviate from human languages, which permit multiple forms for the same meaning and multiple meanings for the same form. But the constraint has to roughly hold for communication to get off the ground, and young children learning language seem to behave as though it does (Markman, 1989).

tionship, that is, that within this set of signs the head represents the location of knowledge.

These kinds of systematic relationships between form and meaning are referred to as **iconicity**. I'll return to the topic of iconicity and **arbitrariness**, the absence of iconicity, in human language later in the paper.

In an iconic language, there is less to learn than in a purely arbitrary language, so learning should be faster and require less storage. This is easily seen by imagining a language with five meanings to be conveyed and a single dimension each for form and meaning. An arbitrary language would require storing separately each of the five form-meaning pairs of values on this dimension, but a completely iconic language with a perfect correlation between the dimensions would only require a single value, a correlation of 1.0. This is illustrated in Figure 1.

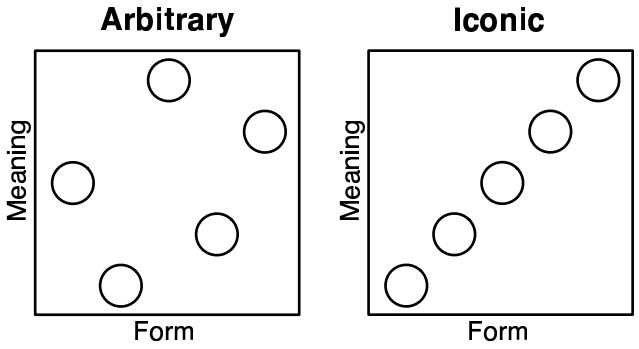


Figure 1: Arbitrariness and iconicity. Two simple languages, each with one form and one meaning dimension and five meanings to be conveyed. Noisy form-meaning pairs are indicated by circles in form-meaning space. In an arbitrary language, there is no correlation between form and meaning. In a perfectly iconic language, form and meaning correlate.

Iconicity can play a further role in the comprehension of the language. If an unknown or poorly learned form is presented in the presence of constraints on the possible meanings for the form, for example, if several candidate meanings are present, then iconicity can add further constraints. For example, if a user of a language knows that loudness in forms that refer to emotions tends to correlate with the strength of the emotion referred to, then for a particularly loud novel form, the user can eliminate candidate emotions that are mild.

How Iconicity Can Interfere

However, this advantage of iconicity should decline as the number of meanings to be associated with forms increases. Increasing the number of form-meaning pairs causes the average distance between these pairs in form-meaning space to decrease. Because of the noise that is part of form and meaning patterns, each form-meaning association occupies a region of the space. In other words, as the number of form-meaning pairs increases, the likelihood that the form regions for two different pairs share the same meaning (homophony) or that the meaning regions for two different pairs share the same form (ambiguity) increases. Obviously both sorts of overlap can interfere with communication; a noisy form pattern might get assigned to more than one meaning category, for example. They also interfere with learning; it will be more difficult to make the proper associations if forms or meanings are sometimes ambiguous.

Now consider how iconicity affects the likelihood of these sorts of overlap. Because iconicity constrains the possible form-meaning associations, it results in a narrowing of the space. This is illustrated in Figure 2. If we imagine the fixed set of meanings that are to be conveyed in the language as non-overlapping channels in the form-meaning space, then the possible forms for each can be viewed as circles (or hyperspheres in spaces of more dimensions) that can be slid back and forth in the channels, resulting in different languages. If we arrange two of these circles so that a portion of one is above a portion of another, we have the sort of overlap that represents ambiguity. There are obviously more ways to arrange the circles and avoid ambiguity in an arbitrary language like the one on the left than there are in a highly iconic language like the one on the right.

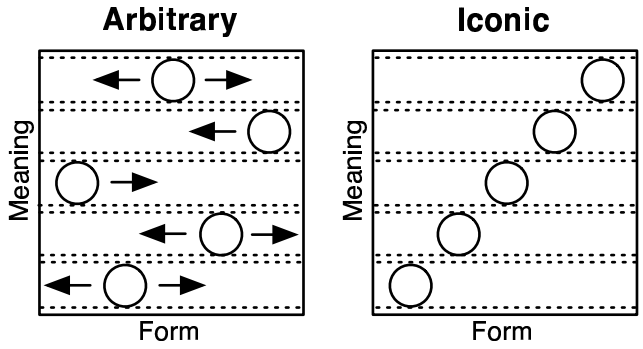


Figure 2: Arbitrariness, iconicity, and vocabulary size. For relatively large vocabularies, iconicity can interfere with communication because of the greater likelihood of overlap between form-meaning pairs. For a given vocabulary size, there are more ways to avoid ambiguity (and homophony) in an arbitrary than an iconic language.

A Simulation

For a learning algorithm that responds to regularities in the association between form and meaning, then, we should observe an interaction between vocabulary size and systematicity in the association (arbitrariness vs. iconicity), as measured by learning error.

To test this idea, I trained several feedforward connectionist networks to learn the associations from a set of meanings to a set of forms. The languages differed on two dimensions, vocabulary size and systematicity in the association. Both forms and meanings were represented by values along three dimensions, with ten possible values for each. Each dimension was represented by ten units, and each input and target value activated a gaussian pattern across the units so that there was the

possibility of some generalization from a value to values close to it.

The "small" languages contained 15 form-meaning pairs, while the "large" languages contained 100 form-meaning pairs. For "iconic" languages, each form-meaning pair coincided on two of the three dimensions, which were randomly selected for each pair. For example, a possible iconic form-meaning pair was: form {3, 2, 8}, meaning {3, 5, 8}. Note that for the iconic languages, there is thus a significant correlation across all three pairs of dimensions. For "arbitrary" languages, the values for each form-meaning pair were selected completely randomly. For each form-meaning pair, the network saw five separate presentations, one with the canonical pair, and four with noisy variations on this pair. For each of these variations, each dimension value was changed by 1 with a probability of 0.2.

Since these were feed-forward networks, they only learned the associations in one direction. Each network contained 30 meaning input units, 30 form output units, and 64 hidden units and was trained using back-propagation. Figure 3 shows the mean square error as training progressed. As can be seen, iconic languages have an early advantage because of the correlations that back-propagation can easily discover. For the small languages, this advantage holds throughout training. For the large languages, however, the network learning the arbitrary language eventually overtakes the one learning the iconic language, apparently because of the proximity of some of the form-meaning pairs to one another and the resulting confusion in the presence of noise.

Note that the potentially adverse effects of iconicity on learning depend crucially on the number of dimensions that are used to represent forms and meanings because the size of the form-meaning space increases with the number of dimensions. For a large enough number of dimensions, iconicity should be superior to arbitrariness, even for a relatively large vocabulary. In fact, if we increase the number of dimensions in the simulation from three to four, the long-term advantage of the arbitrary over the iconic language disappears.

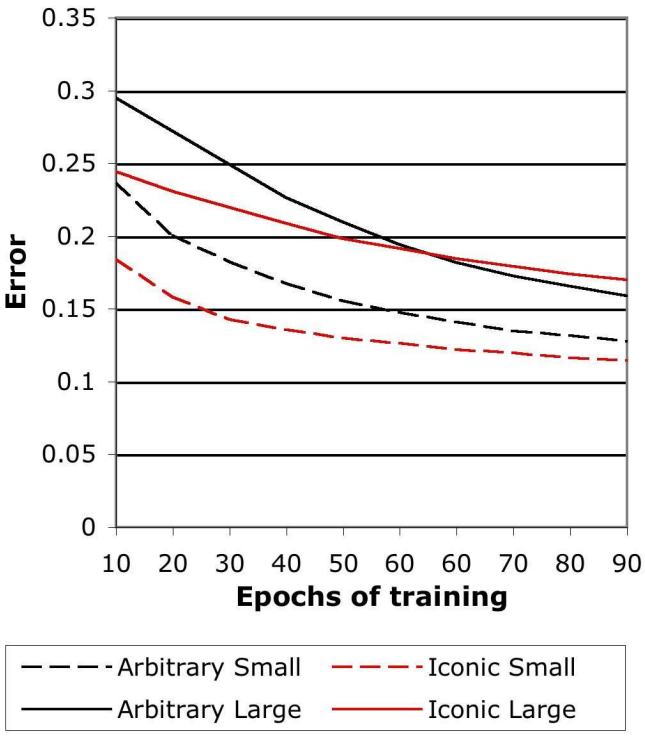


Figure 3: Learning of iconic and arbitrary languages by a feed-forward network. Root mean square error during training is shown for iconic and arbitrary languages consisting of 15 (small) and 100 (large) form-meaning pairs.

Arbitrariness and Competitive Learning

Learning Arbitrary Categories

Let us assume that the communicative demands of the users of the language require forms for a very large number of meanings and that the number of form and meaning dimensions available for representing forms and meanings is small enough that a mostly arbitrary language has a clear advantage over a mostly iconic one.

Now suppose we have some control over the kind of learner that is confronted with this large and mostly arbitrary language. What sort of learning mechanism would be best suited for this task? What matters most is that the different form-meaning pairs be kept distinct from one another. That is, each of these is in effect a separate category. (Since we are now dealing with categories of form-meaning association, it is time to start calling them "words.") Since in an arbitrary language there is little or no regularity to be found between the categories, an algorithm that focuses on within-category regularity, while it ignores between-category regularity, makes sense. Of course, the categories are not specified to the learner in advance; the learner neither knows how many form categories there are nor how many meaning categories there are. Thus the algorithm must be unsupervised.

Competitive learning (e.g., Grossberg, 1987) is such an algorithm (or family of algorithms). It seeks to cluster input patterns on the basis of similarity, and it is oblivious to any regularities that exist between the categories that it finds. It would seem to be well-suited to the task of learning words. But how does it respond to iconicity and arbitrariness?

A competitive learning network has an input layer and an output layer consisting of potential category units. The output layer either has a fixed number of units, representing an upper bound on the number of categories that can be learned, or, in a constructive competitive learning algorithm, the output layer adds new category units in response to error. In the simple version of competitive learning used here, for each input pattern the category unit whose weights best match the input pattern is treated as the "winner" for that pattern. It updates its input weights in the direction of the input pattern. The "losing" units also update their weights in the direction of the input, but with a much smaller step size.

A competitive learning network for the form-meaning

learning task has both form and meaning as inputs feeding into an output layer of category units. During training, an input pattern consisting of a form-meaning pair activates a winning unit, and the weights are updated. Ideally, a single category unit gets assigned to each form-meaning category; that is, each unit ends up representing a word. A single training presentation is illustrated in Figure 4A.

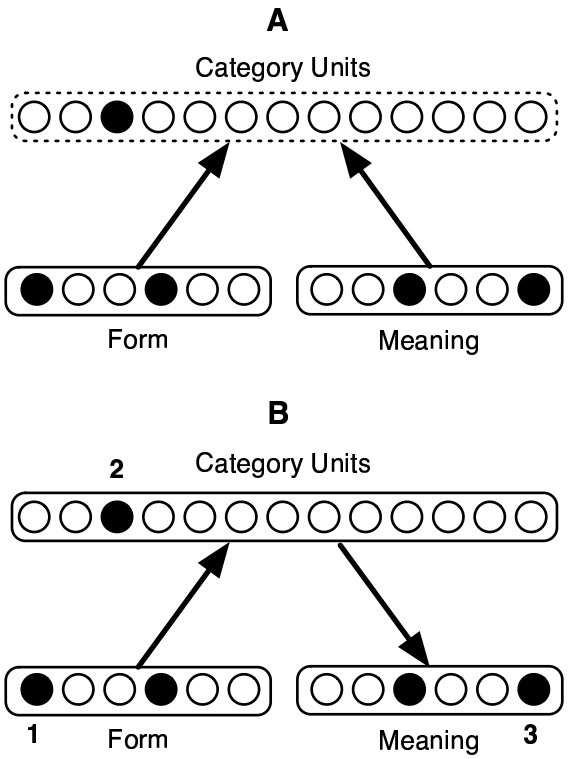


Figure 4: Competitive learning of form-meaning pairs. A. Training. An input pattern, consisting of both form and meaning patterns, is presented to the network, which selects a "winning" category unit, and updates its weights and, to a lesser extent, the weights of other units. In the constructive version of the algorithm used in the simulation, the category layer grows during training (indicated by the dashed border); it adds a new unit whenever error for an input pattern is above a threshold. B. Comprehension. An input pattern, consisting of a form pattern only, is fed to the network (1) and the winning category unit is activated (2). The active category unit activates a pattern on the meaning units (3).

Following training, the network can perform production or comprehension using the trained weights. For comprehension, a form pattern alone is input to the network, and a winning category unit is selected on this basis. This unit then activates the meaning units using the weights learned in the other direction. Production works in the opposite fashion, with meaning as input and form as output. Figure 4B shows how comprehension is implemented.

A Simulation

To test whether competitive learning could elucidate both the advantages and disadvantages of iconicity, I trained a competitive learning network of the type described above on both a completely arbitrary language and a maximally iconic language, in which all form dimensions correlated with meaning dimensions. There were four meaning and four form dimensions and 100 form-meaning pairs in the language. In addition to the form and meaning input layers, the network had a growable layer of category units. At each input presentation, a new category unit was added with a probability based on the error for the input pattern (the distance of the winning category unit from the input). Separate identical networks were trained for 50 epochs on the two kinds of languages. Figure 5 shows the results for several kinds of tests following training.

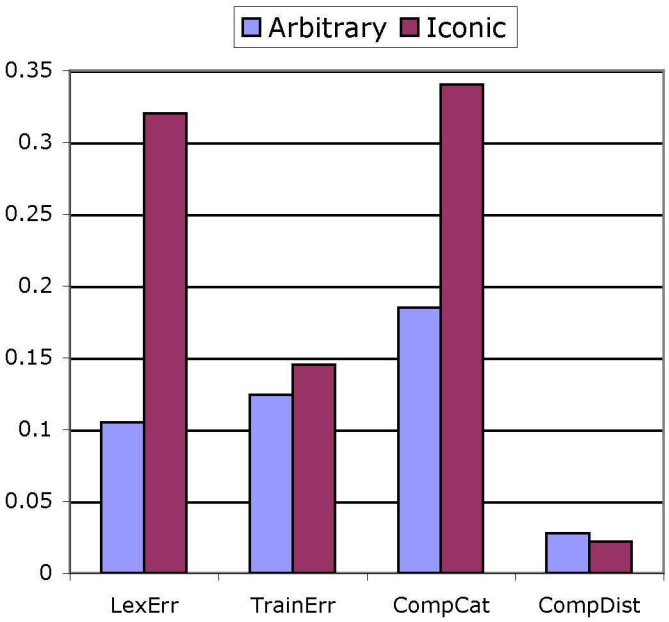


Figure 5: Competitive learning of arbitrary and iconic languages. Results are shown for the proportion of words that are not assigned distinct category units ("LexErr"); the final error on training patterns, that is, the average distance of input patterns from winning category units ("TrainErr"); the proportion of words in comprehension tests for which the meaning output was closer to a meaning category other than the intended one ("CompCat"); and the average distance of the meaning output in comprehension tests from the intended meaning ("CompDist").

The first two columns show tests directly related to the degree to which the networks mastered the languages. The first column shows what proportion of the 100 words became associated with distinct category units during training. Any category unit that ends up representing more than one word will obviously interfere with comprehension or production. For the iconic language there are more units doing double duty because of the greater similarity between the words. The second column shows another measure of learning, the average distance between an input pattern and the category unit that wins

when it is presented following training. The smaller this number, the more successful the network has been in handling all of the words. Again the network trained on the arbitrary language out-performs that trained on the iconic language.

The third and fourth columns represents tests of comprehension of forms, one for each of the words in the training set. There are two ways to test comprehension. One determines whether the meaning that is output is closer to the intended meaning (the one actually associated with the form in the language) than to any other. The result for this test appears in the third column. Again the arbitrary language has an advantage. A second way to test comprehension measures the distance between the meaning that is output and the intended meaning. The result for this test appears in the fourth column. Here the iconic language has a small advantage, one that holds over a range of parameter settings. This can be explained by considering what happens when a noisy or poorly learned form is presented to a network that has learned the iconic language. Even if the category unit that wins for this input is not the appropriate one, that is, the one that would yield the intended meaning, the meaning that is output will not be far off. Somewhat surprisingly, then, even though the iconic language is less well learned, it is more easily comprehended in this sense.

Human Language

What does all of this have to do with human language? Since at least the work of de Saussure (1983), it has been recognized that the association between form and meaning in human language is largely arbitrary. However, in Saussure's work and in other influential work by scholars such as Peirce (1998), iconicity and arbitrariness seem never to have been spelled out clearly enough to admit to any sort of rigorous test. They have always boiled down to "motivation" or "resemblance" or their absence.

The discussion above provides both a formalization of iconicity and arbitrariness and an account of why human language might have a strong tendency to be arbitrary. For whatever reason, we need to distinguish tens of thousands of categories of objects, attributes, states, and events, and the associations between these categories and the forms that convey them in a language need to be stored in a brain and to be learned through presentations that do not make explicit what the categories are. Under these circumstances, the arbitrariness of the form-meaning association helps keep words separate during learning.

Another implication of the discussion above is that word learning and word access in humans is a competitive process, that words are categories. This isn't a novel idea at all. In fact models of word recognition (e.g., Norris et al., 2000) and word access in language production (e.g., Levelt et al., 1999) that are not competitive are the exception. And the fact that competitive learning results in localized representations of words is compatible with the idea that words are the origin of symbolic behavior (Vygotsky, 1978).

But this brings up more questions. First, what about iconicity in human language? It is well-known that, far from being non-existent, iconicity actually thrives in some corners of language (Hinton, Nichols, & Ohala, 1994). It is a property of so-called expressive words, which make up an entire grammatical category in a wide range of languages, including Japanese, Korean, and many languages spoken in Africa, South Asia, Southeast Asia, and the Americas. It is also much more common in sign languages (Taub, 2001) than in spoken languages.

Given what I have claimed, we would expect iconicity in circumstances where the number of words is unusually small or in circumstances where the space of possible distinguishable forms is unusually large. The number of words is small early in first language acquisition, and there is some evidence that in at least one language with a large category of iconic words, Japanese, these words are relatively common in speech to children and they are easier for children to map onto meanings than arbitrary forms are (Yoshida, 2003). That is, they seem to play the role in comprehension that is suggested by the discussion above. Another situation in which a vocabulary is very small is experiments in which which adults have to communicate with one another without speaking. Not surprisingly, subjects in such experiments create highly iconic gestures to represent categories of objects and relations (Oda & Gasser, 2003).

However, expressives survive into the adult language for speakers of languages like Japanese, Tamil, and Zulu. One possible explanation is that these categories are more or less self-contained, existing in a sense in their own space. They tend to be characterized by particular formal properties such as reduplication, and they tend to convey particular categories of meanings such as movements, sounds, and textures. Perhaps expressives fail to interfere with other words because learners place them in a category all by themselves.

But what of sign languages? Although there is no evidence yet that the iconicity of sign languages helps young children pick up the meanings of words, there is lots of anecdotal evidence that adults learn sign languages relatively rapidly, presumably because of the iconicity. But how can we account for the pervasiveness of iconicity in the vocabularies (not to mention the grammars) of these languages? Although there is an apparent tendency towards somewhat less iconicity as these languages change, there is no evidence that the iconicity is disappearing (Taub, 2001). Of course it is possible that sign languages are more iconic than spoken languages because there are more ways to be iconic in the spatial than in the acoustic domain. But that does not explain how all of the iconicity can be tolerated, how the words keep from overlapping in the sense I have discussed. One possibility suggested by the account I've sketched is that the space itself is larger, that the number of dimensions along which signs vary or the number of distinguishable values along these dimensions is greater than it is for spoken word forms. This seems worth investigating.

Finally, how would competitive learning deal with a

language, or a subset of a language, that exhibited some iconicity, along with the more normal arbitrariness? The competitive network discussed above is doomed to being thrown off by the iconicity. Although it might, in the short run, perform better on a comprehension task, as happened in the simulation above, in the long run, it needs to be able to keep words separate from one another. However, there is nothing about competitive learning that restricts it to a single layer of category units. A more flexible network in fact is one that allows for different degrees of granularity in how the clustering of inputs takes place. This is achieved with layers with different numbers of category units or, for constructive networks, with different thresholds for the creation of new category units. The competition among units to classify inputs is only within, not between the layers. Such a network is shown in Figure 6. A network like this was trained on a set of 100 words, again with four form and four meaning dimensions, in which either the first form and first meaning dimension correlated or the second form and second meaning dimension correlated. The other two dimensions of form and meaning were uncorrelated. The larger category layer learned the set of words as before (note that the behavior of this layer is completely unrelated to the behavior of the other), while the smaller layer divided the patterns into two clusters. A comprehension task in a network like this relies on two winning category units, rather than one. That is, it can combine the correlational information embodied in the weights to the smaller layer with the arbitrary associations embodied in the weights to the larger layer.

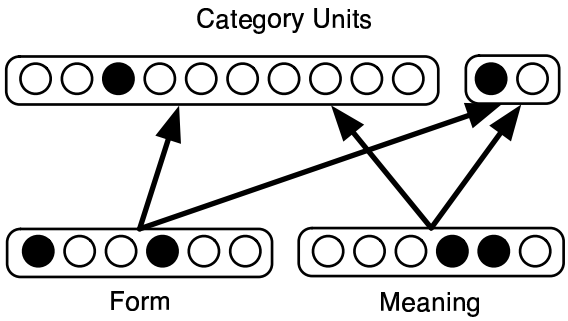


Figure 6: Competitive learning with multiple layers of category units. The number of category units (or the threshold for the creation of new units) in a layer governs the number of categories that it discovers.

Conclusions

Since iconicity seems to make so much sense and since humans are so good at imitation, it might seem surprising that human languages exhibit such overwhelming arbitrariness in the form-meaning relationships that define words. I have tried to show in this paper how the sheer number of concepts we feel the need to talk about inhibits us from making use of this strategy. It's crucial that words be kept separate, and it's easier to do this if there's little or no sense to how forms relate to meanings. This arbitrariness in turn favors algorithms that categorize form-meaning pairings, in short, algorithms that learn words. On this view, words are the local representations that result from the competitive learning of mainly arbitrary form-meaning associations.

But if this is so, how did it or how does it get that way? Did the advantage of being a competitive learner of form-meaning pairings cause our ancestors to evolve this approach to language? Or is this a mechanism that develops in children as they are exposed to a system that mostly fails to be iconic? Investigating the first possibility using evolutionary algorithms and investigating the second through the modeling of early word learning in children are future directions for this project.

References

de Saussure, F. (1983). *Course in General Linguistics*. G. Duckworth, London.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, *11*, 23–63.

Hinton, L., Nichols, J., & Ohala, J. (Eds.). (1994). *Sound Symbolism*. Cambridge University Press, Cambridge.

Levelt, W. J. M., Roelefs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction*. MIT Press, Cambridge, MA.

Norris, D., McQueen, J. J., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

Oda, H. & Gasser, M. (2003). Gesture Language Game — simulating the emergence of linguistic signs.. Paper presented at the 8th International Cognitive Linguistics Conference, Logroño, Spain.

Peirce, C. S. (1998). *The Essential Peirce: Selected Philosophical Writings: Volume 2*. Indiana University Press, Bloomington, IN.

Taub, S. F. (2001). *Language from the Body: Iconicity and Metaphor in American Sign Language*. Cambridge University Press, Cambridge.

Vygotsky, L. S. (1978). *Thought and Language*. MIT Press, Cambridge, MA.

Yoshida, H. (2003). *Iconicity in Language Learning: the Role of Mimetics in Word Learning Tasks*. Ph.D. thesis, Indiana University, Bloomington, IN.

On the Nature of Cognitive Representations and on the Cognitive Role of Manipulations. A Case Study: Surgery

Alberto Gatti (gatti3@unisi.it)
Department of Philosophy and Social Sciences, University of Siena
Via Roma 47, 53100 Siena, Italy

Abstract

The main point of discussion of the present article is the nature of representations, their formal structure and their origin and the cognitive role that manipulations of the environment can have. After having briefly reviewed the perspective expressed by the physical symbol system hypothesis, I take into account surgery as a case study that points out how many manipulative actions performed upon the environment have a cognitive relevance and the importance that the interaction with the environment can have in generating the representations used in cognitive processes and in giving them a formal structure. In the last part of the article I propose a model, that I call *Double Representation Approach*, which tries to give an explanation of the nature itself of representations, of the way they work in the cognitive processes and of certain important human cognitive behaviors.

Physical Symbol System Hypothesis

When first attempts were made to understand human cognition, one of the concepts that emerged as central was that of representation. A classical paradigm in which the notion of representation grew up was the hypothesis that an agent capable of intelligent action must be a physical symbol system (Newell, 1980; Newell & Simon, 1976).

A physical symbol system is a sort of device that contains symbols and symbol structures in memory and can perform processes upon these symbol structures. In more detail, according to the physical symbol system hypothesis, a physical symbol system and, thus, cognition performs three functional processes that occur sequentially and that are controlled by a central information processor. The three functional processes are the following: 1) a symbolic representation of the environment is constructed by means of a perceptual process performed by a perception subsystem; 2) the symbolic representation that has been constructed is delivered to the central processor, which processes it in order to extract information and to be able to select a symbolic expression that stands for an action; 3) an action subsystem decodes the symbolic description of the action and converts it into a concrete action in the environment.

It is important to understand what the terms "symbol" and "physical" mean. According to the classical definition given by Newell (1980), a symbol is an entity that stands for another entity. This kind of relation is called *designation* and its definition, with Newell's words, is:

> *Designation*: An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y. (Newell, 1980, p. 156).

Thus, a symbol is a syntactic element of a code and can be connected to other symbols to form symbol structures.

The term "physical" refers to the need for a physical implementation of a symbolic system in order for it to actually function and to actually operates upon and affect or be affected by the environment.

Following these definitions, we can distinguish three levels of organization in which a cognitive system can be divided: the semantic level, the symbol level and the physical level (Pylyshyn, 1989). At the semantic level, we have the content of knowledge and the goals that a system entertains. At the symbol level, the semantic content of the previous level is encoded by symbolic expressions. Finally, the physical level is constituted by the physical realization of the entire symbol system; in the case of humans, this level is represented by the biological level.

The postulation of a cognitive mechanism that works by means of symbols and symbol structures strictly implies the assumption that cognition takes place by means of internal representations and Newell (1980) considers "representation" as "simply another term to refer to a structure that designates" (Newell, 1980, p. 176):

> X *represents* Y if X designates aspects of Y, i.e., if there exist symbol processes that can take X as input and behave as if they had access to some aspects of Y. (Newell, 1980, p. 176).

Thus, according to the classical symbolic perspective, the central notion is that of representation. Now what we have to pay attention to and to focus on are two characteristics that Fodor and Pylyshyn (1988) indicate as the ones that identify classical symbolic models. Such characteristics are the *combinatorial syntax and semantics of mental representations* and the *structure sensitivity of processes*.

Let us begin with the first concept. Classical symbolic theories distinguish between structurally atomic and structurally molecular representations; structurally molecular representations are constituted by other representations that can be either atomic or molecular and the semantic content of a molecular representation is a function of the semantic contents of its syntactic constituents. According to this perspective, a *Language of Thought* (Fodor, 1975) is postulated, with syntactic components and structural relations between these components.

The second point is the structure sensitivity of processes. What this assumption means is that the principles by which mental representations are manipulated rely only on the structural properties of symbolic representations. More precisely, the formal, syntactic structure of a representation specifies the role of the representation within an inference and can cause the inferential process without reference to

the semantic content. Hence, the mental operations upon symbolic representations are activated only by the form of the representations.

So far, I have tried to delineate the main features of the classical physical symbol system hypothesis. In the following section, by means of a case study applied to the field of surgery, I want to question some aspects that emerge from the physical symbol system hypothesis.

A Case Study: Surgery

The arguments and claims that will be presented in the following two subsections emerge from a case study on surgery that I am conducting in the field in order to analyze the cognitive processes that go on in the work of surgeon. In particular, this study is devoted to analyze which kinds of representations are used in surgery, what the role of physical manipulations is, whether they have a cognitive relevance and how the distribution is of the cognitive processes involved in surgery.

In the following two subsections I want to point out two important cognitive elements: 1) the physical gestures involved in the processes of perception of data from the environment and 2) the formal structure and the origin of the representations used in cognitive processes. I will use surgery as a reference case and I will try to point out some relevant differences with respect to the classical physical symbol system hypothesis. In the first subsection I will take into account the case of a generic objective examination of the abdomen. In the second subsection, I will consider a specific surgical operation: inguinal hernia.

Objective Examination

The first step in the process that brings to a surgical operation is the examination that the surgeon conducts on the patient who feels specific symptoms. After a brief discussion to reconstruct the history of the patient, the surgeon begins what is called objective examination. Objective examination is a process of gathering of diagnostic data from the patient's body which is guided by the four evaluation principles of medical semeiology: inspection, auscultation, palpation, percussion (DeGowin & DeGowin, 1976; Swartz, 2002). In this subsection I take into account the case of a generic abdominal examination. This kind of examination is constituted by a series of evaluation acts that the surgeon accomplishes on the patient's abdomen by means either of external instruments or of parts of the body of the surgeon herself. In addition, this examination involves the cooperation of the patient, who is sometimes asked to make specific actions in interaction with the examination acts of the surgeon.

A generic abdominal examination can be schematized as in the following table 1.

Table 1: Abdominal objective examination.

	Evaluation action	**Means**	**End**
1	Inspection of the abdomen as a whole	Eyes	To evaluate how the aspect and the shape of the abdomen are
2	Inspection of the abdomen after having asked the patient to profoundly breathe	Eyes	To evaluate whether the abdomen moves
3	Auscultation of the abdominal wall before stimulating it with palpation	Stethoscope	To evaluate if there exists an intestinal peristalsis and how it is
4	Superficial palpation of the abdomen	Hands	To evaluate if there are signs of resistance to the abdominal wall that are linked to pathological situations
5	Deep palpation of the abdomen	Hands	To catch the aspects that the various parts of the abdomen can exhibit and that are linked to the contained bowels: -Consistence -Tension -Existence of masses
6	Auscultation of the abdominal wall after the stimulation by means of palpation	Stethoscope	-To evaluate whether, after having touched and moved the abdominal wall, an increase or decrease of the intestinal peristalsis has occurred -To evaluate if there is liquid out of the intestinal loops

The most important element that we can observe in the objective examination and that emerges from the scheme above is the following one: the surgeon uses specific per-

ceptive actions in order to catch and gather specific diagnostic data from the body of the patient. The data that the surgeon gathers are diagnostic signs that suggest a particular diagnosis or several different diagnoses and that help the surgeon in her abductive inference toward a final diagnostic hypothesis (Magnani, 2001). As diagnostic clues, the signs collected by the surgeon can be viewed, from a cognitive point of view, as representations that carry information.

The main feature of these representations is that they are constituted by structured sensations that the surgeon receives in response to her own structured perceptive actions. Therefore, two important aspects emerge: 1) the representations which are the diagnostic signs collected by the surgeon are sensations that the surgeon receives from the environment (the patient's body); 2) such representations are elicited and constructed by perceptive actions by means of which the surgeon interacts with the patient's body. These two points are important because they open the possibility to think of representations and cognitive processes in a new way which is different from the classical symbolic perspective.

According to the physical symbol system perspective, the representations that are used in cognitive processes are symbolic configurations that are inside the head of the cognitive agent. These representations are completely internal and are constituted by the symbols of a single language with a specific syntax. The case of medical examination seen above seems to lead to a more embodied perspective. It seems plausible to state that the representations on which the surgeon relies during an objective examination have, as formal structure, the one constituted by the sensations themselves that the surgeon receives. This formal structure can vary across a great range and, hence, is not the single one of the symbols of a single symbolic language.

Embodiment is also present at the level of cognitive processes. The physical symbol system perspective does not seem to give importance to the perceptive process as a moment in which not only data are simply perceived from the environment, but the environment is inspected and manipulated in specific ways in order to elicit more information. The perceptive actions that the surgeon performs during an objective examination have a strong epistemic value (Kirsh & Maglio, 1994), because they are devoted to examine the patient's body in specific ways so as to obtain specific information. These structured actions structure the sensations that the surgeon receives, that is, they structure her own representations.

A Surgical Operation: Inguinal Hernia

In this subsection I take into account a particular surgical operation as a case to point out some relevant aspects about the formal structure and the origin of the representations used by a cognitive agent. The surgical operation that I consider is the one of inguinal hernia in a male patient (Rutkow & Robbins, 1995; Shwartz, Spencer, Galloway, Tom Shires, Daly & Fisher, 1998; Trabucco & Trabucco, 1998). Inguinal hernia occurs when anatomical elements that are naturally contained in the abdomen enter the inguinal canal. Table 2 shows a schematic description of the main steps of an inguinal hernia operation in a male patient.

Table 2: Main steps of an inguinal hernia surgical operation

1	Incision of the cutis at the level of the inguinal canal
2	Incision of the fascia of the external oblique muscle to have access to the inguinal canal
3	Isolation of the spermatic cord
4	Isolation of the hernial sac
5	Rearrangement of the hernial sac into the abdomen
6	Hernioplasty at the level of the posterior wall of the inguinal canal
7	Suture of the fascia of the external oblique muscle previously cut
8	Suture of the cutis

One of the most relevant aspects of the inguinal hernia operation is that this operation requires a precise knowledge of the anatomy in order to recognize and carefully isolate the various anatomical structures that are found in the inguinal canal in an anatomical situation which has been altered by the hernia itself.

The main cognitive process which is involved in this surgical operation is, thus, the process of recognition. I define recognition as a matching process in which the real situation with which an agent has actually to do matches the salient features of an already defined internal representation that the agent entertains and that represents that identical situation or an analogous situation. Thus, in the case of surgery, a mechanism of recognition of an anatomical structure occurs when the anatomical configuration that the surgeon is confronting matches the internal representation that the surgeon entertains for that anatomical area.

Now, the first contact that a surgeon has with the external aspect of the anatomical structures of the human body is in the study of the illustrative anatomical tables on the anatomy books. This is the first moment in which the surgeon takes an anatomical representation that comes from outside and tries to memorize it, i. e., to bring it inside. I call this process internalization of an external representation. But every surgeon states that there is a difference between the book anatomy and the actual anatomical structures encountered in a real body, especially in those cases in which the anatomy has been altered by the pathological event, as it is in the case of the inguinal hernia. For this reason every surgeon states that recognizing the anatomical structures is a fact of experience.

Experience is a concept that deserves to be analyzed from a cognitive point of view. I have said above that the surgeon internalizes the anatomical representations that she studies on the anatomy books and that there is often a mismatching between these representations and the anatomical structures encountered in a real situation. Therefore, in order for the surgeon to be able to recognize anatomical structures in an actual situation, the book representations of the anatomy that the surgeon entertained must change in order to be in accordance with the anatomical structures actually encountered. Through the direct contact with the real anatomical elements, a process of change and adaptation of the previously internalized representations occurs. This is another

process of internalization, but more complex and slower than the previous one and slightly different from it.

This new process of internalization of external representations occurs by means of the repeated observation of many cases similar to each other and it is similar to a process of abstraction, in which, however, the final abstract representation is constituted by elements that have their deep origin in the experienced external elements. I use for this process the name of experience and I define experience as the process of internalization of external representations and of progressive change and adjustment, through the contact with the environment, of the internalized representations during which such representations acquire a configuration that conveys precise information in an unambiguous way and that can fit several different particular situations.

Also in the case of experienced surgeons there can be situations in which it is difficult to recognize the anatomical structures. This happens when there is such a situation in the operating field that a gap is created between internalized representations of the surgeon and actual configuration. It is in these circumstances that we can see again the cognitive role that manipulations can play. When anatomical data are confused and, therefore, the mechanism of recognition is made difficult, surgeons often make use of manipulations in order to find known anatomical reference elements. These manipulations are devoted to fill the gap between actual situation and internalized representations and, thus, they have an important cognitive relevance. The surgeon, for example, in front of a situation in which she cannot see an anatomical structure that she can usually see, almost always uses her hands to touch in certain areas in order to find anatomical reference elements that she expects on the basis of one of her internalized representations.

In this example is evident the cognitive role played by the manipulation, which examines the internal parts of a human body to construct an embodied representation which be in accordance with the internalized representation of the surgeon herself. At the same time, if the internalized representation of the surgeon is visual, in this example the surgeon is bringing into coordination two different representations of the same information, a visual representation and a tactile representation and this can be taken as an evidence that demonstrates that human cognitive agents are able to handle and, in fact, handle different representational codes and not a single one.

The case of surgery seems to push toward cognitive hypotheses that give the environment and the manipulations that humans perform upon it a predominant role as to the process of generation of the representations used in cognitive processes and as to their formal structures.

Double Representation Approach

In the previous section I have presented a case study and I have advanced some hypotheses about the origin and the formal structure of representations. In the present section I propose a model that tries to explain in a more detailed and more schematic way the nature and origin of representations and some critical cognitive behaviors that we can observe in human beings.

The classical view within cognitive science drew a distinction between what was called functional architecture and what was called anatomical architecture (Pylyshyn, 1989). The anatomical architecture can be considered as the implementational basis on which the functional architecture is realized. The functional architecture, instead, has to do with the algorithms that the mind uses when it carries out cognitive processes. The classical view concentrated its attention especially on the functional architecture and, in some cases, even argued that the physical level can be considered as a matter of implementational details. The functional architecture was described as the place of symbolic processes in which internal symbolic representations was processed. The symbols had an arbitrary relationship with their referents and formed symbolic structures which had a combinatorial syntax and semantics.

I want to revise the distinction anatomical architecture versus functional architecture and take into account the relationships that can take place between anatomical and functional architecture. To do so, I propose an approach that I call *Double Representation Approach*. Such approach locates two different representations as the components of any cognitive process that takes place in the interaction between human agent and environment. The idea is as follows:

1) **First-Level Representation**: this is the *pattern of neural activation* that arises as the result of the interaction between body and environment. This is the representation at the anatomical level and this is not the representation that human agents directly use in their cognitive processes.
2) **Second-Level Representation**: this is the thing for which the pattern of activation stands and this thing is the *sensation* that emerges from the encounter between our receptors and the structures of the environment and that is shaped by the structure itself of the environment. This is the representation at the functional level and this is the representation used by human agents in their cognitive processes.

The First-Level Representation is called representation because it is a pattern that has an analogical relationship with the structure of the environment and, thus, can be considered as an analogical representation of the environment. The Second-Level Representation could be considered as the "face" that we think the world has; in a certain sense it could be considered as the world itself. This assertion comes from the consideration that the world is always given to us through our sensorial perception, which can be considered, at last, as the only possible representation in which we can receive the physical world and the representation which is closest to the physical world. This means that the Second-Level Representation, even though it cannot be defined exactly as the world itself, has a strict relationship with the world. Now, I call representation the Second-Level Representation because the structured configurations of the environment are used in cognitive processes because of their carrying specific information, their representing such information. Therefore, the representations are the form of cognitively relevant information, they embody this information.

Now I want to explain in more detail what consequences the Double Representation Approach has on the way we can account for the way human agents reason. The First-Level Representation has a twofold influence on the way humans reason, the first one is direct, the second one is indirect and is mediated by the Second-Level Representation. We propose the following scheme:

1) The First-Level Representation, as a pattern of neural activation, can be assumed to influence the basic mechanism that underlies any cognitive process, regardless of the specific kind of representation used and of the specific algorithm followed. This basic mechanism can be assumed to be the one constituted by the construction of a pattern of elements and the fixation of this pattern.
2) Thanks to the analogical relationship between the patterns of neural activation (First-Level Representations) and the structures of the environment, the Second-Level Representations which we entertain reflect the structures of the environment. Therefore, human agents use representations that have several different formal structures each of which influences in a different way the reasoning process.

The Double Representation Approach can provide an account not only of the representations that originate in a direct contact between agent and environment, but also of those representations that human agents generate internally in the absence of the environment in order to solve a problem, represent a goal and so forth. The representations that originate in the contact between agent and environment take, at the first level, the form of patterns of neural activation and patterns of neural activation, after adequate training, tend to become stabilized structures and to fix. At this stage the patterns of neural activation no longer need a direct stimulus from the environment for their construction and fixation. In a certain sense they can be viewed as fixed internal records of external structures that can exist also in the absence of such external structures. These patterns of neural activation that constitute the First-Level Representations always keep record of the experience that generated them and, thus, always carry the Second-Level Representation associated to them, even if in a different form, the form of memory and not the form of a vivid sensorial experience. Now, the human agent, via neural mechanisms, can retrieve these Second-Level Representations and use them as internal representations or use parts of them to construct new internal representations very different from the ones stored in memory.

In the case of human beings, whose neural growth, according to the studies in neural constructivism (Clark, 2003; Quartz & Sejnowski, 1997), seems to be strongly environment-dependent, if we assume a strong relationship between neural mechanisms and cognitive processes, we are brought to the conclusion that the representations used in cognitive processes have a deep origin in the experience lived in the environment.

At least three conclusions can be drawn from the discussion above. First, as I have already said, the analogical relationship between First-Level Representations and environment causes the representations that human agents use in cognitive processes, that is, the Second-Level Representations, to be able to have various formal structures, we could say various types of syntax and not a single one. The Double Representation Approach tries to provide an explanation of the mechanisms that would occur in those cases in which it seems evident that humans are handling various types of representations.

Second, the Double Representation Approach seems to explain why human agents accomplish both computations of a "connectionist" type, such as pattern completion or image recognition and computations that use a combinatorial syntax and semantics, such as the ones exhibited in language usage. The First-Level Representation is generated as a pattern of neural activation and, if we assume, as I do, a more direct relation between neural basis and mechanism of reasoning, the mechanisms of connectionist creation of the neural pattern that constitutes the First-Level Representation could sometimes influence in a direct manner the mechanisms of reasoning carried out by means of the Second-Level Representation. This would explain the computations of a connectionist type. But, on the other hand, the First-Level Representation, in virtue of its connectionist character itself, has an analogical relationship with the environment and gives rise to structured sensations, that is, Second-Level Representations, that reflect the structures of the environment. Therefore, the cognitive agent can exploit all the syntactic structures that it finds in the environment and, most important, can follow the computations suggested by these structures. Now, among the syntactic structures that an agent can encounter there are the combinatorial ones and this would explain the combinatorial computations we experience.

Third, the fact that the Second-Level Representation is directly connected to the First-Level Representation and, thus, emerges from the interaction between body and environment points out the importance of the manipulative actions at the cognitive level. We can say that, in many cases, it is the actions of manipulation of the environment that create a specific representation that embodies specific information. In this sense, the Double Representation Approach could give an explanation to all those actions that human beings seem to perform not for achieving a physical goal, but for gathering specific information.

Conclusion

In this article I have first reviewed the classical physical symbol system hypotesis and I have concentrated on the fact that, according to this classical perspective, the representations that humans use in their cognitive processes would be internal symbolic structures of a single language with a specific syntax.

Subsequently, I have taken into account surgery as a case study to point out that the manipulations that humans use to perceive the environment may have a specific cognitive value and that the interaction with the environment plays a direct role in generating the formal structure of the representations used by human cognitive agents in their cognitive processes. The hypothesis that seems to emerge is that hu-

man agents use not a single representational code, but representations that can have multiple formal structures.

Finally, I have tried to construct a model that was able to explain in a more detailed way such hypothesis about the use of representations of multiple formal structures and that was able to provide an embryonic hypothesis about the way human cognition works.

References

Clark, A. (2003). *Natural-born cyborgs. Minds, technologies, and the future of human intelligence*. New York: Oxford University Press.

DeGowin, E. L., & DeGowin, R. L. (1976). *Bedside diagnostic examination*. Indianapolis, IN: Macmillan Publishing Co., Inc.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513-549.

Magnani, L. (2001). *Abduction, reason and science. Processes of discovery and explanation*. New York: Kluwer Academic/ Plenum Publishers.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*, 135-183.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*, 113-126.

Pylyshyn, Z. W. (1989). Computing in cognitive science. In M. Posner (Ed.), *Foundations of cognitive science*. Cambridge, MA: MIT Press.

Quartz, S., & Sejnowski, T. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral ad Brain Sciences*, *20*, 537-596.

Rutkow, I. M., & Robbins, A. W. (1995). Groin hernia. In J. L. Cameron (Ed.), *Current surgical therapy*. St. Louis: Mosby Co.

Shwartz, S. I., Spencer, F. C., Galloway, A. C., Tom Shires, G., Daly, J. M., & Fisher, J. E. (1998). *Principles of surgery*. New York: McGraw-Hill.

Swartz, M. H. (2002). *Textbook of physical diagnosis: history and examination*. Philadelphia, PA: W. B. Saunders Co.

Trabucco, E. E., & Trabucco, A. F. (1998). Flat plug and mesh hernioplasty in the “inguinal box”. Description of a technique. *Hernia*, *2*, 133-138.

Event Related Potentials (ERP) and Behavioral Responses: Comparison of Tonal stimuli to speech stimuli in phonological and semantic tasks.

Miriam Geal-Dor (gealdor@gesher.co.il)
Faculty of Life Science, Bar Ilan University
Ramat Gan, Israel.

Harvey Babkoff (babkofh@mail.biu.ac.il)
Department of Psychology, Bar Ilan University
Ramat Gan, Israel.

Abstract

Event Related Potentials (ERPs) were recorded from 20 young subjects to auditory target stimuli while they were performing three different tasks, using an odd-ball paradigm; 1. Tones: Subjects were instructed to respond to a 2kHz tone, and ignore a 1kHz tone. 2. Phonological: Subjects were instructed to respond only to words that had a specific ending ("f"). 3. Semantic: Subjects were instructed to respond to words that belonged to a specific category (name of alphabetic letters). EEG was recorded from 19 electrode sites. Peak amplitude of the early component (N100) did not differ significantly across the three tasks, while peak latency differed significantly across stimuli. In contrast, the later endogenous component (P300) was stimulus- and task-dependent. P300 latency differed significantly across stimuli and tasks; 327 ms to target tones; 668 ms to the phonological targets; and 706 ms to target words in the semantic task. P300 amplitude was significantly larger to tones than to linguistic stimuli. P300 peak amplitude recorded from electrode sites over the left hemisphere to the tonal target stimuli did not differ significantly from that recorded over the right hemisphere. In contrast, P300 amplitude recorded to both the phonological and semantic targets was significantly larger over the left hemisphere than over the right hemisphere. The present results can aid in our understanding of how humans process linguistic stimuli. These findings emphasize the importance of using similar experimental protocols for a broad comparison of the ERP response to a variety of stimuli and tasks.

Introduction

The process of auditory speech perception requires the use of sensory information in conjunction with linguistic knowledge. Event related potential recordings which have been increasingly used in the research of human cognitive processes, can provide information on the patterns of cortical activity that underlie different modes of processing various kinds of auditory and linguistic information.

The use of P300 for auditory presented tonal stimuli is well known. Studies have compared the ERP responses to tonal stimuli to vowels (Tiitinen et al 1999), syllables (Kayser et al 2001) or words (Lovrich et al 1988) and reported prolongation of latency as well as decrease in amplitude. These differences reflect the involvement of different processes in tonal and speech stimuli.

Using tonal stimuli Polich (1997)) reported an asymmetry in P300 amplitude with right hemisphere dominance specifically at the frontal and central electrode sites. They interpreted the data as reflecting the allocation of attention. Other researchers (Bruder et al 1999, Breier et al 1999) did not observe any laterality effect. Using speech stimuli a left hemisphere advantage was observed for phonemes, syllables (Kayser et al 2001, Alho et al 1998) and word stimuli (Breier et al 1999).

Studies have examined ERP morphology and topography using linguistic stimuli (Novick et al 1985, Henkin et al 2002). There seems to be an agreement among researchers that while phonological processing is characterized by a left hemisphere advantage, semantic processing is less localized, since it involves the activation of distributed networks in the brain (Lovrich et al 1988, Thierry et al 1998, Angrilli et al 2000).

In the present study, we used the oddball paradigm to generate a clear P300 component. We suggest that this paradigm, and specifically the P300 component, is appropriate to compare the ERP to a variety of target stimuli that lie along a continuum of auditory processing, from basic sensory discrimination of auditory features (tones) to cognitive language processing (e.g. phonology and semantics).

Methods

Subjects:

Twenty University students ranging in age from 20-26, mean age 22.5 (10 male and 10 female) participated in the study as part of their course requirement. Written informed consent was obtained, and the Bar Ilan University Ethic committee approved all experiments.

All subjects reported they were right handed, native Hebrew speakers, healthy and had no history of neurological or psychiatric disease. All passed a hearing screening test

performed in a quiet room using the Madsen OB 822 audiometer.

Stimuli:

Three different auditory tasks were tested using the oddball paradigm. One task consisted of tonal stimuli, and the other two tasks consisted of speech stimuli:

Tones: Subjects were instructed to respond to a 2kHz pure tone target and ignore the standard 1kHz tone. The tone duration was 50ms with rise/fall time of 10ms, and an interstimulus interval (onset to onset) of 2 sec.

Speech stimuli: High frequency Hebrew monosyllabic short words were chosen as stimuli. The duration of word stimuli ranged between 450-500ms. The same initial phonemes were used for both targets and nontargets so that discrimination between the targets and nontargets was only possible if the subject attended to the last phoneme. For example: If the target was “kaf” (alphabetic letter), the nontargets were “kal” (easy) or “kar” (cold). In a series of pilot experiments, we attempted to record ERPs using 11different target stimuli. The waveform in the expected P300 window was extremely spread with no clear peak. Consequently, in the present experiment we used three different targets and twelve nontarget stimuli to generate a clear P300 (See figure 1).

Two linguistic tasks were included in the experiment:
Phonological: Subjects were instructed to respond only to words that had a specific ending ("f").
Semantic: Subjects were instructed to respond to words from a specific category (name of Alphabetic letter).

The exact same target and nontarget words were used in the two speech tasks so that we could compare the behavioral and ERP responses to the same target stimuli in the two different linguistic tasks.

The oddball paradigm was programmed on a PC with the Audio task editor, Orgil medical equipment. In all experimental tasks conducted, a total of 180-195 stimuli were presented, thus the probability that a stimulus would be a target was 0.2. Stimuli were presented binaurally at 60 dBSL.

Procedure:

During the experiment subjects were instructed to fixate on a point located 1.5meters distant on the wall facing them, while keeping eye movement, blinks and general body movement to a minimum.

Subjects were instructed to press a button when detecting the target stimuli. A practice run was used to ensure that all individuals understood the task. Presentation order of the different conditions was counterbalanced across subjects. The entire session (of all 11 tasks not all reported here) lasted not longer than 3.5 hours.

The recording system:

The electroencephalogram (EEG) was recorded from 19 sites on the scalp according to the International 10-20 system referenced to back of neck. A ground electrode was placed on the right mastoid. An additional electrode placed below the right eye recorded electrooculogram (EOG) to monitor eye movement. The impedance measured for each electrode was lower than 7k ohm. The EEG program used to collect the data was Ceegraph IV Digital EEG system Biologic Corp. Raw data was continuously recorded with a band pass filter at 0.1-100Hz, sampling rate was 256Hz. Signals were amplified and digitized on line with a 4ms step.

All data underwent analysis using BPM Orgil medical equipment. Recordings were first segmented into epochs that were time locked to the stimuli and extended from 200ms pre-stimulus to 1800ms post-stimulus. Behavioral reaction time and accuracy were measured. The data were referenced to a common 100 ms pre-stimulus base line. Trials containing eye blinks or movements, excessive muscle activity artifacts were corrected or rejected. If more than 15 of the 35-40 sweeps of a given target were rejected for any reason, then all of the data in that condition for that subject was rejected. Thus, each ERP was based on a minimum of 20-25 sweeps.

Recordings to the target were averaged separately from recordings to the standard stimuli. The responses to standards preceding the targets were averaged and used as the comparison. ERPs were originally analyzed for correct response only. Because there were no differences between the averaged ERPs for correctly detected targets and those for all targets, further analysis was based on the later.

In a collateral behavioral experiment eight young naive subjects were instructed to write down exactly what they heard. Target words were cut and segmented in 25 ms intervals from 200ms to 500ms. All the segments were rearranged and randomly presented. The earliest cut off point where at least six subjects recognized the word correctly was defined as the point of identification for that word (e.g. "taf" was identified at 300ms). The results indicated that although the length of the words in the present experiment ranged from 450-500ms, all the words were correctly identified within the range of 275-350ms after word onset. ERP recording analysis were time-locked both to stimulus onset, and also to the point of identification based on the behavioral judgments. Using averaging to behavioral point of identification rather than to the onset of stimuli showed no significant difference in P300 peak latency and amplitude.

The measurements:

Behavioral measures of reaction time and performance accuracy were recorded as well as electrophysiologic

measures. ERP's were quantified in terms of peak latencies and peak amplitudes of the maximum negative or positive values within specific time windows. The time window for the different components was determined by visual inspection of the grand averages over all subjects. N100 was identified as the most negative point between 50 and 180ms post-stimulus. P300 peak amplitude was identified as the maximum positive point between 250 and 450ms for tones and 550 to 900ms for the speech stimuli.

Statistical analysis

Latency and amplitude values as well as behavioral measures were subjected to repeated measures analysis of variance (ANOVA) with 3 levels of task as well as 6 levels of electrode site as within subject factors. The level of significance was set to $p<0.05$.

Correlation tests were preformed between tasks, between behavioral and electrophysiological components.

Results:

Behavioral results:

The accuracy and reaction time data were analyzed (each separately) by a one-way analysis of variance (ANOVA) with task as a repeated variable. Accuracy measured as percent of target detection was not significantly affected by task. Task had a significant effect on reaction times to target stimuli ($F[2,36]=109.426$, $p<0.001$). Post hoc analysis revealed that the response to the target tones was always shorter than to the target speech stimuli ($p<0.001$), with no significant differences in RT within the speech stimuli (Table 1).

Table 1: Behavioral results of accuracy and reaction time averaged from all 20 subjects.

		tone	phonology	semantic
Accuracy (%)	Mean SD	93.74 8.04	93.94 9.87	90.1 11.82
Reaction time (ms)	Mean SD	459.95 87.46	829.67 92.27	869.3 98.72

Electrophysiological results:

Latency: Latency values (N100 and P300) were analyzed separately by a repeated measure analysis of variance (ANOVA) with 3 levels of task as a within subject factor. Both N100 and P300 latencies showed a significant main effect of task (N100 $F[2,38]=12.35$, $p<0.001$; P300 $F[2,38]=217.561$, $p<0.001$) (Table 2). Post Hoc analysis showed that N100 latency to the target in the tonal task was significantly shorter than to targets in both speech tasks ($p<0.001$). However, there were no significant differences in N100 latency to targets in the phonological task versus targets in the semantic task. Post hoc analysis revealed that P300 peak latency was significantly shortest to tonal stimuli ($p<0.001$), and within the speech stimuli, P300 latency to targets in the semantic task was significantly longer than to the targets in the phonological task ($p<0.044$) (Figure 1).

There were no significant correlations between any of the behavioral and electrophysiological measurements.

Table 2: N100 and P300 latency results averaged from all 20 subjects

		tone	phonology	semantic
N100 latency (ms)	Mean SD	91.48 29.14	129.48 25.51	124.4 22.53
P300 latency (ms)	Mean SD	327.5 18.76	668.71 78.4	705.58 74.34

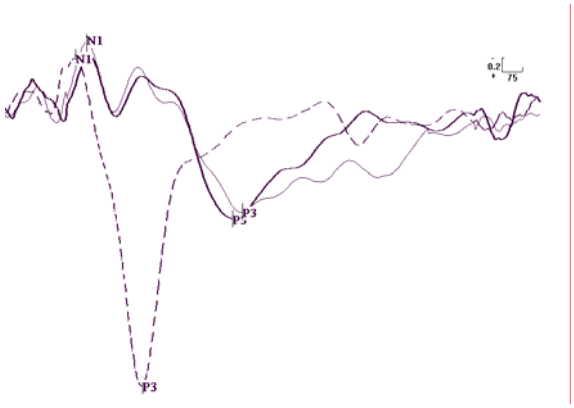


Figure 1: Grand average from all 20 subjects for the 3 tasks recorded at the Pz electrode. As can be seen, N100 latency was shorter for the tonal targets (dashed line) than the phonological (thick) and semantic (thin) targets. The P300 was shortest in latency and had larger amplitude to tonal targets as compared to speech targets.

Amplitude and Topography: A general repeated measure ANOVA with 5 levels for electrode site (frontal, central, parietal, occipital and temporal) revealed P300 amplitude was largest in parietal electrodes (Main effect of electrode site $F[4,76]=9.023$, $p<.001$). Further statistical analyses were performed on selected sets of scalp sites. On the basis of the observed distributions, the statistical analysis of ERP was limited to the central and parietal electrode sites (C4, Cz, C3, P3, Pz, P4).

Peak amplitude of N100 and P300 were analyzed separately by a two-way repeated measures analysis of variance (ANOVA) with 6 levels of electrode site and 3 levels of task as within subject factors.

N100: N100 amplitude showed a main effect of electrode site ($F[5,92]=144.194$, $p<0.001$) and did not show any significant effect of task. Post hoc analysis indicated that N100 peak amplitude was largest over the central electrode sites ($p<0.001$).

The degree of hemispheric asymmetry was computed by subtracting N100 peak amplitude recorded over the right hemisphere from that recorded over the left hemisphere (see Bellis et al 2000 for use of a similar index). As seen in

Figure 2 there was no significant difference in N100 amplitude recorded from the electrode sites over the left hemisphere (c3, p3) as compared to the electrode sites over the right hemisphere (c4, p4), for targets in either tonal, phonology or semantic tasks.

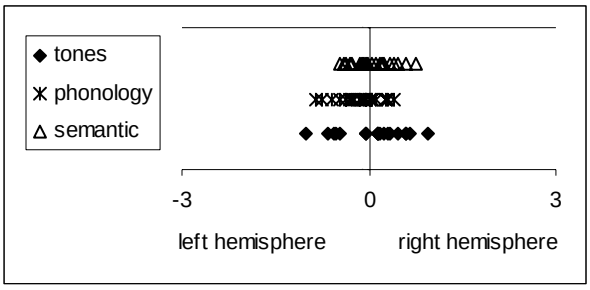


Figure 2: Degree of hemispheric symmetry in N100 amplitude. Results indicate responses were essentially symmetrical across all tasks.

P300: P300 peak amplitude was significantly affected by two of the variables, task and electrode site ($F[2,38]=21.08$, $p<0.001$; $F[5,95]=61.256$, $p<0.001$ respectively) as well as a two-way interaction of task X electrode site ($F[10,195]=3.021$, $p<0.001$). Post hoc analyses showed that the largest P300 amplitude was recorded over the parietal sites ($p<0.001$), and when comparing the tasks the largest P300 amplitude was recorded to targets in the tonal task ($p<0.001$).

P300 amplitude to targets in the tonal task were distributed symmetrically over the electrode sites. There was no significant difference in P300 amplitude recorded to tonal targets from the electrode site over the left hemisphere as compared to the comparable electrode site over the right hemisphere. In contrast, for both the phonological and semantic speech tasks, P300 amplitudes recorded from the parietal electrode site (p3) over the left hemisphere was significantly larger than P300 amplitude recorded from the parietal electrode site (p4) over the right hemisphere (phonology $t[18]=2.551$ $p<0.02$; semantics $t[18]=4.392$ $p<0.001$). (Figure 3).

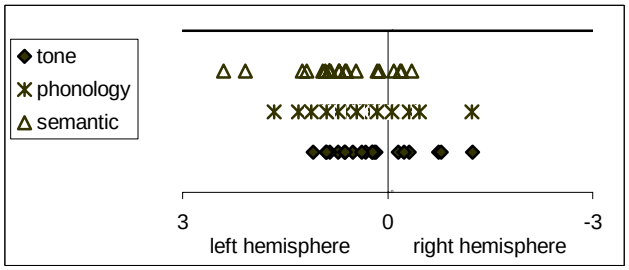


Figure 3: Degree of asymmetry for P300 amplitude. While for the tonal stimuli responses were essentially symmetrical, a significant degree of asymmetry can be seen for both speech stimuli, most pronounced in the semantic task, favoring the left hemisphere.

At the central electrode sites the distribution of P300 was symmetrical.

Discussion

Behavioral

The average accuracy scores for the 3 tasks ranged between 90-94% (Table 1). The high level of accuracy may have caused a ceiling effect and resulted in the inability to differentiate among the three tasks (Henkin et al 2002). In contrast, RT was sensitive to the different tasks. RT to targets in the tonal task was significantly shorter than to the targets in the two speech tasks. Although we did not find significant differences in RT to the targets in the phonological task versus targets in the semantic task, earlier studies did report such a difference (Novick et al 1985, Henkin et al 2002). The present study differs from the two earlier studies in that the construction of the targets in both the phonological and semantic tasks was such that subjects could not differentiate targets from nontargets unless they attended to the last phoneme. This may have presented a more difficult task than either of the earlier studies whose stimulus construction allowed for discrimination of targets from nontargets at an earlier stage of stimulus processing.

N100

The largest N100 peak amplitude was recorded over the central electrode sites. There were no significant differences between N100 recorded to targets and to nontargets in any of the tasks. Furthermore, there was a significant correlation between N100 to target and non-target in each of the tasks. These findings support the hypothesis that N100 represents obligatory primary sensory processing dependant upon the arrival of any stimuli at the auditory cortex, but does not by itself indicate any sort of discrimination or any of the task requirements (Martin et al 1999).

The N100 latency to the tonal stimuli was always shorter than to the speech stimuli, but there were no significant differences in N100 latency between the two speech stimuli. Similar results were previously reported (Wunderlich and Cone-Wesson 2001). Since N100 is an exogenous wave, it is sensitive to changes in the basic physical characteristics of the stimuli.

P300

As noted above, the P300 paradigm was chosen as the experimental technique so that the same electrophysiological components might be compared across a variety of stimuli and tasks.

In the present study P300 latency showed significant differences between the responses to tones and to the speech stimuli (327ms versus 668-706ms). The increase in latency for speech stimuli compared to tonal stimuli was reported previously (Tiitinen et al 1999, Kayser et al 1998). We

tested the hypothesis that the difference in P300 peak latency to tonal targets as compared to speech targets was due to the specific construction of the speech stimuli which required attention to the last phoneme before discrimination was possible. Subjects could only discriminate the target from the nontargets after hearing the last phoneme of the word while they could theoretically begin the process of discriminating the tone target from the tone nontargets beginning with stimulus onset. Furthermore, the duration of the tonal stimuli was 50ms, while the duration of the speech stimuli ranged between 450-500ms. This would mean that the difference in P300 peak latency to targets in the tonal task versus targets in the speech tasks should be directly related to the duration of the speech stimuli necessary to discriminate the words (Woodward et al 1990). In an adjunct experiment, we found that the average word identification point ranged between 275-350ms after word onset. Note the additional amount of time required for identification of speech stimuli (approximately 325ms) coincides with the time difference between P300 latency of tones (327ms) and P300 latency of speech stimuli (668-706ms).

Alternatively, it is possible that processing speech stimuli takes longer than processing tones. Therefore, the difference in P300 peak latency to tone targets versus speech targets also included the differences in processing time to the two types of stimuli (Bentin et al 1999).

The present findings can be related to the ongoing debate concerning the identity of the late ERP potential recorded to speech stimuli within the 500-750 ms time window, the "identity" thesis. Coulson et al (1998) argued that ERPs recorded in complex cognitive tasks are basically identical to (or are just modifications of) waves found in simpler conditions. Particularly, the P600 component of the scalp recorded event-related brain potential related to syntactic violation processing is just a delayed P300 similar to that recorded in simple oddball tasks (both are sensitive to probability manipulations and are similar in their respective scalp distribution). Kotchoubey and Lang (2001) used a paradigm in which subjects discriminated infrequent targets from frequent standards based on a semantic feature (e.g. animals versus other common nouns), this paradigm elicited a positive parietal wave in the 600 ms window frame. They argued that the P600 is an oddball delayed P300 component elicited in a semantic oddball experiment to more complex stimuli.

The alternative view states that there exist specific ERP waves manifesting brain mechanisms of language processing (Osterhout et al 1994, Frisch et al 2003). The late positive wave P600 recorded in response to syntactically anomalous words manifests specific brain mechanisms of syntactic processing.

Comparison between speech tasks: P300 latency to target stimuli in the phonology word tasks was significantly shorter than to targets in the semantic tasks. Similar results have been reported (Novick et al 1985, Cobianchi and Giaquinto 1997)

Topographical distribution

Tones: In our study both N100 and P300 peak amplitude in the tonal task were distributed symmetrically over the two hemispheres. These results are similar to previous reports using pure tones (Breier et al 1999) as well as complex tones (Bruder et al 1999, Kayser et al 2001).

Speech tasks: In the present study, While N100 peak amplitude distribution was symmetrical over the two hemispheres, P300 peak amplitude to the targets in the two speech tasks, was significantly larger when recorded from the parietal electrode over the left hemisphere (p3) as compared to the right hemisphere (p4). Similar results were reported in other studies using phonemes, syllables (Kayser et al 2001, Alho et al 1998) and words (Breier et al 1999). It is important to note that the change from hemispheric symmetry in tonal stimuli to hemispheric asymmetry in speech stimuli was found only for the P300 component and not for the N100 component (compare Fig 2 and 3). This dissociation of the two ERP components further emphasizes their different electrophysiological representations and may point to a dynamic change of hemispheric interaction in the processing of speech stimuli over time.

Phonology vs. semantics: A number of imaging and ERP studies have concluded that while phonological processing is more confined to regions of the left hemisphere, the semantic processing is less localized, since it involves the activation of distributed networks in the brain. (Ferlazzo et al 1993, Cobianchi and Giaquinto 1997, Thierry et al 1998, Angrilli et al 2000, Connolly et al 2001). For example, imaging studies demonstrated that phonological processes are related to Broca's area and the left inferior frontal gyrus (Demonet et al 1992, Becker at al 1999). However, during lexical-semantic tasks there is a wider cortical distribution of activation, not confined only to the left temporal and inferior frontal areas (Zatorre et al 1992, Kareken et al 2000, Zahn et al 2000). In the present study we found hemispheric asymmetry favoring the left hemisphere to targets in both the phonological and semantic tasks and did not find a significant difference in the hemispheric asymmetry of P300 peak amplitude favoring either the targets in the phonological or semantic tasks. These results are in line with several imaging studies (Poldrack et al 1999, Johnson et al 2001) that point to a greater activation of left hemisphere neural systems for both semantic and phonological tasks.

Acknowledgments

This work is part of the Ph.D. dissertation of the first author, was supported by the Schupf scholarship. The study was conducted in the Gonda Goldschmied Medical Diagnostic Research Center. The authors also thank Shlomo Gilat for

the technical assistance and Yury Kamenir for the statistical analysis.

References

Alho K., Connolly J.F., Cheour M., Lehtokoski A., Huotilainen M., Virtanen J., Aulanko R., Ilmoniemi R.J. (1998). Hemispheric lateralization in preattentive processing of speech sounds. *Neurosc. Lett. 258*, 9-12.

Angrilli A., Dobel C., Rocksroch B., Stegagno L., Elbert T. (2000). EEG brain mapping of phonological and semantic tasks in Italian and German languages. *Clin Neurophysiol. 111*, 706-716.

Becker J.T., MacAndrew D.K., & Fiez J.A. (1999). A comment on the functional localization of the phonological storage subsystem of working memory. *Brain and Cognition 41*, 27-38.

Bellis T.J., Nicol T., Kraus N. (2000). Aging affects hemispheric asymmetry in the neural representation of speech sounds. *J. Neurosc. 15*(5), 791-797.

Bentin S., Mouchetant-Rostaing Y., Giard M.H., Echallier J.F., Pernier J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: time course and scalp distribution. *J Cogn Neurosci. 11*(3), 235-60.

Breier J.L., Simos P.G., Zouridakis G., Papanicolaou A.C. (1999). Lateralization of cerebral activation in auditory verbal and non verbal memory tasks using magnetoencephalography. *Brain Topogr. 12*, 89-97.

Bruder G., Kayser J., Tenke C., Amador X., Friedman M., Sharif Z., Gorman J. (1999). Left temporal lobe dysfunction in schizophrenia: event related potential and behavioral evidence from phonetic and tonal dichotic listening tasks. *Arch. Gen. Psychiatry 56*, 267-276.

Cobianchi A. and Giaquinto S. (1997): Event related potentials in Italian spoken words. *EEG Clin. Neurophysiol. 104*, 213-221.

Connolly J.F., Service E., D'Arcy R.C., Kujala A., Alho K. (2001). Phonological aspects of word recognition as revealed by high-resolution spatio-temporal brain mapping. *Neuroreport. 12*(2), 237-43.

Coulson S., King J.W., Kutas M. (1998). ERPs and domain specificity: beating a straw horse. *Lang Cogn Proces 13*(6), 653-72

Demonet J.F., Chollet F., Ramsay S., Cardebat D., Nespoulous J.L., Wise R., Rascol A., Frackowiak R. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain 115*, 1753-68.

Ferlazzo F., Conte S., Gentilomo A. (1993). Event-related potentials and recognition memory within the 'levels of processing' framework. *Neuroreport 4*(6), 667-70.

Frisch S., Kotz S.A., von Cramon D.Y., Friederici A.D. (2003). Why the P600 is not just a P300: the role of the basal ganglia. *Clin Neurophysiol 114*(2), 336-40.

Henkin Y., Kishon-Rabin L., Gadoth N., Pratt H. (2002). Auditory event-related potentials during phonetic and semantic processing in children. *Audiol Neurootol. 7*(4), 228-39.

Johnson S.C., Saykin A.J., Flashman L.A., McAllister T.W., O'Jile J.R., Sparling M.B., Guerin S.J., Moritz C.H., & Mamourian A.C. (2001). Similarities and differences in semantic and phonological processing with age: patterns of fMRI activation. *Aging, Neuropsychology and cognition 8*(4), 307-20.

Kareken D.A., Lowe M., Chen S.H., Lurito J., Mathews V. (2000). Word rhyming as a probe of hemispheric language dominance with functional magnetic resonance imaging. *Neuropsychiatry Neuropsychol Behav Neurol. 13*(4), 264-70

Kayser J., Bruder G.E., Tenke C.E., Stuart B.K., Amador X.F., Gorman J.M. (2001). Event-related brain potentials (ERPs) in schizophrenia for tonal and phonetic oddball tasks. *Biol Psychiatry. 49*(10), 832-47.

Kotchoubey B., Lang S. (2001). Event-related potentials in an auditory semantic oddball task in humans. *Neurosci Lett. 14*, 93-6.

Lovrich D., Novick B., Vaughan Jr. (1988). Topographic analysis of auditory event-related potentials associated with acoustic and semantic processing. *EEG Clin Neurophysiol 71*, 40-54.

Martin B.A., Kurtzberg D., Stapells D.R.(1999). The effects of decreased audibility produced by high pass noise masking on N1 and the mismatch negativity to speech sounds /ba/ and /da/. *J Speech Lang Hear Res 42*, 271-86.

Novick B., Lovrich D., Vaughan HG. (1985). Event-related potentials associated with the discrimination of acoustic and semantic aspects of speech. *Neuropsychologia 23*(1), 87-101.

Osterhout L., Holcomb P.J., Swinney D.A.(1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *J Exp Psychol Learn Mem Cogn. 4*, 786-803.

Poldrack R.A., Wagner A.D., Prull M.W., Desmond J.E., Glover G.H., Gabrieli J.D. (1999). Functional specialization for semantic and phonological processing in the left inferior prefrontal cortex. *Neuroimage 10*(1), 15-35.

Polich J.(1997). EEG and ERP assessment of normal aging. *Electroencephalog. Clin. Neurophysiol. 104*, 244-256.

Thierry G., Doyon B., Demonet J.F. (1998). ERP mapping in phonological and lexical semantic monitoring tasks: A study complementing previous PET results. *Neuroimage 8*(4), 391-408

Tiitinen H., Sivonen P., Alku P., Virtanen J., Naatanen R. (1999). Electromagnetic recordings reveal latency differences in speech and tone processing in humans. *Brain Res. Cogn. Brain Res.* 8(3), 355-63.

Woodward S.H., Owens J., Thompson L.W. (1990). Word to word variation in ERP components latencies: spoken words. *Brain Lang 38*, 488-503

Wunderlich J.L., & Cone-Wesson B.K. (2001). Effects of stimulus frequency and complexity on the mismatch negativity and other components of the cortical auditory evoked potential. *Acous. Soc Am. 109*, 1526-36.

Zahn R., Huber W., Drews E., Erbrich S., Krings T., Willmes K., & Schwarz M. (2000). Hemispheric lateralization at different levels of human auditory word processing: a functional magnetic resonance imaging study. *Neurosci lett. 287*, 195-198.

Analogical Encoding: Facilitating Knowledge Transfer and Integration

Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208 USA

Jeffrey Loewenstein (jeffrey.loewenstein@columbia.edu)
Columbia Business School, 3022 Broadway
New York, NY 10023 USA

Leigh Thompson (leighthompson@kellogg.northwestern.edu)
Kellogg School of Management, 2001 Sheridan Road
Evanston, IL 60208 USA

Abstract

People's ability to recall and use prior experience when faced with current problems is surprisingly limited. We suggest that one reason is that information is often encoded in a situation-specific manner, so that subsequent remindings are limited to situations that are similar to the original both in content and in context. *Analogical encoding*—the explicit comparison of two partially understood situations—can foster the discovery of common principles and allow transfer to new structurally similar situations. This paper addresses two new questions: (1) whether comparison can also improve people's ability to retrieve examples from long term memory; and (2) whether simply providing the common principle would suffice to promote transfer. The results show (1) that not only does comparing examples facilitate transfer *forward* to a new problem, it can also facilitate transfer *backwards* to retrieve an example from memory; and (2) providing a common principle is not sufficient: comparison is still beneficial.

Introduction

The ability to transfer relational knowledge across contexts is of central importance in human cognition. (Gentner, 2003). Yet people do not acquire relational abstractions effortlessly (Chi, Feltovitch & Glaser, 1981; Chase & Simon, 1973), nor do they always apply them when they would be helpful (Gick & Holyoak, 1980). Drawing analogies during learning can address both of these challenges (Catrambone & Holyoak, 1989; Gentner, Loewenstein & Thompson, 2003). We will also examine another role of analogy in learning. Our findings indicate that comparing two structurally similar examples (analogical encoding) not only facilitates transfer to *future* structurally similar cases, but also the retrieval of *prior* structurally similar examples from memory.

An important means of learning is analogical transfer—the use of a prior familiar situation (the base) to solve a novel situation (the target) by mapping the solution from the base problem to the target problem. This kind of transfer has been shown to occur in reasoning and problem-solving (Bassok, 1990; Novick, 1988; Reed, 1987; Ross, 1987). Research on analogical transfer also reveals an Achilles' heel. When people succeed in accessing appropriate prior examples to inform current problems, they perform well (e.g., Pirolli & Andersen, 1985). The importance of prior cases in current reasoning has been argued persuasively in the case-based reasoning literature (Kolodner, 1993; Schank & Riesbeck, 1981). However, people often fail to access prior cases that would be useful, even when they can be shown to have retained the material in memory (Gentner, Rattermann & Forbus, 1993; Gick & Holyoak, 1980). Indeed, people are often unable to solve a problem after having just solved an analogous problem (see Reeves & Weisberg, 1994 for a review).

One way to promote structural transfer is by comparing two initial examples (Gick & Holyoak, 1983). This process capitalizes on the fact that comparison between two exemplars tends to make common relational structure more salient (Gentner & Markman, 1997). We call this *analogical encoding*, to emphasize that one can compare two partly understood examples to derive a common interpretation (Kurtz, Miao & Gentner, 2001). As Gick & Holyoak (1983) demonstrated, comparing two initial examples can facilitate deriving a schema, which in turn facilitates transfer to a structurally-similar problem (Catrambone & Holyoak, 1989). In contrast, people learning from a single case tend to encode it in a context-specific manner, with the result that later remindings are often based on more obvious surface aspects (Gentner & Rattermann, 1991).

We are investigating analogical transfer in the domain of negotiation (Loewenstein, Thompson & Gentner, 2003; Thompson, Gentner & Loewenstein, 2000). Negotiation is a particularly apt arena, for several reasons. Negotiation principles must be applied across many different contexts, making transferability essential. Further, the learning must be applied in real time, often in stressful, competitive situations with the potential for considerable gain or loss. Finally, our participants are highly motivated; they are studying negotiation with a direct interest in raising their job effectiveness.

In a typical negotiation situation, there is a set of issues in which two parties have different preferences, and to which they assign different levels of importance. The goal is to achieve an agreement, with each participant trying to optimize their gain. The negotiation principle we focus on

here is the idea of constructing contingent contracts: agreements whose terms depend on the outcome of a future event (Bazerman & Gillespie, 1999). These contracts allow people to reach agreement despite differences in opinion. Despite their usefulness, untrained negotiators tend not to form contingent contracts. Instead, most negotiators form inefficient compromises (Thompson & Hastie, 1990).

In our studies, the basic method is as follows. We provide highly motivated students (typically MBA students or executives) with materials to prepare for a face-to-face negotiation. Before negotiating, all participants read two brief cases that illustrate a negotiation principle (e.g., a contingent contract) that would be advantageous to use in their face-to-face negotiation. Half the participants (the comparison group) are told to compare the two cases and write out their commonalities; the other half (the separate-cases group) is told to read each case and write out what is important about it. Participants are then paired with someone in the same condition to conduct the negotiation, which is set in a different context than the study cases. We have found that participants who compared the two cases are two to three times as likely to use the negotiation strategy in their subsequent face-to-face negotiations as those who analyzed the same cases one at a time (e.g., Thompson, Gentner and Loewenstein, 2000).

Comparing analogous cases promotes *forward transfer*. That is, it increases the likelihood that the common principle will be retrieved when an analogous situation occurs in the future. One route that has been proposed for this increased transfer may be that schema-abstraction leads to increased matching (Ross, 1989). This possibility follows directly from the assumptions of schema-abstraction. Comparison invites an alignment and re-representation of examples, yielding a common representation that is less context-specific than the initial ones (Gick & Holyoak, 1983; Loewenstein, Thompson & Gentner, 1999). Because this new representation has more general relational representations and fewer potentially conflicting object matches than the initial cases, the match with subsequent cases will be better, making remindings more likely (Forbus et al, 1995; Ross, 1989). However, there is another possible reason for the increase in forward transfer, namely, *learning-to-encode*. Having derived a common representation, people may encode future cases in the domain in a similar manner. (Medin & Ross, 1989). The learning-to-encode account predicts that future cases are likely to match the schema that resulted from the comparison, because they will be encoded in the same way.

Using retrieval to clarify the transfer process. Although most researchers have assumed that both increased matching and learning-to-encode are part of the story, the forward-transfer improvement could be explained solely by learning-to-encode. However, the learning-to-encode account cannot predict any effect on retrieval of cases acquired prior to the schema abstraction. Thus, if learning-to-encode is the main reason for improved transfer, then the effects of comparison should be unidirectional: better learning today will help performance tomorrow, but will be of little help in retrieving examples that were learned yesterday. In contrast, if schema abstraction per se is an important force here, then we will see bi-directional effects: abstracting a schema should aid in retrieval whether it is forwards or backwards.

Experiment 1

In Experiment 1, we tested whether analogical encoding facilitates *both* forward transfer and backwards retrieval. If so, this would suggest a degree of symmetry in the memory retrieval process. That is, it would suggest that an abstract relational structure is better retrieved by future cases and also serves as a better probe for prior cases than a specific case.

We gave all participants two analogous examples. We asked half to compare them, and half to study the examples one at a time. Next, to test whether comparison aids memory *retrieval,* we asked people to recall an example from their own experience that illustrated the same principle as the initial examples. Finally, to test for *transfer*, we asked whether people would use the principle to form better agreements in a subsequent face-to-face negotiation (as in our prior studies).

Our participants were management consultants in a negotiation training program. Given the amount of money and time devoted to this training, there is no question that they were highly motivated to learn. They should also be professionally predisposed to value learning and generalization.

Method

Participants A total of 124 participants aged approximately 25 to 45 years, all full-time professional management consultants working at the same organization, participated through a negotiation training seminar. There were 64 in the comparison condition and 60 in the separate cases condition.

Materials and procedure. Participants read role materials to prepare to be either the buyer or seller in a negotiation role-playing scenario. Just prior to engaging in the role play, participants received a training packet. The first page concerned details about their upcoming negotiation. The next pages contained two cases exemplifying the contingent contract principle. The Comparison group read both cases and then was asked "What is going on in these negotiations? Think about the similarities between these two cases. What are the key parallels in the two negotiations? Please describe the solution and say how successful you think it is." The Separate case group received the following question after each case: "What is going on in this negotiation? Please describe the solution and say how successful you think it is."

For both groups, the next page of the training packet asked participants to recall an example like those they had just read: "Please think of an example, preferably from your own experience, that embodies the same principle as that on

the previous page." We then asked participants to state the source of their examples. Participants were not limited as to time and typically spent 45-60 minutes on these three pages (an indication of their motivation to learn). We saw no time-on-task differences by condition. Then they were paired with someone from the same condition to negotiate. The negotiation case was set in a different context than the training cases, and was designed to afford creating a mutually beneficial contingent contract.

Scoring The negotiated agreements were scored by blind raters as to whether they contained a contingent contract (which was, by design, the optimal solution to the negotiation dilemma). Coders also rated the quality of the participants' initial responses concerning the two cases—that is, the degree to which the contingent contract schema was described—using a 3-point scale: 0 = no elements of the schema were present, 1 = some elements, and 2 = all elements. They also rated whether participants linked the case and principle in any way (as a manipulation check on the condition difference). Finally, coders also rated whether the examples participants recalled were contingent contracts, using the same scale as above, and categorized the source domains in which participants' examples were set. Overall, there was high agreement (87%); disagreements were resolved through discussion.

Results

As predicted, the comparison group was superior to the separate cases group on all three measures: schema quality, likelihood of transfer, and quality of remindings. Making comparisons led to grasping the contingency contract schema from the original examples, which in turn facilitated both linking it to prior examples in memory and using it in a new negotiation situation.

Schema understanding Comparison participants (M = 1.45) articulated the schema better than Separate case participants (M = 0.98), $t(122) = 2.97$, $p < .01$. Another striking finding was that fewer than one in five Separate cases participants linked the two cases in any way, despite the fact that they occurred contiguously and were analogous.

Transfer In their face-to-face negotiations, Comparison participants (69%) were nearly twice as likely to make contingent agreements than Separate case participants (33%), $?^2$ (1, N=62) = 4.22, $p < .05$. As in our previous research, this suggests that comparison facilitates transfer.

Remindings Participants in the Comparison condition (M = 1.25) retrieved better examples of contingency agreements than did participants in the Separate cases condition (M = 0.82), $t(122) = 2.65$, $p < .01$. This suggests that comparison aided people's understanding of the initial cases, thereby better guiding participants' retrievals.

Participants retrieved examples from their own business experience or that of a colleague, and less frequently drew upon examples from the popular press. The examples were mainly from the business domain (as expected—they were in a business training classroom), with the remainder being daily life examples such as betting on sports teams, uncertainty about the weather affecting a vacation activity, arranging a home mortgage, and so forth.

One source of participants' examples had a name within their organization—*value billing*. Value billing is a particular type of contingent contract wherein a consulting firm bills clients a low base fee, with a generous bonus structure based on the outcome of the work. Given that every participant probably knew about value billing, it is striking that most of those who used this example were in the Comparison condition.

Cross-measure associations As expected, schema understanding predicted retrieval performance. The association between articulating the schema and retrieving a match was reliable, ?2 (1, N=124) = 8.68, p < .01. In the transfer measure there was only a modest trend for the sum of a pair's schema ratings to be associated with their transfer ?2 (1, N=62) = 1.70, $p = .19$. However, "high performance pairs" (pairs in which at least one person articulated the schema and retrieved a matching example, and the pair formed a contingent contract—i.e., transferred) were marginally more likely to be in the Comparison condition (47%) than the Separate cases condition (23%), ?2 (1, N=62) = 3.75, $p = .05$.

Distinguishing retrieval from invention. To conclude that there is a comparison advantage for retrieval, it is important to assess whether participants were simply fabricating examples rather than retrieving them. That is, the retrieval advantage for the comparison group could stem simply from their using the derived schema to invent examples, rather than from recalling them. But in this case, we should see the highest proportion of structurally correct "retrievals" from participants who failed to state a source. In fact, the 32 people who did not state the source produced the *lowest* proportion of structural remindings (31%). The proportion was higher for those who stated non-verifiable sources (45%), and highest for those whose source was verifiable (and verified) (68%). The opposite would have been expected on the 'fabricating' account, and hence it seems reasonable to take the participants at their word.

Discussion

Comparing cases yielded consistent advantages for schema abstraction, retrieving a matching example from memory, and transferring to solve a new problem. Although our participants were consultants whose jobs depend on their ability to apply their knowledge in new situations, and who spent considerable time with the training materials, we saw little spontaneous comparison across examples in the group not explicitly told to compare. Nonetheless, a brief instruction to compare was sufficient to advance the

performance of their peers across all three measures of learning.

Another striking pattern is that despite the participants' considerable experience in the business world, over half of them did not recall any examples that were contingent contracts (or structural analogs). In fact, 11% failed to write down any case at all. Our results underscore that (1) transferring from analogous examples can be challenging even for sophisticated and motivated learners (Novick, 1988) (2) analogical encoding can dramatically increase transfer; (3) the benefits of analogical encoding derive in part from inducing a clearer schema for the common principle (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983) and (4) analogical encoding can lead to increased retrieval of prior analogous examples.

Finally, the fact that analogical encoding aids in memory retrieval indicates that the effect of schema abstraction in memory access is bidirectional. The representations that resulted from comparison were *both* more readily retrieved by future analogs than were the separate cases and more effective as probes for prior analogs stored in memory. Thus, although we suspect that the transfer benefits of comparison derive in part from learning-to-encode—i.e., from encoding future examples in a structurally clear manner consistent with the schema—our results indicate that ease of matching must also play a role. The relatively abstract schemas that result from analogical encoding match better with prior examples just as they do with future examples.

Why not just give them the principle? The results of this study and prior work on analogical encoding lead naturally to a further question. If the advantage of mutual alignment is simply that comparing the two examples leads learners to derive the principle, then would learners not fare even better if the principle—in this case, the contingent contract principle—were simply given to them explicitly? We examine this directly in Experiment 2.

Experiment 2

In Experiment 2 we asked Masters of Business Administration (MBA) students to read a case and an abstract principle. If analyzing the principle and elaborating upon it is sufficient for transfer, we should find high rates of transfer in all groups. However, if principles need to be grounded in examples to be comprehensible and generalizable, then those asked to compare the case and principle should show a transfer advantage relative to those who study the example independently of the principle.

Method

Participants A total of 106 MBA students participated in the study, resulting in 27 pairs in the Comparison condition, and 26 pairs in the Separate condition.

Materials and Procedure The materials and procedure were similar to Experiment 1. The training packet did not ask for memory retrievals, and instead of two cases presented people with one case and an abstract description of contingent contracts. Participants in the comparison group were asked to compare the case and principle and specify commonalities, and then describe implications for negotiation. Those in the separate group were asked to read the case and the abstract principle separately, and were asked after each to state its implications for negotiation. Both groups received case and principle on consecutive pages. Participants then engaged in the negotiation with someone else in the same condition.

Scoring As before, coders rated the quality of the contingent contract schemas. They also rated whether participants had paraphrased the case in their responses, whether their responses contained generic advice about negotiation that was unrelated to contingent contracts, and whether they linked the case and principle in any way. They agreed on 93% of their judgments, and disagreements were resolved through discussion.

Results

In their initial descriptions, Comparison participants were more likely than Separate cases participants to articulate the full schema (74% versus 56%) and they less often failed to articulate any of the schema (12% versus 35%), $?^2$ (N=97, 2) = 7.36, $p < .05$. In their face-to-face negotiations, participants who compared the case and principle (44%) were over twice as likely to form contingent contracts as were participants who analyzed the case and principle separately (19%), $?^2$ (N=53, 1) = 3.87, $p < .05$.

The additional ratings of people's individual responses to the training materials showed a further surprising and consistent pattern. Despite the fact that all participants had read and discussed the case and the principle on consecutive pages, almost none of the Separate participants noticed the link between them. Thus, the Separate participants did not appear to notice that the principle was the general statement of what the case exemplified. Comparison participants (88%) were also more likely than Separate participants (34/47, or 71%) to paraphrase the case as they discussed it, $?^2$ (N=97, 1) = 4.24, $p < .05$. Participants in the Separate cases condition were also more likely to give general panaceas as advice (e.g., "it helps to have a good relationship when you're negotiating" or "you want to reach win-win deals") (77%) than comparison participants (31%), $?^2$ (N=97, 1) = 21.06, $p < .001$.

Discussion

Our results lead to something of a paradox. We find that learners who derive a schema through analogical encoding—either by comparing cases, or by comparing a principle with an example—can readily transfer the schema to new cases. Yet learners who are explicitly given the same schema—even along with an example case—cannot. Why? Can we say anything more specific that that "active learning is good"? We suspect that abstract principles are ineffective

because they are less well understood than specific cases (Forbus & Gentner, 1986; Regehr & Brooks, 1993). Indeed, in our study, some people had difficulty re-stating the principles. This is partly because learners may fail to understand the specific terms used, or how they are meant to combine. This is consistent with Ross & Kilbane's (1997) finding that if given an example followed by a principle, people remember the example but forget the principle. Another difficulty in understanding principles is that there are typically many different interpretations of a given relational abstraction. Thus, people may encode the principle in ways that are incompatible with the later example. The joint interpretation of an example and a principle helps overcome these limitations. People better understand the principle if they apply it to an example.

General Discussion

These studies show three learning advantages of analogical encoding. First, drawing comparisons facilitates acquiring an abstract schema. As Experiment 2 showed, it can do so better than studying a statement of the abstraction itself. Second, both studies replicated prior research showing that comparison facilitates applying derived abstractions to solve new problems. Third, analogical encoding of two current cases—that is, analogical encoding of a *probe*—leads to a retrieval advantage in accessing structurally matching cases from long-term memory. Our findings suggest that the second and third of these stem from the first: that it is the possession of clear schemas that facilitates both transfer and retrieval.

Implications for learning and transfer. Our findings have several implications for complex learning. On the dark side, the results of Experiment 1 suggest limitations on even experts' analogical thinking. Over half the participants failed to recall *any* structurally similar example from their own experience, despite the fact that they had considerable experience including specific experience in a particular kind of contingent contract (value billing) that would have qualified nicely. Prior studies suggest that people show more relational transfer in domains that are familiar or in which they possess expertise (Blanchette & Dunbar, 2001; Dunbar, 2001; Novick, 1988). However, as our results show, even for experts relational retrieval can be problematic.

A second rather gloomy finding is the failure of our (highly motivated) participants to spontaneously compare the two cases (Experiment 1) or the case and principle (Experiment 2). Here, as in our prior studies, participants in the Separate cases group almost never noticed the link between the two, despite the fact that they were on consecutive pages. The huge advantage found for the Comparison group, which did compare the two, makes this failure to notice the link all the more telling. It raises the question of how many potentially illuminating comparisons are missed in the course of learning. On the positive side, the relational fluency shown by the Comparison group offers a relatively simple technique whereby learners and educators can improve their understanding and gain relational insight.

Implications for memory retrieval. That analogical encoding can facilitate retrieval is consistent with the point that the match between a specific case and a general abstraction (which has few or no concrete features and therefore few mismatches) is better than the match between two specific cases (unless, of course, the cases are closely similar, with many matches and few mismatches) (Tversky, 1977). Further, it indicates that this advantage holds whether the schema is in the memory bank (as in prior studies of analogical transfer) or in the probe position.

The retrieval effect suggests that people can use a well-articulated principle to retrieve prior examples and reinterpret them as examples of this new abstract structure. This implies a clear mechanism by which reflection can reorganize knowledge. A major question in both child development and the field of expertise is whether and how people's existing knowledge changes as they understand a domain in new depth. To the extent that abstractions can call forth matching cases from memory, the learner may gain a richer understanding of the new abstraction and a re-representation of the prior example in light of the new abstraction. This suggests a means by which new knowledge can connect to existing knowledge and can reorganize that knowledge along more expert lines.

One encouraging implication of our findings is that examples people learn prior to understanding key abstract principles in a domain are not necessarily lost or wasted. Given the increasing demands for adults to learn, this is encouraging news. Teachers can capitalize on people's prior knowledge by encouraging people to recall familiar examples of new principles. We may well rely on learned cases every bit as much as researchers on analogy, categorization and case-based reasoning suggest, but we nonetheless may benefit considerably from interventions in how we encode them.

In conclusion, analogical encoding appears to be a powerful starting point for learning. The resulting understandings may radiate both backwards and forwards.

Acknowledgments

The research was supported by the Office of Naval Research, award number N00014-02-10078, to the first author, a fellowship from the Dispute Resolution Research Center to the second author, and a grant from the National Science Foundation, SES-9870892, to the third author.

References

Bassok, M. (1990). Transfer of domain-specific problem-solving procedures. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 522-533

Bazerman, M. H., & Gillespie, J. J. (1999). Betting on the future: The virtues of contingent contracts. *Harvard Business Review, 77*(5), 155–160.

Blanchette, I., & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition, 29*(5), 730-735.

Brooks, L. R, Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*. *120*(3), 278-287.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*(6), 1147-1156.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Chi, M. T. H., Feltovitch, P. J., Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 313-334). Cambridge, MA: MIT Press.

Forbus, K. D., & Gentner, D. (1986). Learning physical domains: Toward a theoretical framework. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2, pp. 311-348). Los Altos, CA: Kaufmann.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*(2), 141-205.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought*. (pp. 195-235) Cambridge, MA: MIT Press.

Gentner, D., Loewenstein, J., & Thompson L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2) 393–408.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspective on thought and language: Interrelations in development* (pp. 225-277). New York: Cambridge University Press.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability and inferential soundness. *Cognitive Psychology, 25*, 524-575.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*(3), 427-466.

Kolodner, J. L. (1993). *Case-Based reasoning*. San Mateo, CA: Morgan Kaufmann.

Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10*(4), 417-446.

Loewenstein, J., Thompson L., & Gentner, D. (2003). An examination of analogical learning in negotiation teams. *Academy of Management Learning and Education, 2*(2), 119-127.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*(4), 586-597.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189-223). Hillsdale, NJ: Erlbaum.

Novick, L. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 510-520.

Pirolli, P. L., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39*, 240-272.

Reed, S. K. (1987). A structure-mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 124-139.

Reeves, L. M., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115* (3), 381-400.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General, 122*(1), 92-114.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory & Cognition, 13*(4), 629-639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15*(3), 456-468.

Ross, B. H., & Kilbane, M. C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition 23*(2), 427-440.

Schank, R. C., & Riesbeck, C. K. (1981). *Inside computer understanding: Five programs plus miniatures*. Hillsdale, NJ: Erlbaum.

Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical learning improves case-based transfer. *Organizational Behavior and Human Decision Processes, 82*(1), 60-75.

Thompson, L., & Hastie, R. (1990). Social perception in negotiation. *Organizational Behavior & Human Decision Processes, 47*, 98-123.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Recognition effects and noncompensatory decision making strategies

Eric P. Gernaat (gernaat@msu.edu)
Bruce D. Burns (burnsbr@msu.edu)
Department of Psychology & Cognitive Science program
Michigan State University, East Lansing, MI 48824-1117 USA

Abstract

Oppenheimer (2003) challenged the empirical evidence for the recognition heuristic by pointing to the possibility that existing demonstrations may have confounded recognition with a person's existing knowledge. In two experiments we remove the possibility of such a confound by independently manipulating recognition in a way similar to the "overnight fame" paradigm of Jacoby, Kelly, Brown and Jasechko (1989). We found evidence for a recognition effect, but neither compensatory nor noncompensatory decision making strategies seem to be able to completely explain our results. We discuss what modification to these strategies may be necessary.

Introduction

Gigerenzer and Todd (1999) proposed that when people make decisions, rather than using all possible information that they could, they use "fast and frugal" heuristics selected from an adaptive toolbox. Gigerenzer and Goldstein (1996) showed that such heuristics can lead to decisions as good or better at achieving the organisms goals as more resource intensive strategies. This *adaptive rationality* approach challenges the traditional approaches to rationality (see Chater, Oaksford, Nakisa, & Redington, 2003).

Goldstein and Gigerenzer (2002) proposed that the toolbox includes the recognition heuristic, which can be applied "If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion." (p.76) In this sense it is an example of one-reason decision making, though it can also act as a subroutine in heuristics that use information beyond recognition, such as the *Take the Best* heuristic (Gigerenzer & Goldstein, 1996). Such heuristics embody Simon's (1956) concept of *bounded rationality*, which suggests organisms seek enough information for a good decision rather than spending resources on obtaining all relevant information.

However it is apparent from attempts to gather evidence for or against the recognition heuristic that there are some conceptual disagreements. In this paper we present our understanding of the heuristic and its relationship to other concepts of how recognition cues are used, and we present a new methodology for examining recognition as a process.

Previous studies

As evidence for people's use of the recognition heuristic, Goldstein and Gigerenzer (1999, 2002) showed that which German cities Americans recognized affected their choices for which of a pair of German cities was larger. However Oppenheimer (2003) challenged this empirical evidence for the recognition heuristic by suggested that because Goldstein and Gigerenzer used the 30 largest German cities, recognition was confounded with actual city size. Thus recognition itself may be irrelevant as a cue or combined with other cues. Such confounds exist in all their demonstrations because they state that the heuristic should only be found to be used when recognition has a correlation with the correct answer via an ecological valid mediator.

To remove this confound Oppenheimer (2003) presented a cities pairs task, in which recognition could not have a valid correlation because one city was fictional, and thus there was no correct answer. Furthermore the real cities, with which the fictional cities were paired, were either small or famous for reasons unrelated to size. Two studies tested the hypothesis that when subjects were presented with a city they recognized and one they did not, then preference for the recognized city would not be above chance (specifically, he tested if more than 50% of participants chose the recognized city). Both studies found that significantly fewer than 50% of participants chose the recognized cities.

Oppenheimer (2003) suggests that Goldstein and Gigerenzer's (2002) evidence showing that greater than 50% of subjects choose the city they recognized is consistent with them using the recognition heuristic. However because other cues may be confounded with recognition, these results are also consistent with strategies that combine cues together. For examples these other cues could be combined in tallying strategies, weighted additive models, or regression models and produce results consistent with the recognition heuristic. Oppenheimer claims that by demonstrating that cues that should be associated with recognition (cues indicating small in his experiments, those indicating big in Goldstein & Gigerenzer) influences the extent to which subjects chose the recognized city, he has shown that there is no evidence for the recognition heuristic. This interpretation of the data is disputable, partly because exactly what question this data addresses is arguable.

Distinguishing related questions

Experiments on the recognition heuristic have addressed three questions that are related but not necessarily the same:

1) Are there recognition effects on choice? Demonstrating this requires showing that people's choices are influenced by recognizing one option and not the other. This question generates a testable prediction that Oppenheimer's (2003)

data addresses: that subjects should prefer the recognized city at a rate greater than chance. He finds no evidence for this, and claims that Goldstein and Gigerenzer's (2002) evidence for it was due to confounds. However this question is not equivalent to the second.

2) Do people use the recognition heuristic? Oppenheimer (2003) states that "The recognition heuristic posits that individuals will use no information aside from mere recognition to make city size estimations." Goldstein and Gigerenzer almost certainly disagree with this statement. Gigerenzer and Todd (1999, p.32) state that different heuristics may be applied to a choice between two options, and which is applied depends on the person's knowledge. The recognition heuristic is most likely applied when the decider has no information except recognition. *When* it is applied then only recognition information should be used, but there is no claim in Goldstein and Gigerenzer (2002) that the recognition heuristic is always applied. Their own data show that the recognition heuristic is not always applied. Gigerenzer and Todd (1999, p. 32-33) briefly consider the question of how heuristics are selected and suggest that task and available cues determine it. Although their concept of a "toolbox" of heuristics implies that selection is critical, a valid criticism may be that how this is done is an under-developed aspect of their approach. In this they are not unusual as Falk and Konold (1997, p. 305) point out that it hard to make predictions regarding the representatives heuristic because "there is no established procedure for deducing how it will be implemented in a specific task."

3) Is recognition used in a noncompensatory way? Goldstein and Gigerenzer (2002) propose that the recognition heuristic is a noncompensatory algorithm, as are other algorithms proposed by Gigerenzer and Todd (1999). A prototypical noncompensatory algorithms is a lexicographic strategy (see Payne, Bettman, & Johnson, 1993). Cues are arranged in a hierarchy based on validity then starting from the top, one cue at a time is considered until one is found that discriminates between the options. In contrast, a compensatory strategy (such as linear regression or multi-additive models) integrates all cues in making a choice, although they may be given different weightings. A flow diagram of a noncompensatory algorithm in which recognition is the first cue considered is given by Gigerenzer and Goldstein (1996, Figure 2). The stopping rule for such an algorithm is the first cue considered that distinguishes between options. Characteristic of such algorithms is that a decision made on the basis of a higher cue cannot be reversed by cues lower in the order Goldstein and Gigerenzer (p. 82). The recognition heursitc is a special case of a noncompenstory strategy, so recognition need not be the first cue considered in noncompensatory strategies. Gigerenzer and Goldstein and Gigerenzer and Todd discuss various noncompensatory algorithms.

Possibly Oppenheimer's (2003) most interesting claim is that his data showing that people chose the recognized city less than 50% of the time shows that recognition is not used in a noncompensatory way. It is worth examining his arguments for this claim. He argues that if people are only sometimes using the recognition heuristics, then the problematic question of how they decide which heuristic to use is raised. If the task determines it then why was a recognition effect in Goldstein and Gigerenzer's (2002) paired-city task, but not Oppenheimer's (2003)? If the cues presented in the task determine which algorithm is used, then all cues have to be considered, undermining the efficiency of the "one-reason" decision making. Thus it is not as "frugal" in terms of processing and information as Goldstein and Gigerenzer claim.

If more than one cue is available then implicit in the claim that cues are ordered in a hierarchy is that cues must be considered at some level. The defining characteristic of a noncompensatory algorithm seems to be how the cues are combined at the decision point. In compensatory algorithms all available cues are integrated requiring some form of trade-off between those that favor different options, whereas in noncompensatory algorithms at the point of decision only one cue is considered and there are no trade-offs. Whether this means that noncompensatory algorithms are inherently more "frugal" than compensatory algorithms depends on what sort of assumptions are made about cue ordering or cue trade-off processes.

If competing subjective cue validities determine the order in which cues are considered by noncompensatory algorithms, then it would be predicted that the nature of the cues available in addition to recognition would influence the extent to which recognition determined the choice between two options. Thus Oppenheimer's (2003) evidence that other cues may moderate the impact of differential recognition on a choice does not in itself show that recognition is used in a compensatory algorithm.

In both Goldstein and Gigerenzer (2002) and Oppenheimer (2003) recognition was confounded with other information. To more effectively test how recognition is used requires a way to manipulate recognition free of any confounds.

Manipulating recognition

In existing studies of the heuristic, recognition has always been a pseudo-independent variable; that is, the studies have not manipulated recognition but instead examined the impact of what subjects recognized due to life experience.

We manipulated recognition by pursuing Goldstein and Gigerenzer's (2002) suggestion that recognition might be induced in a way analogous to the "overnight fame" effect found by Jacoby, Kelly, Brown and Jasechko (1989). Jacoby, et al. presented participants with a list of unfamiliar names that were included on a later list of names to which participants were asked to respond: famous or nonfamous? They were more likely to choose as famous the arbitrary names from the initial list. We used a similar methodology with small German cities to examine if induced recognition could affect the choice of which of two cities was larger.

To induce recognition we gave participants one of two *induction lists* consisting of eight German cities or towns (Zwickau, Leverkusen, Regensburg, Offenbach, Ulm, Stralsund, Coburg, Dormagen; or, Bochum, Gelsenkirchen, Darmstadt, Krefeld, Schweinfurt, Goslar, Lingen, Iserlohn) plus three irrelevant small cities. Participants were asked to count the number of syllables they thought each city name

contained. We conducted a pilot experiment to confirm that these induction lists could induce recognition. Eight members of the Michigan State University participant pool were given one of the induction lists, and nine the other list. After spending 30 minutes on unrelated tasks, participants were presented with a list of 20 small German cities, including the 16 from the two induction lists. For each city they were asked if they remembered seeing the city on the list they were given earlier. They recognized more cities from their own induction list (M = 85%) than from the list they had not been given (M = 11%), $t(32) = 18.62$, $p < .001$. Therefore the pilot experiment supported our assumption that presenting the syllable list would induce later recognition.

Aims of the experiments

In two experiments using induced recognition we examined recognition effects and how other cues may moderate any effect. This allowed us to address two of the three questions we outlined regarding recognition

1) Are there recognition effects on choice? Oppenheimer (2003) claims that the Goldstein and Gigerenzer's (2002) results are flawed because of confounds, but his experiments intentionally introduce another confound. To clearly establish whether there are recognition effects in the cities-pairs task requires a version with no confounds.

2) Do people use the recognition heuristic? The answer to this question is clearly "not all the time". A more reasonable question is what factors influence the degree to which people use the recognition heuristic? However no one has specified these factors well enough to examine this.

3) Is recognition used in a noncompensatory way? Addressing this question requires examining how other cues moderate the effect of recognition. Goldstein and Gigerenzer (2002) found that adding other cues did not affect the size of the recognition effect, whereas Oppenheimer found that they did. However in neither experiment were cues systematically manipulated within a single experiment. In our experiments we varied the amount of other cues available in order to address how information moderates recognition effects and thus throw light on the issue of whether recognition is used in a compensatory or noncompensatory way.

Experiment 1

In Experiment 1 we induced in Americans recognition of small German cities from an induction list of eight, then presented them with each of these cities paired with one from the alternative list of eight cities. We predicted that participants would be more likely to select the induced city in the pair. We also manipulated giving a positive (has a major league soccer team) or a neutral (the state the city was in) cue to the size of one city in the pair, either the induced or noninduced city. The least amount of extra cues was none, the next least neutral only, the next positive only, and the most extra cues received was both neutral and positive. If recognition is a compensatory cue, then the more information presented, the smaller the effect of recognition should be.

Method

Participants

A total of 256 members of the Michigan State University participant pool participated for partial course credit.

Materials

The two induction lists were the same as those described for the pilot experiment. For the *city-size task*, the eight cities from one induction list were paired with the eight from the other. Thus for each pair, one city was induced (i.e., appeared on the participant's induction list) and one was noninduced (i.e., appeared on the list the participant did not see). For each pair, participants chose which city they thought was larger. To obscure the purpose of the task, Oppenheimer (2003) included some pairs with obvious answers. Similarly we added three pairs containing a well-known large city (Berlin, Frankfurt, Munich).

With each city pair extra cues could appear, though cues were always independent of reality allowing us to freely manipulate information. (No participants pointed out that the information was incorrect.) There were four cue conditions. For the two *cue-none* pairs it was stated for both cities that there was no information available regarding whether it had a soccer team or which state it was in (Germany consists of 16 states). For the two *cue-neutral* pairs soccer team status was said to be unknown but a state was given for one city. For the two *cue-positive* pairs state was said to be unknown but one city was said to have a soccer team. For the two *cue-all* pairs for one city a state was given and it was said to have a soccer team, whereas this information was said to be unknown for the other city. Eight pairs for each participant were necessary because for four pairs the cue conditions were applied to the induced city, and for the other four pairs the cue conditions were applied to the noninduced city.

Note that Goldstein and Gigerenzer (2002) also conducted an experiment in which a city having a major soccer team was used as an extra cue to city size. They taught participants that having a team was indicative of being a large city, but found that it did not appear to alter the effect of recognition based on a between experiments comparison. We did not teach participants about this relationship but instead relied on participants generalizing from their likely awareness that the major American professional sports teams are rarely in small cities.

The same cities were always paired but each pair appeared equally often in each cue and induced condition. Cues appeared in two possible orders (maximum information to none, or vice-versa), and the cities appeared in two different orders. Thus there were 16 versions of the city-size task (four cue by two cue-order by two city-order conditions) that appeared equally often with each induction list.

Procedure

Participants were first given one of the two induction lists and told that their task was to write down the number of syllables in each city name as best they could. They then

spent 30 minutes doing unrelated tasks until they were given one version of the city-size task and asked to choose which city in each pair they thought was larger. After completing this task they were given a questionnaire that asked them: 1) Does the last task you completed relate to any of the previous tasks? 2) Were your responses to the last task you completed affected by any of the previous tasks? For each question they had to answer "yes" or "no" and to explain their answer. These questions were used to determine if participants were explicitly connecting together the induction and city size tasks. Participants who said the syllable task affected them when choosing cities were replaced in the design. This was to protect against the possibility that participants might just select cities because they thought they were supposed to select cities from the induction list. Fewer than 10% of participants were replaced and subsequent analyses found no evidence that the replaced subjects responded differently.

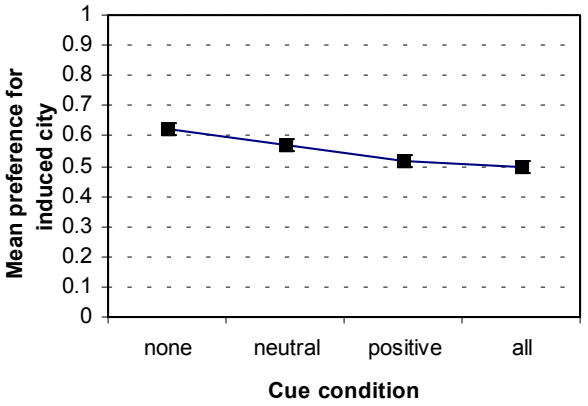


Figure 1: Mean preference for the induced city for each cue level (Bars represent one standard error). Above .50 represent preference towards induced cities.

Results

Figure 1 indicates mean preferences for the induced cities for each cue level. To calculate preferences we examined the two city pairs in each of the cue conditions. If a participant chose the induced city in both pairs (i.e., the pair in which cues were given for the induced city and the pair in which cues were given for the noninduced city) they were assigned 1.0, if they chose the induced city once and the noninduced once then they were assigned 0.5, if they chose the noninduced city both times then they were assigned 0.0. Thus mean preferences range from 0.0 to 1.0. Preferences above .50 indicate a bias towards induced cities, whereas below .50 indicates a bias towards noninduced cities.

Participants had preferences for the induced city above .50 for the cue-none (.62, $t[255] = 5.72$, $p < .001$) and cue-neutral (.57, $t[255] = 3.78$, $p < .001$) conditions, but not for the cue-positive (.52, $t[255] = 1.14$, $p = .258$) or cue-both conditions (.50, $t[255] = -0.14$, $p = .887$). Ordering the four cue condition in terms of amount of extras information given (none, neutral, positive, all) produced a significant linear trend, $t(765) = 5.38$, $p < .001$. Separately calculating the proportion of participants choosing the city for which cues were given, but ignoring whether this was the induced or noninduced city, we found a strong effect of how much information participants were given (cue-none .50; cue-neutral .58; cue-positive .79; cue-all .84), $F(3, 576) = 62.8$, $p < .001$.

Of the 256 participants, 44 indicated that they thought the city size task related to the syllable task, but they were not eliminated from the sample because they said that the syllable task did not affected their responses. However it is possible that effects were driven by just these participants who explicitly remembered the syllable task. Analyzing separately these 44 participants and the other 212, we found the same pattern of results for both groups as we found for the whole sample.

Discussion

When recognition was not confounded with any other information and it was the only cue available, then the city likely to be recognized was selected more than chance. The 62% selection rate for the cue-none condition was not as high as Goldstein and Gigerenzer (2002) report but there could be several reasons for this. First, we did not test which cities participants recognized and thus we did not only analyze pairs for which only one city was recognized, as Goldstein and Gigerenzer and Oppenheimer (2003) did. Based on the recognition pilot data, it could be estimated that this condition would apply for only 77% of our pairs. (To avoid any possible biases, we analyzed all pairs.) Second, our induction procedure may produce recognition that is relatively weak and uncertain compared to that based on experience. Goldstein and Gigerenzer treat recognition as a binary, all-or-none distinction; however creating this distinction is not without error. As the work on eyewitness testimony has amply shown (Wells & Loftus, 1984), recognition and can be uncertain and it may be hard for a person to decide if they recognize something. This may be especially true for foreign, hard to pronounce words briefly experienced once. Our induction procedure allowed us to manipulate recognition, but it may produce recognition with a different character to that in the previous experiments.

The results also showed that the presence of other cues can moderate the influence of recognition. Oppenheimer (2003) suggested that any evidence of such moderation is evidence against noncompensatory decision making of the type Goldstein and Gigerenzer (2002) described. However evidence that one cue moderates the mean effect of another cue does not establish that individuals are integrating cues as in compensatory multi-cue strategies. Our data may be consistent with Goldstein and Gigerenzer's approach in which only one cue is applied at a time but cues form a hierarchy based on validity. When recognition is uncertain other information may easily be seen as more valid.

Of course there is no true validity for either recognition or other information in this experiment, as relaxing that constraint is what allows us to freely manipulate these factors. We are assuming that participants come to the experiment with cue validities, thus it is not surprising if there may be individual differences in how they assign

validity. There may also be individual differences in strategy selection (Fasolo, Miscuraca, & McClelland, 2003).

Experiment 2

One way of investigating how recognition is combined (or not) with other cues is to equalize the amount of extra information given for each city in a pair. Different compensatory decision strategies differ in the way in which they combine cues, but the impact of a cue in these strategies is greatest when that cue differentiates between options. Thus cues that provide equivalent information for both choices should have no impact on decision making. Therefore the moderating effects of extra cues on recognition should largely disappear.

As Gigerenzer and Goldstein (1996) describe noncompensatory algorithms, cues should continue to be considered until one is found that differentiates the options. This stopping rule implies that only differential cues could moderate the effect of recognition because even if a nondifferential cue is higher in the validity hierarchy, it can have no impact on the final decision. Thus a noncompensatory strategy also predicts that when extra cues are not differential then the moderating effect of those cues on the impact of recognition should disappear.

Experiment 2 varied the same cues as in Experiment 1, but instead of giving cues about just one of the two cities, we gave equivalent cues for both. Thus the extra cues never distinguished the two cities, allowing us to test if undifferentiating information still moderated the recognition effect. Both compensatory and noncompensatory approaches make the same prediction: Whatever recognition effect is found for the cue-none condition, the same recognition effect should be found for other cue conditions. If this prediction is violated and cue condition still moderates the recognition effect, then it can point to modifications necessary to these approaches.

Method

Participants

A total of 128 members of the Michigan State University participant pool participated for partial course credit.

Materials and procedure

The exact same procedures and induction lists were used as in Experiment 1. The city-size task was identical except that equivalent cues were given for *both* members of the eight pairs. Thus for the two pairs in the *cue-all* condition a state was given for both cities and both cities were said to have a major league soccer team. For the two pairs in the *cue-positive* condition it was stated that both cities had soccer teams and that the states were not known. For the two pairs in the *cue-neutral* condition a state was given for both cities and it was stated that whether or not these cities had a soccer team was unknown. For the two pairs in the *cue-none* condition it was stated that neither piece of information was known. Thus the design of Experiment 2 was simpler than Experiment 1 but eight pairs were still given. We again counterbalanced the four cue conditions across the city pairs, used two orders of presentation, and two city-orders, which yielded sixteen version of the city task. Each version was presented equally often with each induction list.

Results & Discussion

Figure 2 presents the mean proportions of participants in each information condition who selected the induced city. In the *cue-none* condition the rate of choosing the induced city (.63) was significantly above .50, $t(127) = 4.29$, $p < .001$, but this was not the case for the *cue-all* condition (.54), $t(127) = 1.37$, $p = .175$. Proportions above 50% were almost statistically significant for both the *cue-neutral* (.56), $t(127) = 1.78$, $p = .075$, and *cue-positive* (.56) conditions, $t(127) = 1.91$, $p = .058$. Thus this experiment again found a clear effect of induced recognition when no prior knowledge could be confounded with recognition, especially when no new information about the cities was given.

Overall, there was a significant linear trend for cue condition on the proportions of induced cities selected, $t(381) = 2.15$, $p = .033$. Just as in Experiment 1, giving participants other cues reduced the impact of recognition. However it appears that the volume of information was critical, as this information did not differentiate the options.

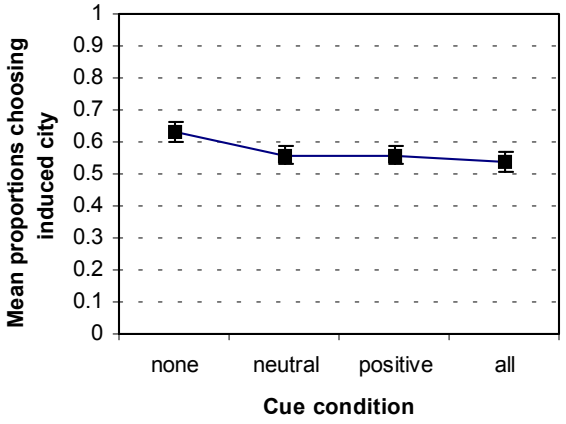


Figure 2: Mean proportions of participant in each cue condition choosing the induced city rather than the noninduced city (Bars represent one standard error). Above .50 indicates a preference for the induced city.

General Discussion

Our results show that recognition effects of the type Goldstein and Gigerenzer (2002) found are not only due to recognition being confounded with other cues. When recognition was unconfounded and was the only available cue the city more likely to be recognized was selected at a rate above chance.

Using within experiment manipulations our results supported Oppenheimer's (2003) finding that the presence of other cues could moderate a recognition effect. However

the results of Experiment 1 are consistent with recognition being part of either compensatory or noncompensatory strategies, as long as recognition is not always the most valid cue. In contrast, the results of Experiment 2 are consistent with the predictions of neither type of strategy, as nondifferential cues should have no impact on decision making. What assumptions would need to be added to these approaches in order for them to explain these results?

Whatever function a compensatory strategy applies to options, the same function would be applied to both cities in a pair. Thus providing the same cue for both cities should just add a constant to the evaluation of both cities. Thus the difference between evaluations will not change. One way to deal with the evidence from Experiment 2 would be to make the total amount of information part of the integration process. Perhaps there is a Weber function for comparing options like there is for comparing perceptual stimuli. This would represent a revision to current compensatory strategies that would yield testable predictions.

Nondifferential cues should be ignored by noncompensatory strategies. However, they could have an impact if the algorithm had a stopping rule that may stop evaluating cues before one is found that differentiates. It seems consistent with bounded rationality that sometimes the evaluation is made by an organism that there is little value in continuing looking for discriminating information. Thus as each nondifferentiating cue is examined there may be a nonzero probability that the organism will decide to stop. If the extra cues may be placed higher in the validity hierarchy than recognition (as Experiment 1 may suggest) then such a stopping rule could lead nondifferential cues to moderate the utilization of differential recognition.

Such a stopping rule might take into account the anticipated cost of evaluating information. This may be particular relevant in our paradigm as the type of uncertain recognition we might have induced may have a cost. Anderson's ACT-R framework emphasizes the cost of any remembering (Anderson, Lebiere, & Lovett, 1998). Recognition is not necessarily more accurate than recall (Tulving & Thomson, 1973), and it may be inaccurate and uncertain (Wells & Loftus, 1984). Thus even a single nondifferential cue may reduce the impact of recognition if it can be higher in the cue hierarchy.

Experiments that examine recognition free of confounds are useful for understanding this heuristic as a processes rather than just a phenomenon, and the importance of doing this in general was pointed to by Gigerenzer (1996). Both the compensatory and noncompensatory approaches to how multiple cues affect decision-making seem under-specified, and thus unable to explain the results of Experiment 2. The paradigm we introduced here has promise for furthering the understanding of heuristics utilizing recognition and the process by which recognition affects choices.

Acknowledgements

We would like to thank Erik Altmann and Tom Carr for helpful comments on earlier drafts. These experiments were done as partial fulfillment of the first author's requirements for the degree of Bachelor of Science (Honors).

References

Anderson, J. R., Lebiere, C., & Lovett, M. (1998). Performance. In J. R. Anderson, & C. Lebiere (Eds.), *The atomic components of thought.* Mahwah, NJ: Lawrence Erlbaum Associates.

Chater, N., Oaksford, M., Nakisa, N., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90,* 63-86.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

Fasolo, B., Miscuraca, R., & McClelland, G. H. (2003). Individual differences in adaptive choice strategies. *Research in Economics, 57*, 219-233.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, *103*, 592-596.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*, 650-669.

Gigerenzer, G. & Todd, P. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, & P. Todd (Eds.), *Simple heuristics that make us smart* (pp. 37-58). New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, & P. Todd (Eds.), *Simple heuristics that make us smart* (pp. 37-58). New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75-90.

Jacoby, L. L., Kelly, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56*, 326-338.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best." *Organizational Behavior and Human Decision Processes, 91,* 82-96.

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition, 90,* B1-B9.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge, UK: Cambridge University Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 359-380.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129-138.

Wells, G. L., & Loftus, E. F. (Eds.) (1984). *Eyewittness testimony: Psychological perspectives.* New York: Cambridge University Press.

Interpersonality: Individual differences and interpersonal priming

Alastair J. Gill (A.Gill@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Annabel J. Harrison (annabelh@cogsci.ed.ac.uk)
School of Philosophy, Psychology, and Language Sciences, University of Edinburgh
7 George Square, Edinburgh, EH8 9JZ UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

We study how Extraversion and Neuroticism influence people's language production in interpersonal interactive situations. A priming study used confederate priming methodology to investigate syntactic priming behaviour. We expected that Extravert sociability would be related to the strength of priming effects, although Neurotic emotionality might also have an effect. Results indicate that Extraversion has no effect, but Neuroticism does have an effect. We discuss possible reasons and suggest further experimentation to investigate this finding. Implications and applications of this work are outlined.

Personality and interaction

Individuals differ in the way they speak and write. Some of those differences are systematic, and can be attributed to apparently deeper differences, such as personality traits, like Extraversion and Neuroticism (or Emotional Stability). Level of Extraversion is intuitively related to sociability and communication, and this is expressed through interpersonal behaviour. However, level of Neuroticism appears to be more related to anxiety and inward focus, and thus having greater influence on solo behavior. In the past, it has been found that both these personality traits do significantly influence an individual's language production behaviour in a variety of contexts (Pennebaker and King, 1999; Dewaele and Furnham, 1999). Recent work has investigated e-mail text, and suggested that even in that genre, there are characteristic sequences of words associated with each end (High or Low) of both dimensions (Extravert or Neurotic) (Gill and Oberlander, 2002, 2003b).

The majority of work on the relations between personality and language production has studied monologue only. Yet most everyday language occurs in the context of interpersonal interaction. So here, we aim to investigate the role of personality upon language use in a dialogue setting.

Studies of conversational behaviour have demonstrated that individuals align with their interlocutors on a number of levels (Pickering and Garrod, in press). The phenomena have been examined from both social and cognitive perspectives. On the social side, a key focus of interest is cooperation and audience design. On the cognitive side, a key focus is coordination and interpersonal priming.

For example, sociolinguistic studies have shown that speakers adopt accent or dialectal variation or a level of lexical density appropriate to their audience. This variation operates at phonological, lexical, and syntactic levels (Labov, 1972; Coupland, 1980; Bell, 1984; Bradac and Wisegarver, 1984). Audience design is regarded as a relatively conscious process over which the speaker has a certain amount of control. It may be a result of co-operativity, affiliation, or willingness to take another's perspective (Haywood, Pickering, and Branigan, 2003).

By contrast, from a cognitive perspective, coordination is viewed as an artifact of the underlying language production mechanisms. For example, it has been argued that references from the comprehension system are recycled to provide output for the production system (Pickering and Garrod, in press). Alignment is found at the lexical level (Brennan and Clark, 1996; Branigan, Pickering, and Cleland, 2000), the conceptual level (Garrod and Doherty, 1987), and the syntactic level (Pickering and Branigan, 1998). Unlike cooperation, such coordination is considered to be largely subconscious.

Coordination therefore provides a more direct insight into underlying processing abilities, and is less prone to outside influence. In approaching the study of personality in dialogue, we therefore use an interpersonal priming paradigm. At the outset, our question is very general: Can differences in interpersonal priming be attributed to personality?

To make this question more specific—and to attempt to answer it—the rest of this paper is structured as follows. First, we introduce a little more background on personality theory. Then, we frame a possible explanation of recent findings on the relations between Extraversion, Neuroticism and language production; this leads to two hypotheses concerning the possible relation between personality and interpersonal priming. We then present the priming experiment which tested these hypotheses. The results were somewhat unexpected, and we conclude by discussing their implications.

Overview

There are a number of approaches to personality (Matthews and Deary, 1998). Two of the most prominent trait theories are the five factor model (Costa and McCrae, 1992), and Eysenck's three-factor PEN model (Eysenck, Eysenck, and Barrett, 1985; Eysenck and Eysenck, 1991). These agree that two main factors are Extraversion (sociability) and Neuroticism (emotional stability). The Five Factor Model sees three further dimensions: Conscientiousness, Agreeableness and Openness; PEN arguably conflates these into one dimension, Psychoticism (tough mindedness). In what follows, we focus on the first two dimensions, common to both models.

The traits can be summarised thus: A typical Extravert tends to be sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert (Low Extravert) is quiet, retiring, reserved, plans ahead, and dislikes excitement; A typical High Neurotic tends to be an anxious, worrying, moody individual. A typical Low Neurotic tends be calm, even-tempered and relaxed (Eysenck and Eysenck, 1991).

Personality and language

Work on personality and language behaviour has studied a range of features. For instance, Extraverts are regarded as talking louder (Scherer, 1978), demonstrating a higher speech rate (Siegman, 1987), and they show less hesitation, but make a higher proportion of semantic errors (Dewaele and Furnham, 2000). At a grammatical level, Extraverts use greater proportions of pronouns, adverbs, verbs (Cope, 1969), which contrasts with the more explicit language of the Introverts and their increased use of nouns, modifiers and prepositions (Dewaele and Furnham, 2000). Additionally, Extraverts demonstrate lower lexical richness in formal situations (Dewaele and Furnham, 2000), whilst analysis of informal e-mail communication has shown highly Neurotic language to be more repetitious (Gill, 2003; Gill and Oberlander, 2003b). At a more content-oriented level, Pennebaker and King (1999), using the Linguistic Inquiry and Word Count text analysis program, showed that broad psychological language categories are related to dimensions of personality variation. For example, they found that when writing about thoughts and feelings, high Neurotics use more negative emotion words and fewer positive emotion words.

However, our interest here is on interaction: dialogue and conversation. Studies using speech act coding have found that Introverts used more hedges and problem talk, namely expressing qualification, and dissatisfaction with one's own activities, while Extraverts expressed more pleasure talk, agreement, and compliments, with content focusing more on extracurricular activities (Thorne, 1987). Extraverts have also been shown to use more self-referent statements, and initiate more laughter (Gifford and Hine, 1994). Gifford and Hine also found that Extraverts talk more, with other studies finding that they use a greater total number of words (Campbell and Rushton, 1978; Carment, Miles, and Cervin, 1965). As would be expected, Extraverts show greater desire to initiate interactions (McCroskey and Richmond, 1990), even in computer-mediated communication (Yellen, Winniford, and Sanford, 1995). Also, Dewaele (2002) finds that in L3 English production, Extraversion (and also Psychoticism) showed a strong negative relationship to communicative anxiety, whilst Neuroticism showed a positive relationship.

Studies investigating hemispheric asymmetry provide a further perspective on this area, for example, Davidson (2001) proposes the relationship between Extraversion and positive affect with approach behaviours, and Neuroticism and negative affect and withrawal behaviours. In the following hypotheses, we explore the implications of personality, affect and approach/withdrawal on priming behaviour.

Hypotheses for interpersonal priming

The likelihood of priming may be affected by the tendency to approach or the tendency to withdraw—or by both.

If Extraversion is associated with approach behaviours, it is natural to expect that higher Extraversion will lead to "more approach", and that this might mean that an individual will coordinate more with their interlocutor. Furthermore, the Extravert's higher drive to gain or retain the conversational floor will mean that less effort can be directed towards detailed language planning. Hence, if their partner has made a lexical or syntactic choice, the High Extravert is likely to re-use that choice, rather than explicitly planning a new one (cf. Gill and Oberlander, 2003a).

If Neuroticism is associated with withdrawal behaviours, it could well be that high levels of this trait result in "more withdrawal" and lower engagement with the interlocutor. Furthermore, the inward (worrying) focus of a High Neurotic might mean that more resources are devoted to inner thought, and fewer to interaction with the environment. Thus, we might expect that such an individual will coordinate less with their interlocutor.

Thus, there is a clear prediction for Extraversion, and a slightly more complex picture for Neuroticism. Of course, it could be that neither Extraversion nor Neuroticism have any effect on coordination or priming.

Method

In syntactic priming, a particular syntactic structure is more likely to be produced given prior exposure to the same structure (Schenkein, 1980). This

phenomenon has been replicated under experimental conditions when speakers say, hear, or read sentences (e.g., Bock, 1986; Pickering and Branigan, 1998; Corley and Scheepers, 2002). Bock and colleagues found that people tended to repeat the active or passive form of a sentence they had just read in describing an unrelated picture (Bock, 1986; Bock, Loebell, and Morey, 1992). In this study we employ the confederate priming method (Pickering and Branigan, 1998): The subject of the experiment takes part in a dialogue game along with a confederate of the experimenter. The game involves matching and describing pictures. Both participants apparently have the same two tasks: to describe a set of pictures so that the other participant can match them, and to verify whether the descriptions that they hear match the picture that they see. However, the confederate's descriptions are scripted.

Participants

Forty University of Edinburgh students who were self-declared native speakers of English were paid to participate in this study. Personality information derived from the NEO-PI questionnaire is as follows: Extraversion M = 51.75 (SD = 12.82), and Neuroticism M = 54.18 (SD = 12.72).

Materials and Design

We prepared two sets of pictures depicting actions. Each set included 12 pictures depicting transitive actions involving an agent and a patient. The entities depicted were chosen to be easily recognisable and nameable. There were two pictures for each of 12 transitive verbs (*bite, chase, dust, hit, kick, lift, poke, pull, push, shoot, touch, weigh*). These 24 pictures comprised the set of targets. The remaining 120 pictures in each set depicted intransitive actions. There were several pictures for each of 20 intransitive verbs. These comprised the filler pictures.

The appropriate verb was printed under each action. Each set of pictures depicted the same range of entities and actions. However, the pairing of entities with actions was different.

We term one set the Subject's Description Set and the other set the Confederate's Description Set. We created ordered pairs of prime and target pictures by pairing each description of a transitive action from the Confederate's Description Set (the prime) with a picture depicting a transitive action from the Subject's Description Set (the target picture).

Half of the prime sentences were assigned active descriptions of the form 'the X verbing the Y', and half were assigned passive descriptions of the form 'the Y being verbed by the X'. An experimental item was defined as the confederate's scripted description of a prime picture plus the subject's target picture paired with it. There were thus two versions of each item: active confederate description and passive confederate description.

We constructed four lists containing 24 experimental items and 120 subject fillers. The confederate fillers were randomly distributed in the remaining gaps. The entities depicted in the target picture were not present in the immediately preceding block (prime plus subject fillers and confederate fillers). The verb also differed between prime and target. Each picture was assigned to either the match or the mismatch condition for the matching task. For the latter, we assigned another picture depicting a different entity doing the same action (thus using the same verb) was assigned. Each list contained 12 experimental items with active prime descriptions and 12 with passive prime descriptions. Exactly one version of each item appeared in each list. Hence, Prime Type (active vs. passive) was manipulated within subjects and items. The dependent measure was the proportion of descriptions of target pictures produced with a passive structure.

Procedure

The Subject's Description Set was presented to the subject via a computer program.The order of the pictures was randomised for each subject, with between four and eight filler items intervening between each experimental item. A divider prevented the subject from seeing the confederate or his computer screen. The experimenter told the subject and the confederate that the experiment was investigating how well people communicate when they cannot see each other. Their tasks were alternately to describe the pictures to the other participant, and to match their picture to the other participant's descriptions. When it was the subject's turn to match, the confederate would see a sentence appear on his screen which he would read aloud and then press space bar, at which point a picture would appear on the subject's screen. The subject was instructed to say "yes" or "no" (or ask for repetition) and to press the Z key for "no" and the M key for "yes" according to whether the picture matched or mismatched the description. When it was the subject's turn to describe, a picture would appear on the subject's screen and the confederate would say "yes" or "no" (or ask for repetition) and press the Z key or the M key according to whether the picture on his screen matched or mismatched the description. Throughout the session, the experimenter and confederate acted as if the confederate was a genuine subject (e.g., the confederate asked questions about the task). Before the experiment, there was a practice session with two filler items each, after which the subject could ask for clarification if necessary. The confederate also gave the first description. Hence the confederate's description of a prime always immediately preceded the subject's description of a target. Both dialogue participants wore a lapel microphone. The experimental session was recorded on audio tape and subsequently transcribed.

Table 1: Proportion of Passive target responses after active and passive primes and degree of priming

Group	Nos.	PP	AP	Priming
Low E	8	.1363	.0300	10.6
Mid E	27	.2015	.0270	17.5
High E	5	.1500	.0480	10.0
Low N	5	.1160	.0480	6.8
Mid N	28	.2271	.0261	20.1
High N	7	.0486	.0343	1.4
Total	40	.1820	.0302	15.2

We coded the first response that the subject produced; 3 target responses that described the agent as the patient and the patient as the agent were excluded. We coded the remaining target 957 responses as passive if the patient was described as being verbed by the agent and as active if the agent of the action was described as verbing the patient.

An analysis of variance (ANOVA) was conducted, with prime type (active vs. passive) as a within subjects factor and Neuroticism (Low [> -1 s.d. of the mean], Mid [< 1 s.d. of the mean], High [$> +1$ s.d. of the mean]) as a between subjects factor.

Results

Proportions of passive target responses following passive and active primes are reported in Table 1; these are described by personality type of participant, and also for the group overall. Here we can see that in both cases the Mid groups appear to show greater priming. However the High and Low Neurotic groups appear to show even lower levels of priming than for Extraversion.

Turning now to our analysis of variance, and here the ANOVA revealed a significant effect of prime type (active vs. passive) on the proportion of passive forms used (F_1 (1,37) = 6.63; $p < 0.05$; F_2 (1,23) = 97.01; $p < 0.05$).

A significant interaction was found between Neuroticism (Low, Mid or High) and prime type (F_1 (1,37) = 3.68; $p < 0.05$). Post-hoc Tukey tests revealed that both the High N and Low N groups primed significantly less than the Mid N group ($p < 0.05$). No interaction was found between Extraversion and prime (F_1 (1,37) = 0.60; $p > 0.1$).

Discussion

We found a reliable effect of syntactic priming of active and passive structures in a dialogue task. This confirms our expectations and replicates previous syntactic priming found in dialogue (e.g., Pickering and Branigan, 1998) and with active vs. passive forms (e.g., Bock, 1986).

Additionally, our results demonstrate that Neuroticism is related to the degree of syntactic priming for passive constructions; Extraversion is not. We now relate these results to our hypotheses. For Extraversion, we proposed that higher levels of Extraversion would lead to an increase in priming. Here we found that the Mid group primed more, however this result was not significantly different to that of the Low and High groups. In this case we therefore accept the null hypothesis that Extraversion is not related to levels of priming. For Neuroticism, we find that the Low and High groups primed significantly less than the Mid group. Comparing this result directly with our Neuroticism hypothesis creates a tension: We proposed that the High group would less likely to prime due to an inward focus and thus withdrawal from their partner. To address these findings, we therefore reframe our Neuroticism hypothesis as follows: as before, we claim that the High group are less likely to prime due to inward focus, but that the Low group are also less likely to prime, since they are less concerned with monitoring themselves in relation to their interlocutor. In this case—as in our results—the extreme High and Low levels of the trait have an inhibitory effect on priming, and the Mid trait levels represent a facilitating effect.

We acknowledge that such explanation is relatively speculative, and further experimentation will be required to test this hypothesis. For example, the NEO-PI questionnaire divides Neuroticism into 6 facets: anxiety, angry hostility, depression, self-consciousness, impulsivity, vulnerability. It may be that these may relate more specifically to withdrawal or threat-monitoring, in which case these could be related to the priming information. However, we expect that a larger experimental population would be required for such work. For Extraversion, no significant pattern emerges, however we propose that the extremes are similarly inhibited by over- or under-other-directedness.

Turning now the significance of our findings, and they have several important implications. At a theoretical level, they provide more data about personality behaviour in dialogue contexts, which extend previous research using monologue data. Additionally this can better inform our understanding of personality in relation to models of language production.

Our results also contribute to the dialogue and priming literature which, for example, acknowledge that individuals often behave differently, but that systematic variation has mainly been examined in sociological terms. Here we have presented data which shows real and important differences between individuals in conversational behaviour, and highlights the potential role of personality in priming experimentation, more generally.

Finally, our findings can be used to directly inform dynamic computer interface technology, which could allow linguistic alignment in a realistic way. For example, Nass, Moon, Fogg, and Reeves (1995) have shown that computer users viewed their ma-

chine more favourably when it mirrored their personality. On the basis of work reported here, we are closer to being able to represent personality at the conversational, interactive level. We therefore anticipate that this will lead to more convincing artificial agents and intelligent dynamic computer interfaces.

These findings also nicely complement those presented by Branigan, Pickering, Pearson, McLean, and Nass (2003), in which computer users syntactic align with a pre-programmed computer interface, whether they believed this to be another person or an 'unintelligent computer'. Therefore, if such an 'unintelligent computer' was to project personality, we may expect it to vary its degree of priming—in addition to its lexicon—depending upon the sort of personality *it* may wish to project.

Conclusion

We have used experimental priming data to investigate the influence of personality on interpersonal language behaviour. Proposing hypotheses which suggested both Extraversion and Neuroticism influence linguistic coordination, here we found that the less interpersonal trait—Neuroticism—surprisingly influenced priming, whilst Extraversion did not. Given our finding that priming is facilitated by moderate Neuroticism, but inhibited by more extreme levels, we explain this in terms of withdrawal by building upon a previously proposed model of personality and language production. Issues regarding the significance and potential implications of this study are also discussed.

Acknowledgements

The first and second authors gratefully acknowledge support from the UK Economic and Social Research Council and the School of Informatics. We also express our gratitude to Holly Branigan, Sarah Haywood, Alan Marshall, Janet McLean, Martin Pickering and Matt Watson for help and advice with the study.

References

Bell, A. (1984). Language as audience design. *Language in Society*, **13**, 145–204.

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, **18**, 355–387.

Bock, J. K., Loebell, H., and Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, **99**, 150–171.

Bradac, J. and Wisegarver, R. (1984). Ascribed status lexical diversity, and accent: Determinants of perceived status, solidarity, and control of speech style. *Journal of Language and Social Psychology*, **3**, 239–255.

Branigan, H., Pickering, M., and Cleland, A. (2000). Syntactic coordination in dialogue. *Cognition*, **75**, B13–B25.

Branigan, H., Pickering, M., Pearson, J., McLean, J., and Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 186–191.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Memory and Cognition*, **22**, 1482–1493.

Campbell, A. and Rushton, J. (1978). Bodily communication and personality. *British Journal of Social and Clinical Psychology*, **17**, 31–36.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to Intelligence and Extraversion. *British Journal of Social and Clinical Psychology*, **4**, 1–7.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counselling Psychology*, **16**, 1–19.

Corley, M. and Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin and Review*, **9**, 126–131.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Coupland, N. (1980). Style-shifting in a Cardiff work-setting. *Language in Society*, **9**, 1–12.

Davidson, R. J. (2001). Toward a biology of personality and emotion. *Annals of the NY Academy of Sciences*, **935**, 191–207.

Dewaele, J.-M. (2002). Psychological and sociodemographic correlates of communication anxiety in L2 and L3 production. *International Journal of Bilingualism*, **6**, 23–28.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**, 509–544.

Dewaele, J.-M. and Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, **28**, 355–365.

Eysenck, H. and Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Garrod, S. and Doherty, G. (1987). Saying what you mean in dialogue. *Cognition*, **27**, 181–218.

Gifford, R. and Hine, D. W. (1994). The role of verbal behaviour in the encoding and decoding of

interpersonal dispositions. *Journal of Research in Personality*, **28**, 115–132.

Gill, A. (2003). *Personality and Language: The projection and perception of personality in computer-mediated communication*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. and Oberlander, J. (2003a). Looking forward to more extraversion with n-grams. In L. Lagerwerf, W. Spooren, and L. Degand, editors, *Determination of Information and Tenor in Texts: Multiple Approaches to Discourse 2003*, pages 125–137. Stichting Neederlandistiek & Nodus Publikationen, Amsterdam & Münster.

Gill, A. and Oberlander, J. (2003b). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; Neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461.

Haywood, S., Pickering, M., and Branigan, H. (2003). Co-operation and co-ordination in the production of noun phrases. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 533–538.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennysylvania Press, Philadelphia.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

McCroskey, J. and Richmond, V. (1990). Wilingness to communicate: A cognitive view. *Journal of Social Behaviour and Personality*, **5**, 19–37.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, **43**, 223–239.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Pickering, M. and Branigan, H. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, **39**, 633–651.

Pickering, M. and Garrod, S. (in press). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*.

Schenkein, J. (1980). A taxonomy for repeating action sequences in natural conversation. In B. Butterworth, editor, *Language production*, volume 1, pages 21–47. Academic Press, London.

Scherer, K. R. (1978). Inference rules in personality attribution from voice quality: The loud voice of extraversion. *European Journal of Social Psychology*, **8**, 467–487.

Siegman, A. W. (1987). The tell-tale voice: Nonverbal messages of verbal communication. In A. Siegman and S. Feldstein, editors, *Nonverbal behaviour and communication*, pages 642–654. Erlbaum, Hillsdale, NJ.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, **53**, 718–726.

Yellen, R., Winniford, M., and Sanford, C. (1995). Extraversion and introversion in electronically-supported meetings. *Information & Management*, **28**, 63–74.

Multisensory enhancement of localization with synergetic visual-auditory cues

Martine Godfroy (mgodfroy@imassa.fr**),**
Cognitive Science Department, Institut de Médecine Aérospatiale du Service de Santé des Armées,
B.P.73, 91223 Brétigny sur Orge Cedex, France

Corinne Roumes (croumes@imassa.fr)
Cognitive Science Department, Institut de Médecine Aérospatiale du Service de Santé des Armées,
B.P.73, 91223 Brétigny sur Orge Cedex, France

Abstract

Enhanced behavioral performance mediated by multisensory stimuli has been shown using a variety of measures, including response times, orientation behavior and even simple stimulus detection. In the particular case of the study of saccadic response to unimodal or bimodal stimuli, Corneil et al. (2002) were able to show that the bimodal visual-auditory saccades benefited from the accuracy of visual saccades at saccadic response time (SRTs) typical of auditory saccades. However, there has been little evidence of multisensory mediated improvement in stimulus localization. Recently, Hairston et al. (2003) shows improvement in visual-auditory localization performance (variability) for induced myopia while no benefit was reported for normal vision. Using a similar experimental design, taking into account two space dimensions, azimuth vs. elevation, we examined the ability of human subjects to localize visual, auditory and combined visual-auditory targets for stimuli considered optimal for the given task. The results showed significant improvement in bimodal localization when compared with the more accurate modality, visual, as measured with multi criterion data (precision, dispersion and orientation of the response patterns). Furthermore, the 2D analysis of combined visual-auditory target localization performance, for azimuth and elevation response components, underlines the role of the auditory system in the determination of the response characteristics. The data suggested that visual-auditory localization performance benefited from the "*best of the two worlds*" (Corneil et al., 2002), in that it was improved only in the horizontal plane, and restricted to the response criterion where audition is more reliable than vision.

Introduction

The literature dealing with intersensory perception first dealt with the phenomena of sensory illusions, the most well known being the ventriloquism effect (Howard and Templeton, 1966) and the McGurk effect (McGurk and McDonald, 1976). Both these "on-line" effects result from discrepancies, either spatial and/or temporal between the two unimodal components of the stimulation. The much more ecological situation, in which visual and auditory signals are synergetic, i.e. in terms of spatial and temporal congruence, has been rarely investigated systematically in a localization task. Furthermore, to our knowledge, taking into consideration the two dimensions (azimuth and elevation) of the observer's perceptive field for a multimodal localization task was never explored. In addition, the simultaneous presentation of spatially congruent visual and auditory cues was mostly studied considering detection of a target (Frasinetti et al., 2002), orientation toward a target (Stein et al., 1988, 1989) or reduction in response latencies (Hugues et al., 1994; Frens et al., 1995, Colonius & Arndt, 2001) rather than purely localization capability. When shown, increase in precision of the localization was restricted to the analysis of an angular value, expressing the stimulus-response discrepancy in polar coordinates. The purpose of this experiment was to evaluate multisensory integration in a two-dimensional localization task and qualify the nature of a cross modal benefit that could be obtained when the spatial information in the two modalities was convergent. We suggested a separate analysis of the localization performance for the azimuth and elevation components of the response, as a function of target double pole coordinate system in which the origin coincides with the center of the head. This procedure should reveal the contribution of the auditory modality into the bimodal localization performance, given the initial differences in coding the position of an auditory target in azimuth (Interaural Time and Level differences) and in elevation (monaural spectral shape cues). Indeed, as a consequence of this specific coding, auditory resolution differs in the horizontal and the vertical dimension while the visual resolution, associated to a retinotopic coding, is isotropic in space. The investigation of criterion we assumed to be relevant for the task was performed. Centering, precision, dispersion and orientation of the responses were successively examined to determine a potential benefit and the modal contribution of a bimodal visual-auditory target presentation.

Materials and methods

Participants

Ten adults, aged 22 to 50 years, took part in the experiment. They all had a minimum of 20/20 visual acuity (if need be, corrected). Their audiometric capacities were also normal, with age related variations. All were naïve regarding the setup configuration (number and positions of the auditory sources).

Experimental setup

The participant sat in darkness in the center of an acoustically transparent semi-cylindrical vertical screen, 120 cm in radius and 145 cm high, with the head maintained by a chin-rest, as shown in Figure 1. A Liquid Crystal Display Philips Hopper SV10 video-projector was hung above and behind the observer, 245 cm from the screen, providing a 80° horizontal x 60° vertical green light field of view of 1.5 cd.m^{-2} average luminance (Fig. 1). The color green (coordinates of the 1931 CIE system x = 0.267; y = 0.640) was used for the background and for the visual stimulus, and made it possible to obtain a maximum signal to noise contrast and maximum background homogeneity, given the characteristics of the optic device. A PC (Pentium III 300 MHz) equipped with a 128 SoundBlaster sound card and a Matrox G400 (32MB) video card generated the stimuli. It was connected to the video-projector on the one hand, and to the loudspeakers via an audio switch and its Velleman K8000 control module, on the other hand. Thirty five 10-cm-diameter loudspeakers (Fostex FE103 Sigma) were laid out behind the screen in a 7 x 5 matrix, with a 10° step. The speaker positions were defined in a two-dimensional polar coordinate system with the origin at the straight-ahead fixation position. Eccentricity in the perceptive field was referred in relation to this coordinate system. The speakers were positioned at azimuths 0°, ±10°, ±20°, ±30° and elevations 0°, ±10°, ±20° (Figure 2).

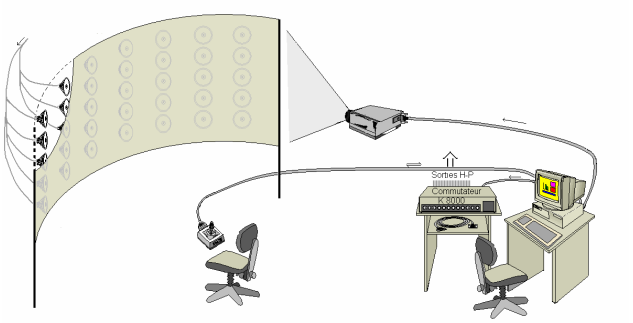


Figure 1: The experimental setup.

Visual stimuli consisted of a spot of light (100ms, 20 cd.m^{-2}), subtending a 1° of visual angle and auditory stimuli consisted of a pink noise burst (broadband noise, constant intensity per octave), 100ms duration (20ms fade-in and fade-out), at 49dB as measured at the subject's ear or hearing position, against a 38dB background noise (precision integrating sound level meter Brüel and Kjaer Model 2230). The device allows the precise superimposition of the visual and auditory stimulation for a combined presentation to the target, where the spot of light is exactly located at the center of the loudspeaker's cone. To perform localization judgments, participants used a track-ball, allowing for movements along all directions. Figure 2 describes the succession of the events in trial.

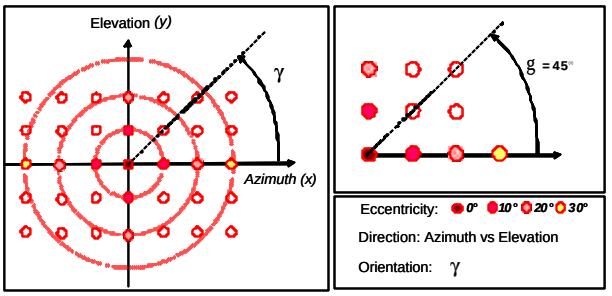


Figure 2: Definition of the independent variables used in the analyses and characterizing the target position. *Eccentricity* refers to the distance of the target from the center of the 2D perceptive field, *Direction* allow transforming target and response *Orientation* (?) in a two components position (azimuth and elevation).

1. At the beginning of each trial, a fixation cross was presented at the center of the screen, at (0°, 0°) coordinates, for 500 to 1500 ms for acquisition.
2. At the extinction of the cross, the visual, auditory or bimodal visual- auditory stimulus was presented randomly at one of the 35 positions during 100ms. The picture illustrates a - 20° to 0° visual stimuli.

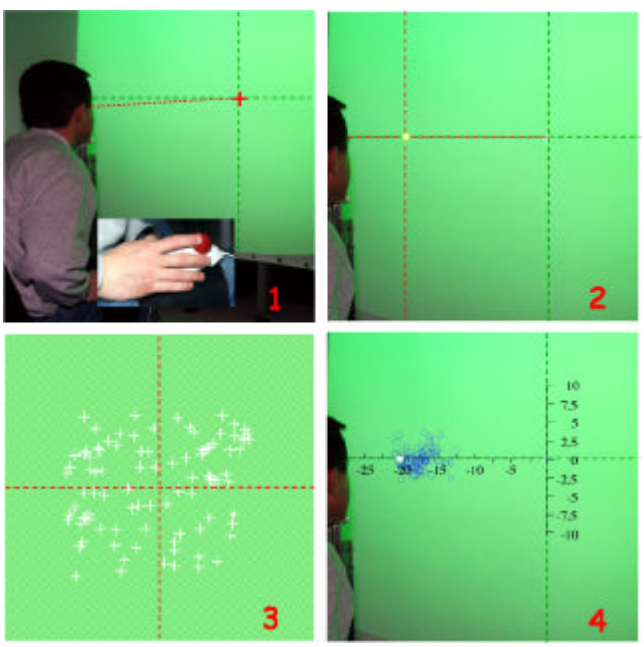


Figure 3: The experimental paradigm. 1. Presentation of a fixation cross at the center of the screen. 2. A visual stimulus at (-20°, 0° coordinates). 3. All the possible cursor position for the -20°, 0° target position. 4. Each dot stem from an individual localization response.

3. After the target disappears, a response cursor, associated to the further manipulation of the track-ball, appears randomly inside a 20° imaginary circle whose center is the position of the target with a minimum of 2.5° distance from it in both axes (azimuth and elevation).

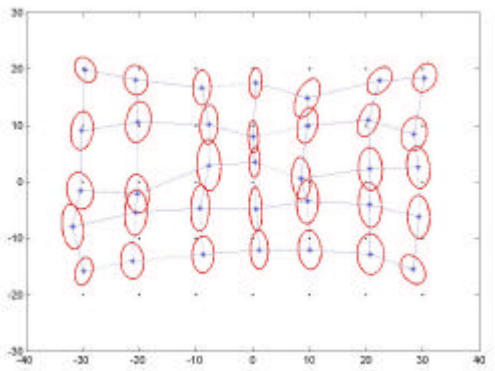
Subjects were instructed to localize the target as accurately as possible while pointing this cursor towards the perceived location of the target, the temporal constraint being secondary. The picture 4 illustrates the response distribution of the 10 subjects and 10 repetitions for the given location of the visual stimulation. The experiment consisted of 6 experimental sessions with 10 repetitions of each stimulus combination (3 stimulus conditions [Visual, auditory, bimodal] at 35 locations [7 azimuth values, 5 elevation values] presented in pseudo-random order) for a total of 175 trials per session, with a 1.5s inter-trial interval.
Prior to testing, 20 practice trials were performed to make the participant familiar with the task and the manipulation of the track-ball. The session lasted about 30 min. and a minimum 24-hour delay was observed between two sessions.

Data analysis

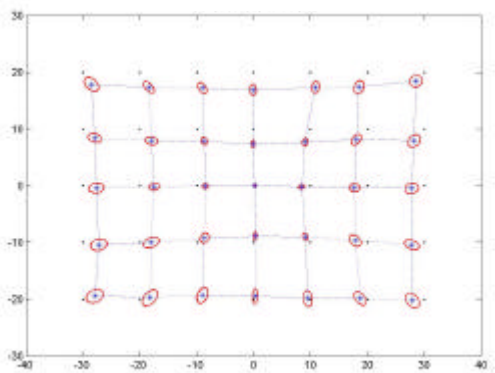
Localization errors were calculated as the difference, in degrees, between the localization judgment and the actual target location. Taking into consideration the azimuth and elevation components of the response, *centering* and *precision* of the responses were calculated from the raw data. *Centering* refers to the mean response, which the sign denotes a tendency to overshoot (positive values associated to errors eccentric to the target in reference to the reference coordinates) or undershoot (negative values associated to errors central to the target). *Precision* evaluate the amount of discrepancy (absolute value) from target to designation. The distribution of the response patterns were computed using a procedure of regression analysis for obtaining the regression slope that determines the major *orientation* of the response distribution. Estimation of the maximum and minimum variance of the distribution along the slope axis and the perpendicular one, respectively noted *b* and *a*, were used for *dispersion* analysis. By extension, in reference with Hofman et Van Opstal (Hofman et Van Opstal, 1998), a characterization of the response patterns under a geometrical approximation, i.e. ellipses, did allow a better comparison within and between modalities than the traditional methods using a two-dimensional discontinuous space analysis (Oldfield et Parker, 1984). In this way, the analysis of dispersion and orientation of the patterns would provide complementary data to those obtained with the use of the horizontal and vertical axis of the 2D coordinate system. To analyze the data, multiple 2-way within subjects ANOVAs were performed according to the specific hypothesis: Statistical comparisons were structured to examine the main effect of target modality (visual, auditory, combined visual-auditory) and target location (eccentricity range [0°, 10°, 20° and 30°] and direction [azimuth versus elevation]) as well as the possible interaction between the variables.

Results

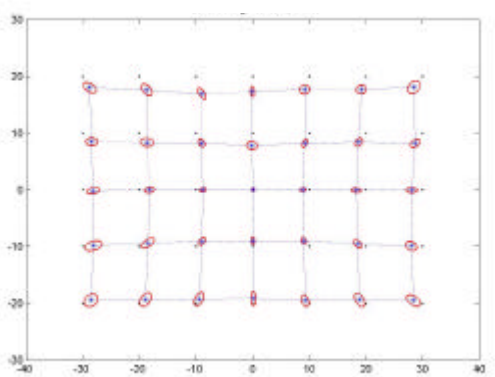
The results only consider here the comparison of response localization between modalities while a preliminary work was performed on unimodal data to ensure the validity of the results. We shall now successively describe the data using the four variables mentioned in section Data Analysis.



Auditory condition



Visual condition



Combined visual-auditory condition

Figure 4: Responses patterns as approximated by ellipses for the 3 conditions and the 35 target positions.

Precision of the responses

A short look at the approximated data for each condition of presentation of the target for the 35 positions tested (Figure 4) underlines the specificity of the auditory system in terms of localization capability and the relative similar localization behavior between the visual and the bimodal conditions.

Table 1: Precision of localization between conditions

Modality	Azimuth		Elevation		A/E	$F_{1,3486}$	*P*
	Error	SD	Error	SD			
Auditory	**2,96°**	**±2,6°**	**6,15°**	**±4,94°**	**A<E**	**1159**	**<0,0001**
Visual	**1,87°**	**±1,7°**	**1,86°**	**±1,63°**	**A=E**	**0,119**	**0,7304**
Bimodal	**1,53°**	**±1,5°**	**1,62°**	**±1,49°**	**A<E**	**7,7**	**0,0055**

A more detailed analysis of mean of errors confirmed this first impression. A repeated measures ANOVA showed that the effect of modality condition is significant in azimuth ($F_{2,336}$=87.2; p<.0001) and in elevation ($F_{2,336}$=23.316 p<.0001). The much more interesting result concerned the significant improvement in bimodal localization compared to the visual one in azimuth, (Scheffe test, p=0.0302) but interestingly, not in elevation (Scheffe test, p=0.8355). When looking at the within-modality variations between error in azimuth and error in elevation, expressed by the Azimuth/Elevation precision relationship (A/E in table 1), it appears that the gain obtained in the bimodal condition follows the difference in precision of the auditory condition (with statistically significant values). This result is an argument for audition playing a structuring role in intersensory processing for a spatial task.

Centering of the responses

One of the most well known characteristics of the auditory system is concerned with the differences in accuracy between azimuth and elevation, in relation to the differences in the initial information extraction process in the two directions of space (Oldfield & Parker, 1986; Hofman & Van Opstal, 1998). As a consequence, there is a strong response bias in the elevation responses, with a central compression of the auditory space related to a systematic undershoot of target eccentricity in this direction. No observable or statistical improvement in centering was obtained between the visual and the combined audio-visual conditions. On the other hand, the localization of an auditory target in azimuth is much less biased by eccentricity than for the visual and bimodal conditions, as illustrated in Figure 5. In this direction, the reduction of error is at the maximum when the direction of the visual and the auditory biases are in opposition of signs. When the sign of the bias is identical, no visible effect is observed. A statistical comparison between the visual and bimodal results fails to show any improvement, probably due to the arithmetic mean performed on data expressed in polar coordinates. Despite the lack of significance, the results did again suggest that the contribution of the auditory modality did enhance performance.

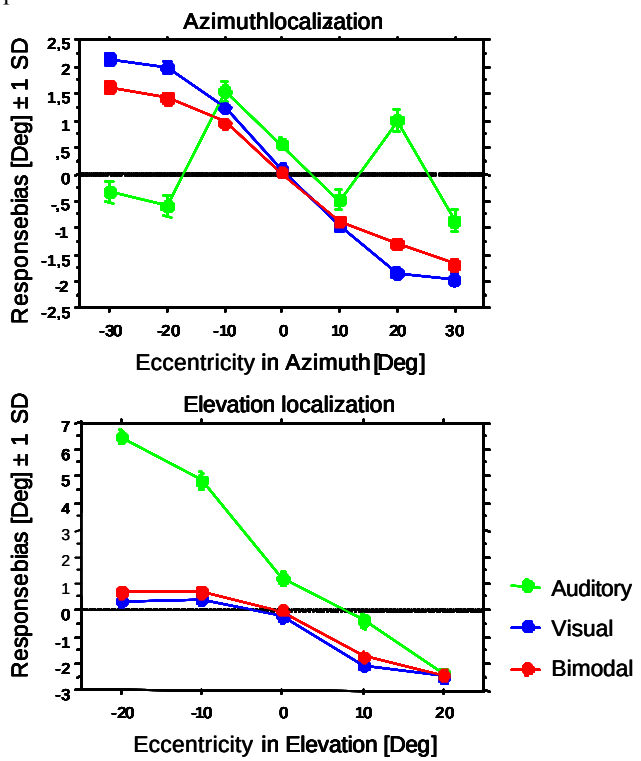


Figure 5: Centering of the responses for azimuth and elevation components of the localization responses. Improvement in performance is only visible in the horizontal plane.

Dispersion of the responses

The diverse responses are compared on the two characteristic axis of the responses patterns, *a* and *b* (Cf. Data analysis).

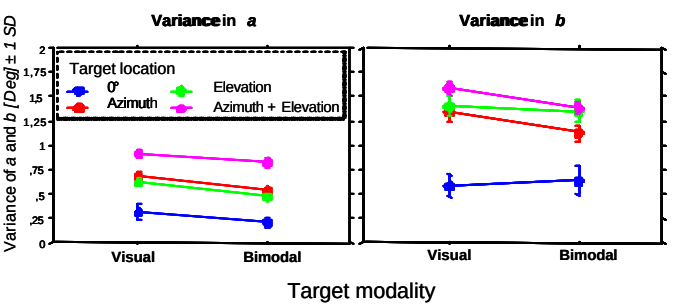


Figure 6: left: Decrease in variance in a between the visual and bimodal condition for all target locations. Right: Decrease in variance in b between the visual and bimodal condition for the targets that didn't belong to the median sagittal plane (0° and Elevation).

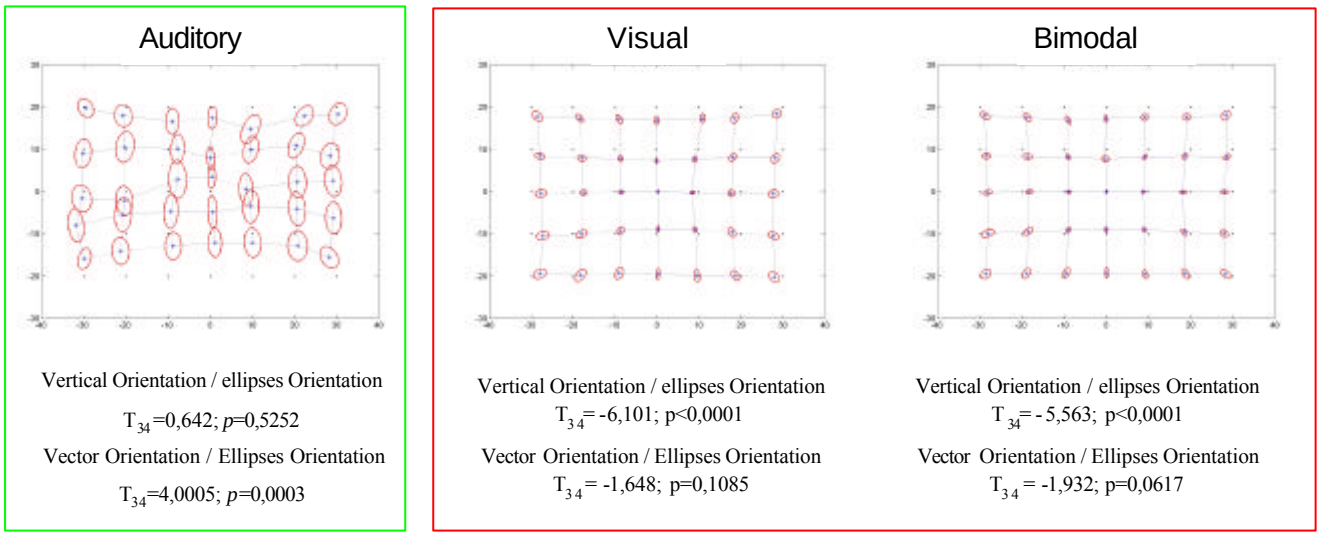


Figure 8: Orientation of the responses in relation with target location in the 2D perceptive field. In the auditory condition, the response patterns are vertically oriented (90°-180° axis) while visual and bimodal response patterns exhibit a vector distribution with the ellipses oriented centrifugally (toward the center of the perceptive field).

The minimum variance axis, *a*, diverges significantly according to the modality condition for target presentation (repeated measures ANOVA: ($F_{2,338}$=43.055 p<.0001), with the comparison of visual and bimodal conditions being also significant ($F_{1,169}$=23.356 p<.0001). Similar results are obtained for the *b* axis, with a slightly different behavior in relation to target location, expressed by the belonging or not to a specific plane (0°, Azimuth, Elevation, or combined eccentricity in Azimuth and Elevation). Indeed, only the targets that are not located on the median sagittal plane (0°and Elevation only) did benefit of a significant variance reduction (Figure 6).

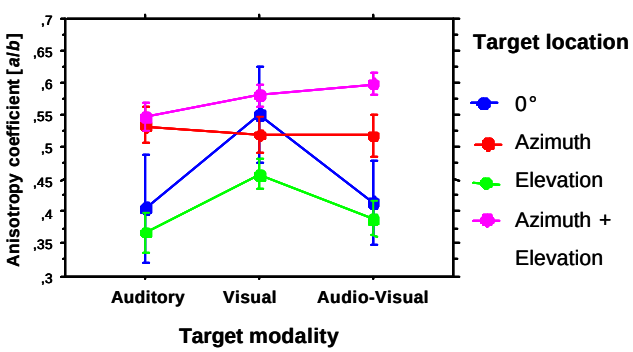


Figure 7: Anisotropy coefficient variations according to target modality and location in the 2D perceptive field. Note that the coefficient varies in the same way for auditory and bimodal conditions.

We also calculated an anisotropy coefficient, corresponding to the *a*/*b* ratio (a "1" value corresponding to an homogeneous distribution along the two axis), and looked at the variations of this coefficient according to the target location in space. It can be seen in Figure 7 that the value of the coefficient follows the same variations in the auditory and the combined visual-auditory condition. Once again, these data pointed out the role of the auditory modality into the multimodal spatial perception, not only in performance improvement, but also in representation structuring.

Orientation of the responses

At this point, we shall remember that the orientation of the responses distributions are determined by the slope of the regression analysis computed for the 35 tested target positions and the 3 modalities. In each condition, the calculated orientation is compared to two models of sensory coding: an auditory coding using a Cartesian coordinates system on one hand, and a vector coding, which can reflect a saccadic component in the response, on the other hand. The data shown in Figure 8allow mentioning that auditory response patterns are vertically oriented while visual and bimodal response patterns exhibit a vector distribution with the ellipses oriented centrifugally. These observations could reflect a possible different role of the saccadic system according to the target modality and the presence vs. absence of visual information in the perceptive field.

Discussion

This study investigated the localization performance to visual, auditory, and bimodal stimuli distributed throughout the 2D perceptive field. The result of the current study illustrates a significant multisensory

enhancement of localization performance in precision and dispersion. Through a quantitative approach, the data allowed to parameterize the different dimensions which describe the perceptive field of an ideal observer and to attest to the relative contribution of each sensory modality into the bimodal perception. The results argue for an integrative process applying for synergetic presentations of visual and auditory stimuli, and cues considered as well suited for the given task. For all that, our result did not refute the very ecological principle of the "inverse effectiveness rule" (Stein & Meredith, 1993). They just underline the structuring role of the auditory system only when it is more reliable than the visual system, what can be shown only by the comparison in performance for the two directional components (azimuth and elevation) of the response. It is a strong argument to say that sensory integration in a localization (spatial) task rests on a tendency to optimization. Looking at the data obtained by Corneil et al. (2002), showing that bimodal visual-auditory saccades were at least as accurate as visual saccades, but also generated at saccadic response times (SRTs) shorter typical of auditory saccades, our result also go in the way of a very similar neural process applying. This tendency to optimize shall be considered as an economic and ecological process that drove the Central Nervous System (CNS) to use the sensory systems in relation with the specific contribution they can have. In the case of a localization task (spatial task), and given the reliability of each sensory system, we demonstrated an improvement in centering and a part correction of the variance attributed to audition, an increase in precision and possibly in structure of representation for vision.

Acknowledgments

We would like to acknowledge the assistance of A. Bichot, L. Pellieux, J. Plantier and P.M.B. Sandor in the technical phases of the work. Special thanks to G. Cooper for his daily encouragement. Financial support was provided by IMASSA.

References

Colonius, H. & Arndt, P. (2001). "A two-stage model for visual-auditory interaction in saccadic latencies." *Perception & Psychophysics, 63*: 126-147.

Corneil, B.D., van Wanrooij, M., Munoz, D.P., van Opstal, A.J. (2002). "Auditory-visual interactions subserving goal-directed saccades in a complex scene." *Journal of Neurophysiology, 88,* 438-454.

Frasinetti, F. et al. (2002). "Enhancement of visual perception by crossmodal visuo-auditory interactions." *Experimental Brain Research, 147,* 332-343.

Frens, M.A.; van Opstal, A.J. & van der Willigen, R.F. (1995). "Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements." *Perception & Psychophysics, 57(6),* 802-816.

Hairston, D.W. et al. (2003). "Multisensory enhancement of localization under conditions of induced myopia." *Experimental Brain Research , 152,* 404-408.

Hofman, P.M. & Van Opstal, A.J. (1998). "Spectro-temporal factors in two-dimensional human sound localization." *Journal of Acoustical Society of America 103,* 2634-2648.

Howard, I.P. & Templeton, W.B. (1966). *Human spatial orientation*. New York, NY: Wiley.

Hugues, H.C.; Reuter-Lorenz, P.A.; Nozawa, G. & Fendrich, R. (1994). "Visual-auditory interactions in sensorimotor processing: saccades versus manual responses." *Journal of Experimental Psychology: Human Perception and Performance 20(1),* 131-153.

McGurk, H. & McDonald, J. (1976). "Hearing lips and seeing voices." *Nature, 264,* 746-748.

Oldfield, S.R. & Parker, S.P. (1984). "Acuity of sound localisation: a topography of auditory space. I. Normal hearing conditions." *Perception, 13(5),* 581-600.

Stein, B.E.; Huneycutt, W.S. & Meredith, M.A. (1988). "Neurons and behaviour: the same rules of multisensory integration apply." *Brain Research, 448:* 355-358.

Stein, B.E.; Meredith, M.A.; Huneycutt, W.S. & McDade, L. (1989). "Behavioural indices of multisensory integration: orientation to visual cues is affected by auditory stimuli." *Journal of Cognitive Neuroscience 1(1),* 12-24.

Stein, B.E. & Meredith, M.A. (1993). *Merging of the senses*, The MIT Press, Cambridge, MA.

Expressions Related to Knowledge and Belief in Children's Speech

Andrew S. Gordon (gordon@ict.usc.edu) and Anish Nair (anair@usc.edu)
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA

Abstract

Children develop certain abilities related to Theory of Mind reasoning, particularly concerning the False-belief Task, between the ages of 3 and 5. This paper investigates whether there is a corresponding change in the frequency of linguistic expressions related to knowledge and belief produced by children around these ages. Automated corpus analysis techniques are used to tag each expression related to knowledge and belief in a large corpus of transcripts of speech from normally developing English-learning children. Results indicate that the frequency of expressions related to knowledge and belief increases steadily from the beginning of children's language production. Tracking of individual concepts related to knowledge and belief indicates that there are no clear qualitative changes in the set of concepts that are expressed by children of different ages. The implications for the relationship between language and the development of Theory of Mind reasoning abilities in children are discussed.

A Developing Theory of Mind

Among the most interesting of human cognitive abilities are those concerning how we understand and reason about the minds of others. The term *Theory of Mind* is used pervasively throughout the cognitive sciences to refer to the set of abilities that enable people to reflect introspectively on their own reasoning, to empathize with other people by imagining what it would be like to be in their position, and to generate reasonable expectations and inferences about mental states and processes.

Within the research area of developmental psychology, Theory of Mind has been studied as a set of cognitive abilities that progressively emerge in children. A standard experimental instrument for studying children's Theory of Mind abilities is the false-belief task. In a standard version of this task (Wimmer & Perner, 1983), the child is introduced to two characters, Maxi and his mother. Maxi places an object of interest into a cupboard, and then leaves the scene. While he is away, his mother removes the object from the cupboard and places it in a drawer. The child is then asked to predict where Maxi will look for the object when he returns to the scene. Success on this task has been criticized as neither entirely dependent on Theory of Mind abilities nor broadly representative of them (Bloom & German, 2000), however its utility has been in reliably demonstrating a developmental shift. Wellman et al. (2001) analyzed 178 separate studies that employed a version of this task, finding that 3-year-olds will consistently fail this task on the majority of trials by indicating that Maxi will look for the object in the location to which his mother has moved it. 4-year-olds will succeed on half the trials, while 5-year-olds will succeed on the majority of trials. Call & Tomasello (1999) demonstrate that these results are consistent across verbal and non-verbal versions of this task.

Children's developing performance on the false-belief task is particularly interesting when couched within the larger debate concerning *maturation* and*. conceptual change* in cognitive development. Like every other cognitive ability that emerges in childhood, performance on Theory of Mind tasks is likely due to a complex combination of maturing innate abilities and knowledge learned through experience. Still, understanding the relative importance of these two factors may have some utility in evaluating two types of cognitive process models that have been proposed to account for human Theory of Mind abilities.

First, *Theory Theory* hypothesizes that Theory of Mind abilities are computed by prediction and explanation mechanisms by employing representation-level knowledge about mental attitudes (Gopnik & Meltzoff, 1997; Nichols & Stich, 2002). Second, *Simulation Theory* argues that Theory of Mind abilities are computed by imagining that you are in the place of the other person, then inferring their mental states by monitoring the processing that is done by your own cognitive abilities (Goldman, 2000). With respect to the development of Theory of Mind abilities in children, each of these theories would emphasize different things as most important. Theory Theorists would argue that the acquisition of mental models of commonsense psychology would play the most important role, a view consistent with a conceptual change model of development (e.g. Bartsch & Wellman, 1995). In contrast, Simulation Theorists would look instead for a maturational change that enabled children to take the perspective of other people or in the monitoring of one's own mental state, a view consistent with a modularity model of development (e.g. Scholl & Leslie, 2001).

One approach to investigating this issue is to look for evidence of the acquisition of mental models of commonsense psychology in the language that children use in everyday conversation. The contemporary view of natural language understanding and generation presupposes that the meaning of verbal expressions are representational in nature, and that these underlying representations are the same ones that would be manipulated for the purposes of inference (e.g. explanation and prediction). By tracking the production of children's speech that references commonsense psychology concepts, we can look for some correlation between linguistic competency with commonsense psychology concepts and emerging Theory of Mind abilities.

In this paper, we explore the progressive use of expressions that reference commonsense psychology concepts in children's language. The approach that we take in this investigation is to employ automated corpus analysis techniques developed within the computational linguistics research community, where the aim is to construct computer programs to reliably recognize every possible way of expressing a concept within a given language. In using automated corpus analysis techniques, we were able to quickly analyze each of the datasets within the CHILDES linguistic corpus (MacWhinney, 2000) that contained transcriptions of normally developing, monolingual English-learning children.

The specific interest that we had in conducting this research concerned the acquisition of a linguistic competency for concepts related to knowledge and beliefs, as they are the most relevant to the false-belief task described earlier. By examining the correspondences between these linguistic competencies and the ages in which children acquire cognitive competencies in Theory of Mind tasks, our aim is to provide an additional point of evidence that can be used in arguing for or against the competing models of the cognitive processes that underlie these abilities.

The Theory of Mind in Language

The Theory of Mind in Language project at the University of Southern California is an effort aimed at developing a large-scale lexical-semantic resource for the automated annotation of commonsense psychological concepts expressed in English text. This resource is being authored as a set of local grammars, encoded as finite-state transducers that can be applied to large text corpora for concept-level tagging and markup. Associated with each unique concept tag in the resource is a local grammar that has been hand-authored with the aim of recognizing every possible way that the concept could be expressed in the English language. The application of these local grammars to text documents produces an annotated text, where each English expression that is recognized as referencing a commonsense psychological concept is tagged. The following paragraph (from William Makepeace Thackeray's 1848 novel, *Vanity Fair*) provides an example of the output of the application of this lexical-semantic resource.

> Perhaps [*partially-justified-proposition*] she had mentioned the fact [*proposition*] already to Rebecca, but that young lady did not appear to [*partially-justified-proposition*] have remembered it [*memory-retrieval*]; indeed, vowed and protested that she expected [*add-expectation*] to see a number of Amelia's nephews and nieces. She was quite disappointed [*disappointment-emotion*] that Mr. Sedley was not married; she was sure [*justified-proposition*] Amelia had said he was, and she doted so on [*liking-emotion*] little children.

The tag set that is being used in this lexical-semantic resource was developed first through the large-scale analysis of strategies, defined as the abstract structural commonalities that exist between analogous planning cases (Gordon, 2002). 635 concepts resulting from this analysis (grouped into 30 representational areas) were related to a Theory of Mind, which constitutes the broadest cognitive science specification of a representational Theory of Mind to date. Gordon & Hobbs (2003) describe how this tag set is modified through the process of examining the breadth of English language expressions that are related to a given representational area, among the 30 in the complete set. Gordon et al. (2003) describes the process of constructing local grammars for each of the concepts in the revised tag set, encoded as finite-state transducers, and describes an evaluation to determine the effectiveness of this approach at automatically recognizing every English expression that refers to the concept tag in written text.

One of the 30 representational areas described in this previous work, Managing Knowledge, specifically deals

Managing knowledge (37 concepts)

He's got a logical mind (*managing-knowledge-ability*). She's very gullible (*bias-toward-belief*). He's skeptical by nature (*bias-toward-disbelief*). It is the truth (*true*). That is completely false (*false*). We need to know whether it is true or false (*truth-value*). His claim was bizarre (*proposition*). I believe what you are saying (*belief*). I didn't know about that (unknown). I used to think like you do (*revealed-incorrect-belief*). The assumption was widespread (*assumption*). There is no reason to think that (*unjustified-proposition*). There is some evidence you are right (*partially-justified-proposition*). The fact is well established (*justified-proposition*). As a rule, students are generally bright (*inference*). The conclusion could not be otherwise (*consequence*). What was the reason for your suspicion (*justification*)? That isn't a good reason (*poor-justification*). Your argument is circular (*circular-justification*). One of these things must be false (*contradiction*). His wisdom is vast (*knowledge*). He knew all about history (*knowledge-domain*). I know something about plumbing (*partial-knowledge-domain*). He's got a lot of real-world experience (*world-knowledge*). He understands the theory behind it (*world-model-knowledge*). That is just common sense (*shared-knowledge*). I'm willing to believe that (*add-belief*). I stopped believing it after a while (*remove-belief*). I assumed you were coming (*add-assumption*). You can't make that assumption here (*remove-assumption*). Let's see what follows from that (*check-inferences*). Disregard the consequences of the assumption (*ignore-inference*). I tried not to think about it (*suppress-inferences*). I concluded that one of them must be wrong (*realize-contradiction*). I realized he must have been there (*realize*). I can't think straight (*knowledge-management-failure*). It just confirms what I knew all along (*reaffirm-belief*).

Figure 1. Example expressions for 37 concepts related to Managing Knowledge.

with the concepts surrounding knowledge and belief, including assumptions, contradictions, justifications, logical consequences, truth, falsehood, and the mental processes associated with these commonsense psychological entities. Gordon et al. (2003) describe 37 concept tags associated with this area, which is presented here in Figure 1. The evaluation described in this previous work indicated that the lexical-semantic resources associated with this specific subset of the tag set was effective at identifying 83.92% of the expressions associated with these tags in English written text (recall score), and that 92.15% of the tagged expressions would be judged as appropriate by a human rater (precision score).

References in the CHILDES Corpus

As a corpus of analysis, we utilized the CHILDES database of children's speech (MacWhinney, 2000), a collection of transcripts from a wide variety of psycholinguistic studies conducted largely in the 1980s. Specifically, we analyzed the transcripts from the 42 research studies that contributed data of normally developing monolingual English-learning children.

To facilitate the analysis of this dataset according to the age of the children, individual files were generated containing only the transcripts of speech produced by a single child for each of the transcript files (a total of 3001 individual files). The total number of words in each file was calculated and the age of the child (in months) was recorded. There were 3,347,340 words transcribed in these files from children ranging in age from 11 to 87 months.

Figure 2 presents a histogram of the number of words in the files associated with each age of the children. The notable spike that appears in this figure is due to a large dataset that exists within the CHILDES database contributed from a study by Hall et al. (1984). The groups of children are collectively identified only as being between the ages of 54 and 60 months without differentiation, so all of this dataset was used for evidence at the low end of this range. More significantly, Figure 2 reveals that comparatively little data exists within the CHILDES corpus of normally developing monolingual English-learning children after the age of 5 years (60 months). Although the available data should allow for the observation of some interesting trends throughout the age range of the corpus, some caution is necessary when drawing strong conclusions about children older than 60 months.

In order to enable comparisons between children and adults, each of the analyses were also conducted on the CALLHOME American English Speech data collected and transcribed by the Linguistic Data Consortium (1997). The CALLHOME database consists of transcripts of 120 unscripted telephone conversations (302,083 words) between native speakers of English, where the callers average 38.875 years in age (σ=16.14).

Given these text corpora, two sets of analysis were conducted. Both of these analyses involved the use of tag frequency as data points. To compute tag frequency, the local grammars described in the previous section were applied to a corpus in order to find every expression within the corpus that should be tagged with the concept associated with each local grammar. The number of tagged expressions was then divided by the number of words that were searched to compute each frequency data point. In the first analysis, the frequency of all expressions of the 37 concepts related to knowledge and belief were tabulated for each of the datasets. In the second analysis, the frequencies of expressions related to each individual concept (of the 37 total) were tabulated.

No attempt was made to filter the results of the application of the local grammars to improve precision, and no evaluation was conducted to estimate the recall rate on these corpora. However, after reviewing the resulting tags we believe that the precision and recall scores obtained on these corpora are only marginally less than was achieved in the evaluation of written text tagging conducted by Gordon et al. (2003) using the same set of local grammars.

Frequency of all expressions related to knowledge

The first analysis that we conducted was to apply all of the local grammars for the 37 concepts related to knowledge and belief to each of the data files corresponding to children of different ages. In all, there were 18,283 tags produced

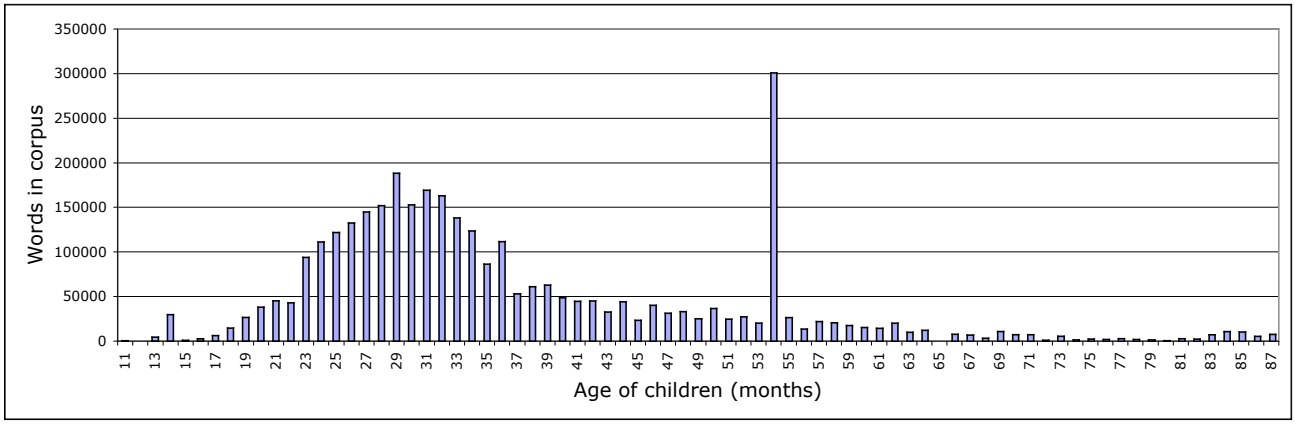


Figure 2. Number of words in corpus by age of children in months

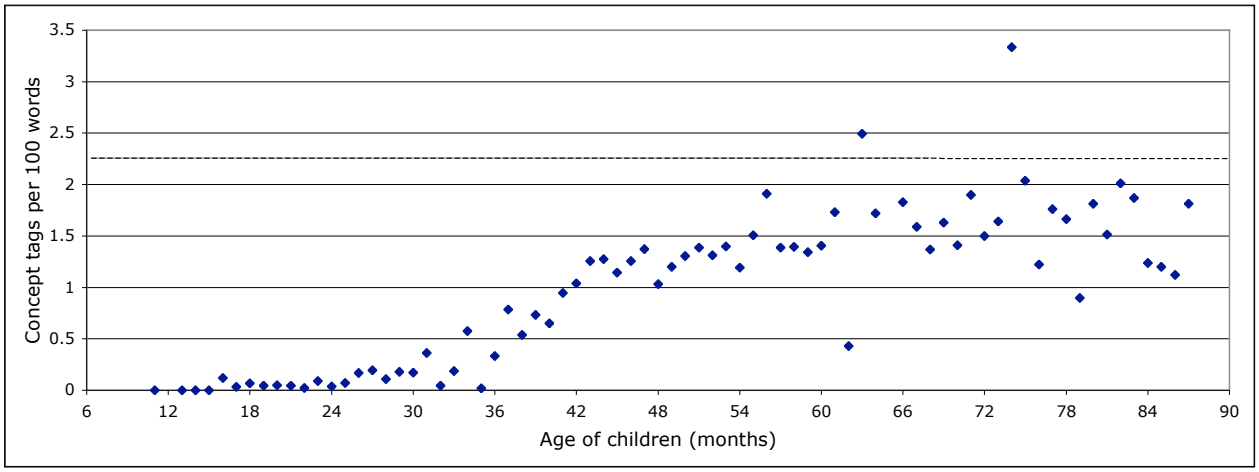


Figure 3. Frequency of all expressions related to knowledge and belief by age of children

through the application of these local grammars, with only 19 of the 37 tags appearing in the data. Nearly half of these tags were for the concept of a justified proposition (9113 tags), while the remaining half was dominated by tags to the concepts of belief (3150 tags), contradictions (3485 tags), and partially-justified propositions (1483 tags).

Applying the full set of local grammars to the CALLHOME data set produced 6775 tags, yielding a frequency of 2.24 reference per 100 words of speech. 21 of the 37 tags were assigned to this data, with the highest frequencies going to the concepts of justified proposition (3172 tags), contradiction (1551 tags), belief (1000 tags), and partially-justified proposition (493 tags).

Figure 3 presents a graph of the frequency per 100 words of speech for all expressions related to the concepts of knowledge and belief based on the age of the children (in months) of the analyzed data. As a point of comparison, the frequency for the CALLHOME data (2.24) is also indicated on the graph as a dashed horizontal line. The data on the graph can be described by the linear function $y=0.0281x - 0.3914$, where the correlation statistic (r^2) is 0.7021.

The results indicate that expressions related to knowledge and belief do not appear at the beginning of children's speech production, but increase in frequency in a strongly linear manner from 30 months (2.5 years) until 48 months (4 years), when the frequencies of these expressions are roughly half of what is observed in adult conversational speech.

Frequency of expressions of individual concepts

In the second analysis, we individually applied each of the 19 local grammars that produced at least one tag in the corpus to each of the transcript data for children of different ages. The primary purpose of this analysis was to track the relative increase in frequency for each concept over the developmental period where a change in Theory of Mind abilities is evident (between 36 and 60 months of age).

	Total tags	24 mo.	30 mo.	36 mo.	42 mo.	48 mo.	54 mo.	60 mo.	CallHome
add-assumption	499	0.27	0.65	1.34	3.31	0.91	4.22	2.59	5.16
assumption	22	0	0	0	0	0	0.17	0	0.07
belief	3150	0.72	3.21	3.67	17.67	18.71	12.39	24.63	33.1
bias-toward-disbelief	26	0	0	0	0	0	0.17	0.65	1.22
check-inferences	4	0	0	0	0	0	0	0.65	0.23
consequence	1	0	0	0	0	0	0	0	0.26
contradiction	3485	0	0	0	22.09	29.27	30.43	27.22	51.34
false	69	0	0	0	1.1	0.6	0.37	0.65	0.26
ignore-inference	1	0	0	0	0	0	0.03	0	0
justified-proposition	9113	2.07	13.02	28.35	51.46	45.87	61.42	73.23	105
knowledge	5	0.09	0	0	0	0	0	0	0.53
managing-knowledge	26	0	0	0	0	0	0.33	0	0.36
partial-domain-knowledge	23	0	0	0	0.22	0	0.2	0	0.6
partially-justified-proposition	1483	0.27	0	0	6.85	6.04	7.77	8.43	16.32
reaffirm-belief	19	0	0.26	0	0	0	0.13	0	0.36
realize	229	0.18	0.13	0.09	0.22	1.81	1.3	2.59	5.63
true	67	0	0.07	0	0	0	0.47	0	2.02
unjustified-proposition	51	0	0	0	1.1	0	0.07	0	0.93
world-model-knowledge	10	0	0	0	0	0	0	0	0.73

Figure 4. Frequency of expressions of individual concepts related to knowledge and belief (tags per 10,000 words)

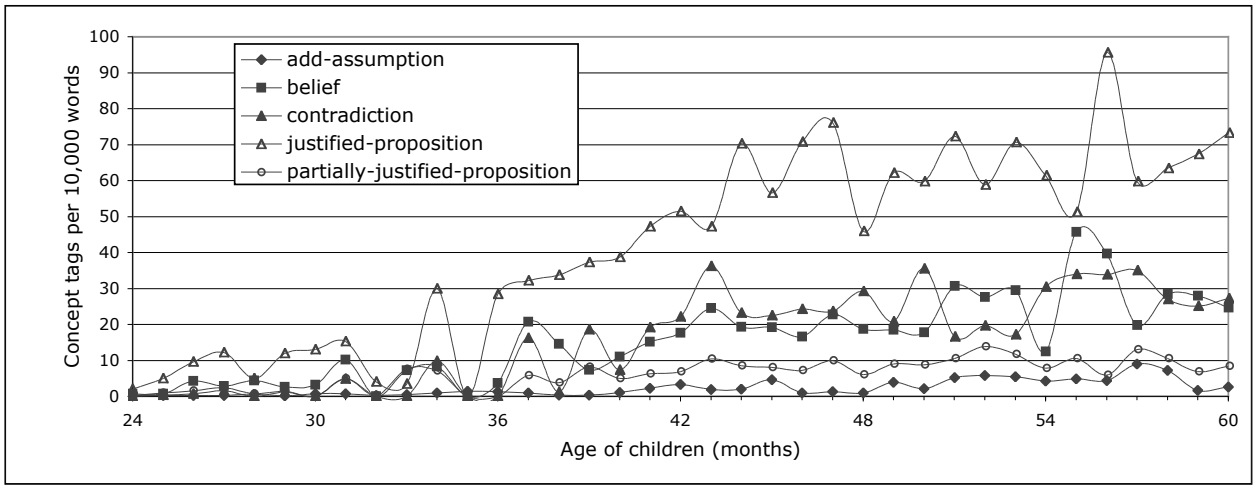


Figure 5. Frequency of expressions of the 5 most frequent individual concepts by age of children

Figure 4 presents a chart of the results of this analysis. Each concept is listed with the total number of tags assigned in the corpus and frequencies of occurrence within the data sets for 24, 30, 36, 42, 48, 54, and 60 month-old children, along with the CALLHOME frequency for the concept. Figure 5 further describes the results of this analysis by charting the growth in frequency of expressions related to the 5 most frequent concepts tagged in the corpus (add-assumption, belief, contradiction, justified-proposition, and partially-justified-proposition) between the ages of 24 and 60 months.

The results indicate that the increases in overall frequency of expressions related to knowledge and belief can be attributed to a steady increase in expressions related to a handful of concepts, particularly the concepts of belief, contradiction, and justified-propositions. This steady increase begins at 24 months and continuing past 48 months, when the frequencies of these expressions are roughly half of what is observed in adult conversational speech evidenced by the CALLHOME corpus. There is no evidence of any qualitative change in the sorts of concepts related to knowledge and belief that are expressed by children of different ages.

Discussion

The overall purpose of this study was to determine the relationship between a linguistic competency in the production of expressions related to knowledge and belief and children's developing Theory of Mind abilities, particularly during the age range where children acquire competency on the false-belief task (between 3 and 5 years in age). In this section, we will consider the results of our analysis with respect to this purpose.

First, there is no evidence to suggest a qualitative change in the *frequencies* that children express concepts related to knowledge and belief between the ages of 3 and 5. Looking first only at the frequency of all expressions related to knowledge and belief we see that children between the ages of 3 and 5 are continuing a steady increase in frequency that started at the beginning of their speech production. The sparse data that we have for children older than 60 months suggests that this gradual increase begins to level off after this point. If we had seen a non-linear shift in the frequencies of expression between 3 and 5, then an argument could have been made relating linguistic competency to Theory of Mind abilities. Finding no such shift, one could reasonably infer that the developing linguistic competencies that children have for expressions related to knowledge and belief are unrelated to their reasoning abilities in Theory of Mind tasks.

Second, there is no evidence to suggest a qualitative change in the *concepts* that children express related to knowledge and belief between the ages of 3 and 5. Looking at the individual frequencies for each of the 19 concept tags that were assigned to the corpus we see that the handful of concepts that account for the vast majority of tags increase in frequency at a constant rate from the very beginning of children's speech production. Very few expressions appear in this data related to other concepts that appear with slightly higher frequencies in adult discourse, and there is no evidence that linguistic competency is acquired for these concepts during this period of time either. If we had seen a change in the concepts that were being expressed between 3 and 5, then a different argument could have been made relating linguistic competency to Theory of Mind abilities. Finding no such change one could again reasonably infer a lack of a direct relationship between language-use and acquired reasoning abilities.

Together these two points argue against a strong relationship between linguistic competencies for expressions related to knowledge and belief and children's developing Theory of Mind abilities. This argument is particularly important in evaluating cognitive models that assume that Theory of Mind abilities and language abilities are enabled by representational mental models of the same type. If we assume that the sophistication of children's representational theories of knowledge and belief is closely related to the way that children express these concepts in language, then

there is little evidence to suggest that these representational theories change at all between ages of 3 and 5, when competency on the false-belief task develops. Accepting this assumption, the evidence in this paper would argue against any strong *conceptual change* account of Theory of Mind abilities where competency on the false-belief task is due solely to the acquisition of more sophisticated representational mental models. This evidence would argue instead for a *maturational* account, where competency on the false-belief task can be attributed to the development of new cognitive abilities for taking the perspective of other people or in the monitoring of one's own mental state between the ages of 3 and 5. One strong counterargument that could be made against this line of reasoning concerns the differences in the linguistic competencies between language production and language understanding. In analyzing transcript data consisting of words uttered by children, this study can make no claims regarding the linguistic competency that these children might have for *understanding* expressions related to knowledge and belief during the relevant periods of development.

Conclusions

The availability of large corpora of transcripts of children's speech production has afforded researchers the opportunity to investigate a wide variety of issues related to language acquisition. This paper has demonstrated that specific issues related to the acquisition Theory of Mind abilities can also be addressed using these corpora. By employing automated techniques for the tagging of expressions related to commonsense psychology we have been able to efficiently analyze data sets that are larger than could have been reasonably tackled given limited resources.

The specific interest of this paper was to determine if there was evidence for change in the linguistic competency in expressions related to knowledge and belief during developmental periods associated with acquired competency in the false-belief task (between 3 and 5 years of age). By using automated corpus analysis techniques, expressions related to knowledge and belief were identified across all datasets within the CHILDES corpus containing speech from normally developing monolingual English-learning children. By charting the frequencies of these expressions at different ages, it is evident that children steadily increase the frequency of expressions related to knowledge and belief at a constant rate from the beginning of their speech production. By tracking the production of expressions related to individual concepts, no qualitative changes in the conceptual content of these expressions over time is evident. These results argue against a strong relationship between linguistic competencies for expressions related to knowledge and belief and children's developing Theory of Mind abilities.

Acknowledgments

This paper was developed with funds of the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

References

Bartsch, K. & Wellman, H. (1995) *Children Talk About the Mind*. New York: Oxford University Press.

Bloom, P. & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. Cognition 77, B25-B31.

Call, J. & Tomasello, M. (1999) A Non-Verbal False-Belief Task: The Performance of Children and Great Apes. *Child Development* 70(2):381-395.

Goldman, A. (2000) Folk Psychology and Mental Concepts. *Protosociology* 14, 4-25.

Gopnick, A. & Meltzoff, A. (1997) *Words, Thoughts, and Theories*. Cambridge, MA: Bradford, MIT Press.

Gordon, A. & Hobbs, J. (2003) Coverage and Competency in Formal Theories: A Commonsense Theory of Memory. *Proceedings of the 2003 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University, March 24-26, 2003.

Gordon, A. (2002) The Theory of Mind in Strategy Representations. *Proceedings of the Twenty-fourth Annual Meeting of the Cognitive Science Society (CogSci-2002)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, A., Kazemzadeh, A., Nair, A., & Petrova, M. (2003) Recognizing Expressions of Commonsense Psychology in English Text. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*.

Hall, w. Nagy, W., & Linn. R. (1984) *Spoken words: Effects of situation and social group on oral word usage and frequency*. Mahwah, NJ: Erlbaum.

Linguistic Data Consortium (1997) CALLHOME American English Transcripts. LDC catalog no. LDC97T14. http://www.ldc.upenn.edu

MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk (Vol. I & II)* Mahwah, NJ: Lawrence Erlbaum Associates.

Nichols, S. & Stich, S. (2002) How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness. In Q. Smith and A. Jokic. (eds.) *Consciousness: New Philosophical Essays*. Oxford University Press.

Scholl, B. & Leslie, A. (2001) Minds, Modules, and Meta-Analysis. *Child Development* 72(3): 696-701.

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. Child Development 72, 655-684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13, 103-128.

Strategy Constancy Amidst Implementation Differences: Interaction-Intensive Versus Memory-Intensive Adaptations To Information Access In Decision-Making

Wayne D. Gray, Michael J. Schoelles, & Christopher W. Myers
Cognitive Science Department
Rensselaer Polytechnic Institute
[grayw, schoem, myersc] @rpi.edu

Abstract

Over the last two decades attempts to quantify decision-making have established that, under a wide range of conditions, people trade-off effectiveness for efficiency in the strategies they adopt. However, as interesting, significant, and influential as this research has been, its scope is limited by three factors; the coarseness of how effort was measured, the confounding of the costs of steps in the decision-making algorithm with the costs of steps in a given task environment, and the static nature of the decision tasks studied. In the current study, we embedded a decision-making task in a dynamic task environment and varied the cost required for the information access step. Across three conditions, small changes in the cost of interactive behavior led to changes in the strategy adopted for decision-making as well as to differences in how a step in the same strategy was implemented.

Introduction

In the 80's and 90's, Payne, Bettman, and Johnson (1993) showed that decision-makers trade-off efficiency of their decision making strategy for the effort it requires. They attempted to quantify the cognitive effort of decision making by counting the number of steps that different strategies required for the same decision. The conclusion of this work was that people adapt to a wide variety of conditions to find a strategy that is about as accurate as it needs to be for as little cognitive effort as possible.

As interesting, significant, and influential as Payne, et al.'s work was, its scope was limited by three factors; the coarseness of how effort was measured, the confounding of the costs of steps in the decision-making algorithm with the costs of steps in a given task environment, and the static nature of the decision tasks studied.

First, the *elementary information processes* (EIPs) that Payne et al. used to count steps were neither elementary or steps. By today's standards EIPs such as "reading value, comparing two values or storing a result in long-term memory" (Todd & Benbasat, 2000) would be analyzed as a series of more fundamental cognitive, perceptual, and action operations. Furthermore, the count of steps was not based on an analysis of the decision-making process executed by a human, but stemmed from task analyses of the minimum number of steps a perfect agent would require to execute the algorithm. The step count did not consider the mis-steps or re-steps taken by a boundedly rational agent as they skipped a step or forgot an intermediary product, and then backed up and redid a number of steps to recover.

Second is the confounding of the costs of a step in the decision-making algorithm with the costs associated with how a step is implemented in a given task environment. Research has shown that the organization, form, and sequence of information influences strategy selection (for example, Fennema & Kleinmuntz, 1995; Kleinmuntz & Schkade, 1993; Schkade & Kleinmuntz, 1994). Other research has looked at how individual differences in working memory capacity interact with interface design to affect performance on decision-making tasks (Lohse, 1997). Other studies have looked at how the design of decision aids may have unintended consequences for the decision strategies that people adopt (Adelman, Miller, & Yeo, 2001; Benbasat & Todd, 1996; Rose & Wolfe, 2000; Todd & Benbasat, 1994, 1999, 2000). At least one study has investigated how the cost of information access affects strategy selection (Lohse & Johnson, 1996).

The third limit on the scope of Payne, et al.'s pioneering work is that the decision-making tasks they used were static, not dynamic. Although time constraints were sometimes introduced (Payne, Bettman, & Luce, 1996), these were extrinsic, not intrinsic to the decision-making task. For example, subjects were told to work quickly, timed, or rewarded for fast performance. Such extrinsic time pressure differs from tasks where the information, options, and criteria for decision-making change over time (Adelman, Bresnick, Black, Marvin, & Sak, 1996) or in which an early step in decision-making may result in changes to the task environment (Ehret, Gray, & Kirschenbaum, 2000). Hastie (2001) has characterized these dynamic situations as entailing a series of "linked decisions in a dynamic, temporally extended future" and has marked understanding this type of decision making as one of his 16 challenges for decision-making research in the 21st century.

The current paper reports empirical data from the first of a planned series of experimental and modeling efforts to extend the scope of decision-making research. In the study reported here, decision-making was embedded as an integral part of a dynamic classification task. Subjects' goal was to score as high as possible on the classification task while maximizing performance on the decision-making task. This initial study focuses on the ways in which varying the cost of interactive behavior affects the decision-making process. Specifically, across three between-subject conditions, we introduced modest differences in the cost of information access and studied how these differences affected the mix of cognitive, perceptual, and action operations for acquiring and comparing information.

Method

Subjects

Forty undergraduate students participated for approximately five hours each. Seven failed to complete the study. Subjects were either given course credit or were paid $5.00 per hour of participation and a $5.00 per hour completion bonus. Subjects were run individually.

Task

The experimental task was a preferential choice decision-making task embedded in the Argus Prime simulated radar-operator task environment (Schoelles & Gray, 2001a). Argus Prime is a complex but tractable simulated task environment (Gray, 2002) that we have used in a variety of studies (see, e.g., Gray & Schoelles, 2003; Schoelles, 2002; Schoelles & Gray, 2001b).

Classification Task. For the classification task, the subject must assess the threat value of each target in each sector of a radar screen (depicted in Figure 1). The screen represents an airborne radar console with ownship at the bottom. Arcs divide the screen into four sectors; each sector is fifty miles wide. The task is dynamic since the targets have a speed and course. A session is scenario driven; that is, the initial time of appearance, range, bearing, course, speed, and altitude of each target are read from an experimenter-generated file. The scenario can contain events that change a target's speed, course, or altitude. New targets can appear at any time during the scenario.

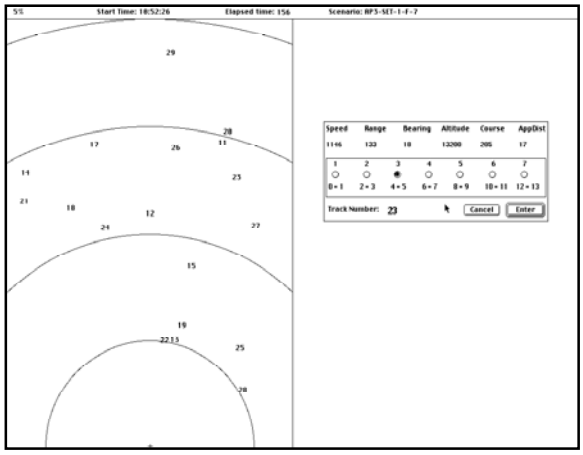


Figure 1: Argus Prime Radar Screen (left) and Information/Decision Window (upper right).

The subject selects (i.e., hooks) a target by moving the cursor to its icon (i.e., track number) and clicking. When a target has been hooked, an information window appears (on the upper-right of the display) that contains the track number of the target hooked and the current value of target attributes such as speed, bearing, altitude, and course. The subject's task is to combine these values, using an algorithm that we have taught them, and to map the result onto a 7–point threat value scale (at the bottom of the information window).

Targets must be classified once for each sector that they enter. If a target leaves a sector before the subject can classify it, it is considered incorrectly classified and a score of zero is assigned. A running score that indicates percentage of targets correctly classified is shown in the upper-left of the display. For this study, each Argus Prime scenario lasted 12-min. During this period a subject had the opportunity to calculate the threat value of between 70 and 90 targets.

The Decision–Making Task (DMT). The decision-making task (DMT) was added to Argus Prime for this study. As discussed in the Procedure section, subjects were introduced to the DMT after an hour of training and a second hour of practice on the classification task.

Each scenario proceeded until the subject had classified 8 targets. At this point, a DMT presented the subject with 4 or 6 targets for which he or she had already calculated the threat value. All groups were given the identification number for each of the DMT alternatives in a *target-column* that appeared in the lower right of the display (this area is blank in Figure 1). The subject's task was to determine which target had the highest threat value and select that target by clicking on its number in the target-column. The DMT ended and the classification task resumed when the subject clicked the CHOOSE button located below the target-column.

On making a correct choice, feedback was given via a simulated explosion, the chosen aircraft was removed from the radar screen, and the overall percent score for decision-making on that scenario was increased. If the participant chose the incorrect target, the participant's overall percent score for that scenario was reduced. A running average of DMT performance was presented to the right of the classification score. After classifying or re-classifying 8 more aircraft, another DMT was presented. This sequence continued until the end of each scenario.

Procedure

Subjects were randomly assigned to one of three DMT conditions: Table, 0-Second Lockout (0-Lock), or 2-Second Lockout (2-Lock). We were most interested in differences between the two lockout conditions, with the Table condition providing a measure of how high decision-making performance could be in this task environment under near optimal conditions.

As in other Argus Prime studies, subjects were trained for 1-hr on the Argus Prime classification task. They then practiced this task during their second hour by performing four scenarios in which the classification task was the only task. After the fourth scenario, subjects were given a short break and were then instructed on the DMT task. Training on the DMT took approximately 10-min. During the last 8 scenarios (5 through 12), subjects continued doing the classification task while being interrupted to perform the DMT.

The more time spent on the DMT, the more likely it would be that a target would cross a sector boundary

without being classified. Such unclassified targets were assigned a score of zero. Hence, time on the DMT decreased time available for classification. This, in turn, placed pressure on the subjects to perform the DMT quickly.

The three between-subject conditions differed in their cost of information access. As it was unclear to us how demanding the DMT would be in the Argus Prime task environment, the Table condition provided near minimum access costs. For this condition the numeric threat value for each target was listed in the target-column next to the target's identification number. Subjects simply scanned the target-column for the highest threat value (a 1–7 scale).

In contrast, to obtain a threat value, the 0-Lock and 2-Lock groups had to locate the target on the radar screen and move the cursor to it. Similar to a "tool-tip", the threat value then appeared next to the target. For 0-Lock, the threat value appeared as soon as the cursor moved to the target. For 2-Lock, the threat value appeared after a 2-s delay.

Results

Our focus is on process measures; namely, how the cost of information access affects the combination of cognitive, perceptual, and action operators required to implement the information access step in decision-making. For these comparisons, we focus on the two lockout conditions as we have not yet analyzed the eye movement data required to infer process in the Table condition. However, before discussing the process measures we look at outcome measures for both classification and decision-making. For these outcome measures, the Table condition provides a baseline against which to compare the effect of increased access costs on outcome.

Classification

All subjects received four practice scenarios of Argus Prime with the classification task only, followed by 8 scenarios where performance on the classification task was interrupted by the decision-making task.

An analysis of variance (ANOVA) that looked at classification performance over blocks of scenarios (scenarios 1–4, 5–8, and 9–12) yielded a significant main effect of block, F (2, 30) = 3.1, p = 0.0597, MSE = 2313. Performance improved from a mean of 56% during practice to 66% in the first four DMT scenarios to 72% in the final four DMT scenarios (see Figure 2).

All conditions were treated the same through the initial training and initial four practice scenarios. Hence, performance on the four practice scenarios provides an opportunity to determine whether the subjects in the three conditions were of roughly equal ability (as per the assumption of random assignment of subjects to condition). A second ANOVA was conducted on scenarios 1–4. As judged by the classification scores there were no differences among the three groups (F < 1). Any difference in classification performance during the 8 DMT scenarios will be regarded as due to the DMT manipulation.

A third ANOVA focused on classification performance during the 8 DMT scenarios (scenarios 5 through 12). Classification scores varied significantly between conditions [F (2, 30) = 7.0, p = 0.003, MSE = 3118.9], Table = 79%, 0-Lock = 64%, and 2-Lock = 65%. Planned comparisons showed that this difference was localized in the Table versus 0- and 2-Lock comparison (p = .0008) with no difference between 0-Lock and 2-Lock (F < 1). Performance increased from scenario 5–8 to 9–12 [F (1, 30) = 33.6, p = .0001, MSE = 1167] but this effect did not interact with DMT condition (p = 0.12).

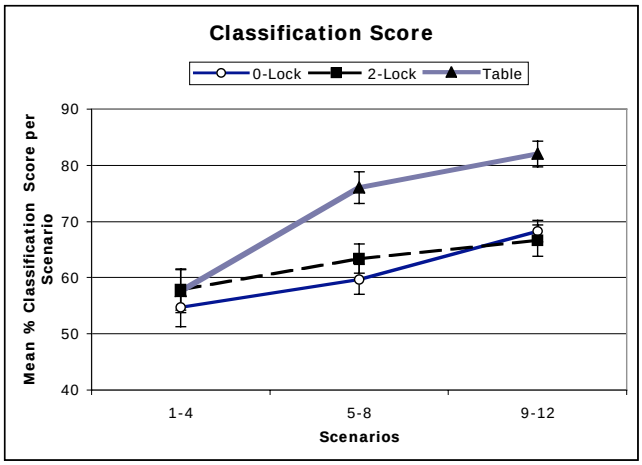


Figure 2: Classification Score across Practice scenarios (1–4) and DMT scenarios (5–8 and 9–12). (Error bars show the standard error.)

Summary of Classification Performance. The three groups were equal in their classification performance during practice (scenarios 1–4) and each continued to improve through the first and then second set of DMT scenarios (5–8 and 9–12). However, once the DMT began, the two lockout conditions performed lower on the classification task than the Table condition. As discussed below, the Table condition spent much less time on the DMT than did the lockout conditions. Hence, we believe the difference in classification performance is simply attributable to the difference in time spent by the three groups on the classification task.

Decision-Making Task (DMT)

Outcome Measures. Although performance on the decision-making task was uniformly high (see Figure 3), there was a significant difference between conditions [F (2, 30) = 10.4, p = .0004, MSE = 0.05] with Table being almost perfect (0.98) followed by 0-Lock (0.94) and then by 2-Lock (0.91). Planned comparisons showed the difference between Table and the two lockout conditions to be significant (p = .0003) and the difference between 0-Lock versus 2-Lock to be marginally significant (p = .064). The influence of number of choices (DMT-4 versus DMT-6) was also significant [F (1, 20) = 14.25, p = .0007, MSE = .043]. The interaction of number of choices with condition

was marginally significant, F (2, 20) = 2.71, p = .08, MSE = .008.

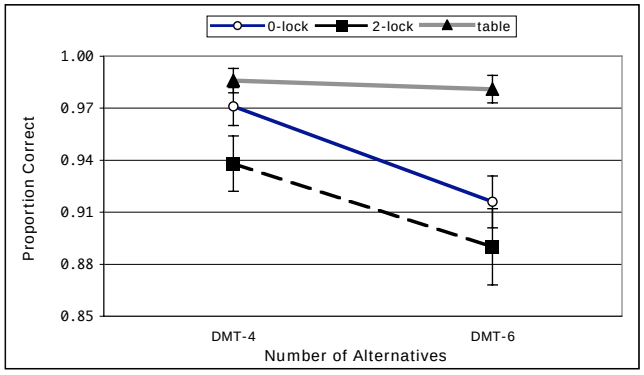


Figure 3: Proportion Correct Choices in Decision Making Task by Number of Alternatives (DMT-4 and DMT-6) and Interface Condition. (Error bars show the standard error.)

A second outcome measure is the time per DMT. This measure yields a significant effect of condition [F (2, 30) = 27.45, p = .0001, MSE = 4953] with Table spending a mere 2.7-s per DMT, 0-Lock spending 16.5-s and 2-Lock spending 23.6-s per DMT. The effect of number of targets per DMT was significant (p = .0005); however, this effect is constrained by a significant interaction of condition by DMT number [F (2, 30) = 3.21, p = 0.054, MSE = 83.8]. This interaction reflects the near asymptotic performance of Table in both DMT-4 (2.6-s) and DMT-6 (2.7-s) whereas both of the Lock groups showed a healthy increase in time from DMT-4 to DMT-6 (14.2 to 18.8 for 0-Lock and 20.8 to 26.3 for 2-Lock).

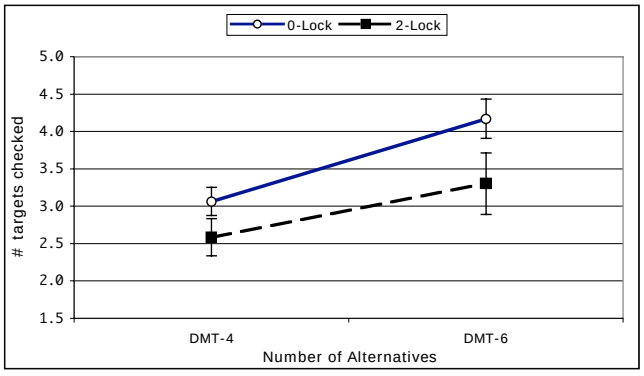


Figure 4: Number of different targets checked per DMT for 0-Lock and 2-Lock. (Error bars show the standard error.)

Of course the 2-Lock condition was locked out for 2-s for each check they made. To determine the contribution of lockout time to the difference between 0-Lock and 2-Lock we subtracted 2-s for each check or recheck made by 2-Lock. With this adjustment, time per 0-Lock versus 2-Lock was no longer significant (F < 1), leaving only a significant main effect of DMT target number [F (1, 20) = 15.45, p = .0008, MSE = 397].

Process Measures. Our first process measure is total number of targets that were checked at least once per DMT. Clearly, if subjects were doing a thorough job this number would be 4 for DMT-4 and 6 for DMT-6. Although we do not have this information for the Table condition, we do have it for the two lockout conditions (see Figure 4) and it is not surprising to find a significant main effect of number of alternatives [F (1, 20) = 33.87, p = .0001, MSE = 18.34] with DMT-4 checking an average of 2.82 targets versus 3.74 for DMT-6. However, this absolute increase masks a relative decrease as DMT-4 checked 72% of their targets versus 62% for DMT-6.

More interesting for our purposes is the difference in number checked across the two lockout conditions. Although 0-Lock checked slightly more targets than 2-Lock (3.62 versus 2.94) this difference was not significant (p = .24). No other comparisons were significant.

Our second measure of process is the number of rechecks per DMT. If a threat value was checked once, how likely was it to be rechecked? As the proportion correct and number checked varied between DMT-4 and DMT-6, we were somewhat surprised that the number of rechecks was constant (F < 1). It is somewhat less surprising that more rechecks were done for 0-Lock than for 2-Lock [F (1, 20) = 44.63, p = .0001, MSE = 4.57]. However, it does surprise us that the 2-Lock condition made almost no rechecks (see Figure 5). None of the interactions were significant (F < 1).

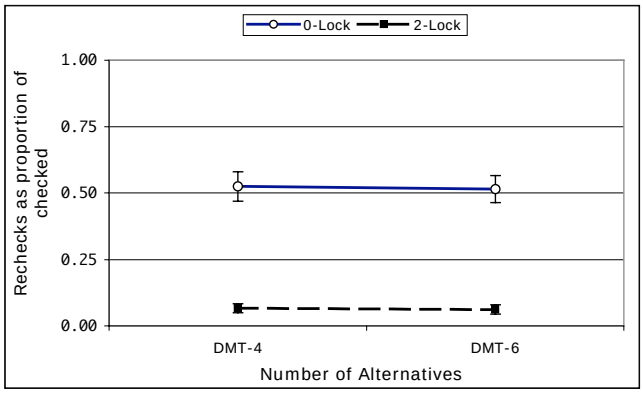


Figure 5: Rechecks as a proportion of those checked at least once. (Error bars show the standard error.)

Our third process measure is the time per check or recheck. We know from Figure 4 that 0-Lock performed more checks per DMT than 2-Lock. However, after subtracting 2-s for each check, the analysis of time per DMT showed that 0-Lock spent as much time per DMT as did 2-Lock. Hence, the time per check must be greater for 2-Lock than 0-Lock. We tested this conjecture in our final process analysis.

Time per check or recheck (after subtracting 2-s for each check made by 2-Lock) yielded a significant difference between lockout conditions [F (1, 20) = 7.98, p = 0.01, MSE = 374]. Even after subtracting 2-s per check, 2-Lock spent over twice as much time per check as 0-Lock (7.2-s versus 3.1-s). Interestingly enough, no other comparisons were significant—neither number of alternatives (DMT-4 versus DMT-6, F < 1) nor any interactions.

Discussion of Results

The Classification results suggest that the three between-

subject conditions (Table, 0-Lock, and 2-Lock) were equivalent as measured by their performance during the four practice scenarios. Performance on classification increased across the eight scenarios in the decision making part of the study. This increase suggests that subjects were still taking the classification task very seriously.

The classification score differences between conditions during scenarios 5-12 appear to reflect the differences in time spent on the decision-making task. As the Table condition spent < 3-s per DMT compared to 16-s for 0-Lock and 24-s for 2-Lock they had more time to devote to the classification task. (Although there were significant differences between groups on the mean number of DMTs per scenario, these differences were small – Table = 3.4, 0-Lock = 2.9, and 2-Lock = 2.7 DMTs per scenario.)

All groups did well on the decision-making task though the Table group did the best. Table also spent much less time per decision than did the other two groups. The time to locate and move the mouse to the screen position of the target contributed to time spent per check by each of the Lock groups. However, although the search and movement costs were similar for 0-Lock and 2-Lock, the 0-Lock group made more rechecks than did the 2-Lock and this difference was constant across DMT-4 and DMT-6. Likewise, after subtracting time for the 2-s lockout, time per check was over twice as great for 2-Lock than for 0-Lock. What factors can explain these patterns?

Discussion

The current study addresses three limits to traditional research on the tradeoff of effectiveness for efficiency in decision-making. First, rather than counting the steps required for an expert agent to execute a decision-making algorithm, we counted the actual steps taken by human subjects during the process of decision-making and measured the duration of those steps. This approach provides better evidence for what people actually do when they make a decision and exposes important intermediary steps not captured by the traditional approach. For example, the current data reveals that the 0-Lock group performs many rechecks of threat value during decision-making. The necessity to check a step more than once implies that the requirement to hold the currently highest threat value in memory while searching for another target is an important sub-step that is affected by memory limits.

Second, by varying the interface design of the decision-making task, we have begun to disentangle the cost of a step in a decision-making algorithm from the cost due to how a step is implemented in a given task environment. It is obvious that the 0-Lock and 2-Lock conditions required more visual search and more motor movement than did the Table condition. In addition, the necessity to search for the next target while holding the currently highest threat value and its target identification number in memory adds a significant cognitive cost to the lockout conditions as compared to the Table.

Of great interest to us is that the additional 2-s per check imposed on the 2-Lock condition seems responsible for the vast differences in process and the slight differences in outcome between lockout conditions. 0-Lock rechecked more targets per decision-making trial while spending half as long on each check or recheck than did 2-Lock. Although there was nothing preventing subjects in the lockout conditions from rechecking the same number of targets or spending the same amount of time per check and recheck, they differed on both of these measures. Apparently differences in lockout costs led the two groups of subjects to adopt two different solutions to the problem of comparing a new threat value to the currently highest threat value. Encoding of location is a fairly automatic outcome of locating a target on a screen (Ehret, 2002). For 0-Lock, after a target had been found once, the cost of reacquiring that target was relatively low. This low reacquisition cost led 0-Lock to adopt a strategy of minimum memory encoding (as judged by the time spent per check) and more reliance on rechecks. For the 2-Lock group, the 2-s lockout did not simply add a delay in the time to access threat value, it also added 2-s to the retention interval for previously encoded threat values as well as for previously encoded target locations. As time for retrieving an item from memory varies with its activation level, we interpret the additional time per check of 2-Lock over 0-Lock as reflecting additional time spent retrieving old information from memory as well as a longer encoding time in anticipation of a longer retention interval.

Third, our experiment helps to move decision-making studies from static to more dynamic paradigms. Time spent on the decision-making task took time away from performing the classification task. Subjects had spent the first two hours of the study learning and practicing the classification task. During the last three hours we encouraged them to continue working hard on classification and to attempt to improve their performance. The data indicate that all groups improved their classification performance throughout the 8 decision-making scenarios.

The pressure to do well on the classification task apparently led subjects in the lockout conditions to satisfice on the decision-making task. As reported earlier, only 72% of the DMT-4 targets and 62% of the DMT-6 targets were checked on any given decision-making trial.

Summary & Conclusions

The study shows that small changes in the cost of interactive behavior may lead to changes in the strategy adopted for decision-making as well as to differences in how a step in the same strategy is implemented. The low cost of scanning the target-column for threat values led the Table condition to use all of the data to achieve near perfect performance in decision-making. In contrast, the lockout conditions satisficed by using less than 100% of the target data.

Although the two lockout conditions did not differ in the amount of information accessed, the differences in lockout time led each group of subjects to implement the information access step in very different ways. The 0-Lock

group adopted an interaction-intensive procedure that made good use of perceptual–motor operations to minimize memory load. In contrast, the 2-Lock group adopted a memory-intensive procedure that maximized memory load and minimized lockout time per alternative. The different procedures adopted by the different groups reflect an adaptation of cognition, perception, and action to the cost structure or soft constraints (Gray & Fu, 2004) of the task environment.

Acknowledgments

The work reported was supported by a grant from the Air Force Office of Scientific Research AFOSR #F49620-03-1-0143. Thanks to Chris R. Sims and Vladislav D. Veksler for running subjects as well as many other contributions to this project.

References

Adelman, L., Bresnick, T., Black, P. K., Marvin, F. F., & Sak, S. G. (1996). Research with Patriot Air Defense officers: Examining information order effects. *Human Factors, 38*(2), 250–261.

Adelman, L., Miller, S. L., & Yeo, C. (2001). Testing the effectiveness of icons for supporting distributed team decision making under time pressure. *Manuscript submitted for publication.*

Benbasat, I., & Todd, P. (1996). The effects of decision support and task contingencies on model formulation: A cognitive perspective. *Decision Support Systems, 17*(4), 241–252.

Ehret, B. D. (2002). Learning where to look: Location learning in graphical user interfaces. *CHI Letters, 4*(1), 211–218.

Ehret, B. D., Gray, W. D., & Kirschenbaum, S. S. (2000). Contending with complexity: Developing and using a scaled world in applied cognitive research. *Human Factors, 42*(1), 8–23.

Fennema, M. G., & Kleinmuntz, D. N. (1995). Anticipations of effort and accuracy in multiattribute choice. *Organizational Behavior & Human Decision Processes, 63*(1), 21–32.

Gray, W. D. (2002). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. *Cognitive Science Quarterly, 2*(2), 205–227.

Gray, W. D., & Fu, W.-T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28*(3).

Gray, W. D., & Schoelles, M. J. (2003). The nature and timing of interruptions in a complex, cognitive task: Empirical data and computational cognitive models. In R. Alterman & D. Kirsch (Eds.), *25th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology, 52*, 653–683.

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science, 4*(4), 221–227.

Lohse, G. L. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology, 16*(6), 297–308.

Lohse, G. L., & Johnson, E. J. (1996). A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes, 68*(1), 28–43.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.

Payne, J. W., Bettman, J. R., & Luce, M. F. (1996). When time is money: Decision behavior under opportunity-cost time pressure. *Organizational Behavior and Human Decision Processes, 66*(2), 131–152.

Rose, J. M., & Wolfe, C. J. (2000). The effects of system design alternatives on the acquisition of tax knowledge from a computerized tax decision aid. *Accounting Organizations and Society, 25*(3), 285–306.

Schkade, D. A., & Kleinmuntz, D. N. (1994). Information displays and choice processes: Differential effects of organization, form, and sequence. *Organizational Behavior & Human Decision Processes, 57*(3), 319–337.

Schoelles, M. J. (2002). *Simulating Human Users in Dynamic Environments.* Unpublished doctoral dissertation, George Mason University, Fairfax, VA.

Schoelles, M. J., & Gray, W. D. (2001a). Argus: A suite of tools for research in complex cognition. *Behavior Research Methods, Instruments, & Computers, 33*(2), 130–140.

Schoelles, M. J., & Gray, W. D. (2001b). Decomposing interactive behavior. In J. D. Moore & K. Stenning (Eds.), *Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 898–903). Mahwah, NJ: Lawrence Erlbaum Associates.

Todd, P., & Benbasat, I. (1994). The Influence of Decision Aids on Choice Strategies - an Experimental-Analysis of the Role of Cognitive Effort. *Organizational Behavior and Human Decision Processes, 60*(1), 36–74.

Todd, P., & Benbasat, I. (1999). Evaluating the impact of DSS, cognitive effort, and incentives on strategy selection. *Information Systems Research, 10*(4), 356–374.

Todd, P., & Benbasat, I. (2000). Inducing compensatory information processing through decision aids that facilitate effort reduction: An experimental assessment. *Journal of Behavioral Decision Making, 13*(1), 91–106.

Functional Interactions Affect Object Detection in Non-Scene Displays

Collin Green (cbgreen@ucla.edu)
John E. Hummel (jhummel@psych.ucla.edu)
Department of Psychology, 1285 Franz Hall
University of California, Los Angeles
Los Angeles, CA 90095

Abstract

Two experiments suggest that functional relations influence the processing of visual stimuli. Experiment 1 demonstrated that participants are more accurate to detect targets engaged in functional interactions with related items than when they are simply surrounded by those items. Experiment 2 demonstrated that the accuracy of visual search in a non-scene display is affected when distractor items can be grouped functionally versus when distractor items are simply semantically related to each other. Overall, these data suggest that functional relations between objects affect the allocation of visual attention and by consequence, the processing of natural scenes and other structured visual stimuli.

Introduction

An important aspect of semantic knowledge about objects concerns function. The very identity of an object often hinges upon its intended use. The experiments presented here explore the possibility that participants performing object search tasks may be sensitive to functional relations among the objects being searched. This work is based on the idea that natural scenes are mentally represented in terms of the functional groups they comprise (Green & Hummel, 2004). For example, a coffee shop may be defined as a place where it is possible to make, buy, sell, and drink coffee. The objects associated with these activities (a table and chair in certain arrangement suggest dining) form the basic units of the scene definition.

While scene categories are difficult to define in terms of the objects present (the same objects may form different types of scenes by virtue of different arrangements), or in terms of the spatial layout only (the identities and meanings of objects have bearing on scene categorization), a function-based scene representation may provide a consistent, flexible, and useful definition (see Green & Hummel, 2004, for a more thorough discussion).

In both experiments presented here, the presence of functional relations (the presence of meaningful structure) in the stimulus was expected to improve performance: In Experiment 1, we expected the processing of a target object in a functional relation to be facilitated (relative to a target adjacent to the same objects, but not interacting with any of them). In Experiment 2, we expected that functionally meaningful relations would effectively unitize pairs of distractor objects, making search more efficient than when such objects must be rejected one by one.

Experiment 1

Experiment 1 investigated whether functional interactions would affect observers' ability to detect and locate target objects in non-scene displays. The experiment required observers to indicate whether a named target object was present in a masked, briefly-presented array of twelve line-drawn objects. We manipulated whether the search array contained a distractor object semantically associated to the named target, and whether the target and associated distractor (if both were present) were interacting.

In general, the addition of an associated distractor object to a search array impairs performance in visual search. Moores, Laiti, & Chelazzi (2003) found that when participants searched for a target, distractor objects semantically associated with the target had the effect of reducing accuracy and increasing latency relative to when distractors were not associated with the target.

In the current experiment, we expected a similar result. Overall performance should be lower when an associated distractor object is present in the search array relative to when no associated distractor is present (though Auckland, Cave, & Donnelly, 2004, find evidence for the opposite effect). However, it remains unclear whether such effects interact with relational information in guiding visual search. Specifically, there is reason to believe that the introduction of functional interactions between targets and associated distractors will modulate the impairment caused by target-distractor associations, to some degree.

Riddoch, Humphreys, Edwards, Baker, & Willson (2003) found that functional interactions facilitated the processing of the interacting objects. Their subjects were parietal patients who showed extinction when trying to report the names of two simultaneously-presented objects. When objects were presented together but were not interacting, the patients could reliably report the name of one object, but not both. When the objects were positioned to interact, patients showed increased ability to accurately report the name of the second object. This suggests that functional relations may in fact play a special role in the processing of visual stimuli.

Experiment 1 brought together the two results mentioned above, combining semantic associations between targets and distractors with functional interactions between targets and distractors in a single experiment. Based on the results of Moores, et al. (2003) and Riddoch, et al. (2003)

we expected that introducing a semantically associated distractor to a search array would impair target detection, but that this effect would be reduced when the target interacted with the associated distractor.

Method and Materials

Stimuli All materials were presented on a Macintosh iMac personal computer running the SuperLab application. Stimuli were composed of black and white line drawings (some taken from Snodgrass & Vanderwart (1980), others created specifically for this work) that depicted everyday objects.

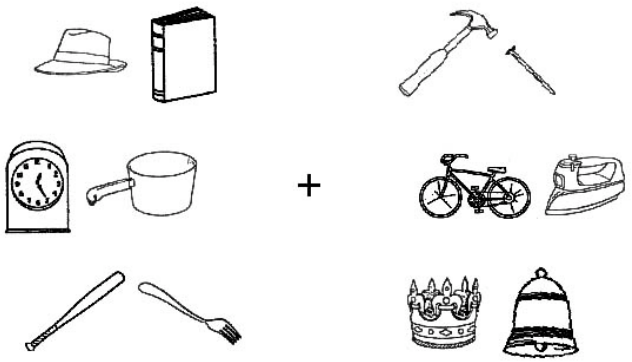


Figure 1: Typical stimulus from Experiment 1. Here the target (hammer) is interacting with a related distractor item (nail).

All search arrays employed the same basic layout (see Figure 1). A fixation cross was centered in the stimulus array. Twelve objects (each approximately 2.3° visual angle in width/height) were arranged around the fixation cross in two concentric circles. The inner circle had a radius of approximately 4.5° visual angle, and the outer circle had a radius of approximately 7.9° visual angle. Six objects were centered on the inner circle, with objects located at 45°, 90°, 135°, 225°, 270° and 315° from vertical. The six remaining objects were centered on the outer circle. Objects on the outer circle were placed horizontally in line with objects on the inner circle. In this way, the twelve objects made up six pairs. This layout placed paired objects closer to each other than to any other object in the array.

/\Critically, we manipulated the presence and position of a related distractor item in the search display. On target-present trials, the target could appear with no semantically-related distractors (the *target-only* condition), with a semantically-related distractor in a location not adjacent to the target (*non-adjacent*), paired with, but not interacting with the target (*adjacent*), or paired with and interacting with the target (*interacting*). Catch trials were presented in which the related distractor was present without the target (*distractor-only*), and in which neither the related distractor nor the target were present (*none*).

Each participant completed 24 randomly-ordered trials, with each trial using a different target object. Each participant saw one of 24 counterbalanced sets of stimuli. Across counterbalancing sets, every target object appeared in every condition equally often.

Participants Participants were 40 undergraduate psychology students at the University of California, Los Angeles. Participants took part in the experiment as part of a research requirement for a psychology course.

Procedure Participants were instructed to look for named target objects and indicate (a) whether the target object was present and if so, (b) its location in the search display. The participant was given a description of how each trial would proceed and what responses were required. The participant viewed a single practice trial, with the experimenter providing a verbal description of what was happening at each step and how the participant should respond. The experimenter emphasized that accuracy was important in all responses, but that the speed of response mattered only during the detection task.

Each trial proceeded as follows: First, a word naming the target object appeared in the center of the computer screen in black 24-point Arial font and remained on the screen until the participant pressed a key. Then, a fixation cross appeared in the center of the screen. After 750ms, the fixation cross was replaced by a search array. The search array was visible for 250ms and was subsequently masked until response or until the trial timed out (2500ms after search array onset).

The participant indicated whether or not the target object was present in the search array by making a key press (yes or no) as quickly and accurately as possible. After response, the participant was presented with an labeled layout of the search array and was asked to indicate the location of the target object appeared (or to verify that the target object did not appear) by pressing a letter on the keyboard. This response was not speeded. A 1000ms inter-trial interval during which the computer screen was blank preceded the next trial.

Results

Accuracy and response time (RT) data were analyzed using within-subjects ANOVAs. Trials upon which detection RT exceeded 2500ms were counted as errors. Error trials were excluded from all RT analyses.

Detection Accuracy Accuracy data (***d'***) from the detection task are presented in Table 1. The main effect of stimulus condition on detection accuracy only approached significance ($F(3,117) = 1.927$, $MSE = 0.621$, $p = 0.129$). However, planned comparison indicated that mean ***d'*** in the Interacting condition was significantly higher than mean ***d'*** in the Adjacent condition ($t(39) = 3.242$, $SE = 0.126$, $p = 0.002$). This comparison is the most revealing with respect to the effect of functional interactions, as the only difference

between the Interacting and Adjacent conditions is the orientation of the associated distractor object. Of the four conditions, only the Interacting condition produced performance significantly different than chance ($t(39) = 1.895$, SE = 0.168, $p = 0.0325$ one-tailed).

Detection Response Time RT data are presented in Table 1. Mean RT on the detection task did not vary across conditions ($F(3,111) = 0.537$, MSE = 48600, $p = 0.658$). No pair-wise comparisons yielded significant differences.

Table 1: Summary of response time and accuracy data from Experiment 1.

Condition	Detection d'	Detection RT	Localization accuracy
Interacting	0.319 (SE = 0.168)	1340 ms (55)	0.558 (0.048)
Adjacent	-0.091 (0.174)	1388 (61)	0.488 (0.062)
Non-Adjacent	0.024 (0.188)	1398 (54)	0.431 (0.057)
Target Only	0.078 (0.141)	1362 (57)	0.502 (0.061)

Localization Accuracy Accuracy data for the localization task are presented in Table 1. As a measure of localization accuracy, we report the probability that the correct location would be chosen given that a target was present and the observer attempted to localize the target. That is, we excluded target-absent trials, and trials where the observer made a localization response indicating that the target did not appear in the search array.

There was no main effect of stimulus condition on localization accuracy ($F(3,105) = 0.911$, MSE = 0.108, $p = 0.439$). No pair-wise comparisons were significant.

Discussion

Though weak, these results do suggest that functional relations influence the processing of objects in non-scene displays. This effect obtained even though the task did not require participants to use (or even notice) the functional relations in the stimuli. Indeed, it may be argued that functional information is not useful in this task. Only one sixth of the trials each participant saw contained a meaningful functional relation between the target and related distractor item. One would not expect the pattern of results observed were the processing of functional relations effortful.

If functional information influences the allocation of visual attention during simple search tasks, then it seems plausible that the guidance of visual attention during search of natural, structured scenes is also influenced by such information. Heuristics about what kinds of objects should appear together in scene could help the visual system to efficiently deploy attention, and facilitate the processing of scene-consistent stimuli. The finding that functional relations affect the processing of simple visual stimuli also suggests that visual representations may include abstract, functional information.

Experiment 2

Experiment 1 suggests that functional relations do influence the processing of visual stimuli during a search task. The presence of a functional relation involving the target object and an associated distractor object increased detection accuracy relative to when an associated distractor was adjacent to, but not interacting with the target. That result does little to discriminate between the possibility that interacting objects are processed more efficiently than other objects and the possibility that functional interactions capture visual attention.

Experiment 2 sought to decide between these explanations. In this experiment, distractor objects engaged in functional interactions, and the number of functional groupings in the search array was varied parametrically. If functional groups capture attention, then one would expect the addition of interacting distractor pairs to impair performance, performance suffering increasingly as more interacting pairs are added. On the other hand, if objects engaged in functional relations are processed more efficiently than objects not engaged in functional interactions, then one would expect performance to improve as more interactions are introduced to distractor objects. Search time in non-scene displays is a function of the number of distractor items present (Biederman, et al., 1988). If functionally interacting objects form perceptual groups, then adding interactions among distractors while holding the total number of display objects constant should effectively reduce the number of perceptual units that must be searched. As a result, displays with more interactions should yield superior search performance.

Method and Materials

Stimuli All materials were presented on a Macintosh iMac personal computer running the SuperLab application. Stimuli were composed of a subset of the black and white line drawings used in Experiment 1.

The experimental trials were divided into four conditions: zero functional interactions (the *0i* condition), one interaction (*1i*), two interactions (*2i*), or three interactions (*3i*). In addition to the four experimental conditions, two control conditions were run to provide baseline search performance measures. All search arrays in the four experimental conditions employed the same basic layout (see Figure 2). A fixation cross was centered in the stimulus array. Eight objects (each approx. 2.3° visual angle in width/height) were arranged around the fixation cross in two concentric circles. The inner circle had a radius of approximately 4.5° visual angle, and the outer circle had a radius of approximately 7.9° visual angle. Four objects

were centered on the inner circle, located at 45°, 135°, 225°, and 315° from vertical. The four remaining objects were centered on the outer circle. Each object on the outer circle was placed horizontally in line with an object on the inner circle. In this way, the eight objects made up four pairs. As before, objects within a pair were closer to each other than to any other object in the array.

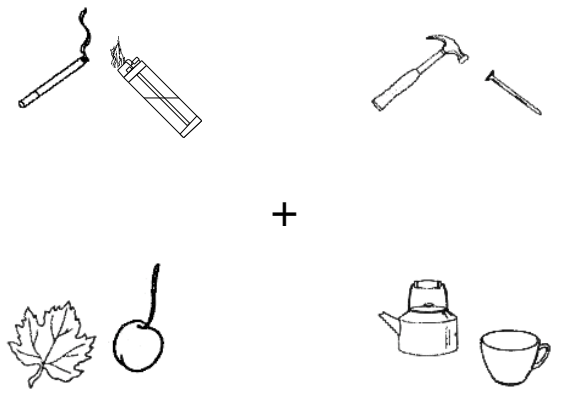


Figure 2: Typical stimulus from Experiment 2. Here, the target (leaf) is accompanied by seven distractor objects, of which four are engaged in interactions (lighter-cigarette, hammer-nail). The other distractor pair (kettle-cup) is related but not interacting. This is an example of a *2i* stimulus.

The identity and orientation of distractor items in the search display were manipulated. In each array, one distractor object was paired with the target object (or a lure). The remaining six distractor objects were organized into three pairs. The distractor objects in each pair were semantically associated and capable of entering into a functional interaction. In this experiment, participants performed trials in which there were zero, one, two, or three of these distractor pairs were actually arranged to interact.

In the control-five object condition (the *5c* condition), the target (or lure) was accompanied by four distractor objects that were unrelated to the target, and unrelated to each other, for a total of five objects in each array. In the control-eight object condition (*8c*), there were seven distractor objects unrelated to the target and each other, for a total of eight objects in each search array. Objects in the *8c* condition were arranged in accordance with the layout depicted in Figure 2. Objects in the *5c* condition occupied five of the eight positions (randomly selected) in the standard search array for this experiment.

Each participant completed 228 randomly-ordered trials. Each participant saw every target object in every condition, but no one target appeared with the same distractor objects in more than one array.

Participants 40 undergraduate psychology students at the University of California, Los Angeles participated in the experiment as part of a research requirement for a psychology course.

Procedure The procedure in Experiment 2 was identical to that of Experiment 1. On each trial participants viewed a target label and a briefly-presented search array which was masked. Participants made a speeded response indicating the presence or absence of the target object, and then a non-speeded location response. Instructions were identical to those in Experiment 1.

Results

Response time (RT) and accuracy data were analyzed using within-subjects ANOVAs. Trials on which detection RT exceeded 2500ms were counted as errors. Error trials of all types were excluded from all RT analyses.

Detection Accuracy Accuracy data (***d'***) are presented in Table 2. There was a main effect of stimulus condition (including control conditions) ($F(5,195) = 4.750$, MSE = 0.17, $p < 0.001$). There was also a significant effect of condition among the four experimental conditions (*0i*, *1i*, *2i*, and *3i*) ($F(3,117) = 3.485$, MSE = 0.164, $p=0.018$). Detection accuracy was significantly higher in the *5c* condition than in the *8c* condition ($F(1,36) = 4.152$, $p = 0.04$).

Planned comparisons indicated that accuracy in the *1i* condition was significantly worse than in the *0i*, and *3i* conditions, but not the *2i* condition. The *0i*, *2i*, and *3i* conditions were not significantly different than each other.

Trend analysis indicated that there was a significant increasing linear trend in detection accuracy across the *1i*, *2i*, and *3i* conditions ($F(1,39) = 8.684$, MSE = 0.169, $p = 0.005$). In addition, there was a significant quadratic trend across the *0i*, *1i*, *2i*, and *3i* conditions ($F(1,39) = 6.085$, MSE = 0.151, $p = 0.018$).

Detection Response Time RT data are presented in Table 2. There was no significant difference in RT across the six experimental conditions, ($F(5,200) = 1.199$, $p=0.311$).

No pairwise contrasts were significant, but participants were marginally faster to accurately respond in the *5c* condition, than in the *8c* condition ($F(1,40) = 3.137$, $p = 0.084$).

Localization Accuracy Accuracy data for the localization task are presented in Table 2. As in Experiment 1, we report the probability that the correct location would be chosen given that the target was present and the observer attempted to localize the target.

There was a significant main effect of stimulus condition on localization accuracy ($F(5,195) = 38.649$, MSE = 0.006, $p < 0.001$). There was also a main effect of stimulus condition across the *0i*, *1i*, *2i*, and *3i* conditions ($F(3,117) = 3.522$, MSE = 0.005, $p = 0.017$). Post-hoc analysis indicated that localization was significantly more accurate in the *3i* condition than in the *0i* condition, and that a

significant linear trend existed across the *0i*, *1i*, *2i*, and *3i* conditions ($F(3,117) = 9.597$, $MSE = 0.005$, $p = 0.004$).

Table 2. Summary of Reaction Time and Accuracy Data for the Detection Task in Experiment 2.

Condition	Detection RT	Detection d'	Localization accuracy
0i	812 ms (SE = 30)	1.40 (.092)	0.845 (.019)
1i	822 (34)	1.17 (.099)	0.880 (.019)
2i	833 (34)	1.36 (.115)	0.872 (.017)
3i	820 (32)	1.44 (.093)	0.899 (.014)
5c	810 (32)	1.61 (.102)	0.676 (.014)
8c	831 (32)	1.42 (.100)	0.825 (.018)

Discussion

The results of Experiment 2 suggest several possibilities. If the only meaningful difference among the four experimental conditions is the decrease in accuracy observed in the *1i* condition, then the data support an attention-capture account. Specifically, if the presence of a single functional group among distractors impairs performance, then it is possible that the participant's attention was drawn to the functional interaction (which never contained the target) and so the actual target was detected less often. One could explain the disappearance of this effect in the *2i* and *3i* conditions if the ability of a functional group to capture attention is dependent on its uniqueness in the display. Adding multiple functional interactions among distractors may "wash out" such an effect by bringing the average salience of the display elements closer to the maximum salience of any one element (i.e., the salience of a functionally interacting pair is farther from the mean salience of the display when one interaction is present than when multiple interactions are present). This explanation is somewhat unsatisfying, and the trend analyses performed suggest a more interesting alternative.

There was a significant linear trend among the *1i*, *2i*, and *3i* conditions, and a significant quadratic trend among those and the *0i* condition. The shape of these data suggest that the addition of functional interactions among distractor objects did not strictly hurt performance (as predicted by an attention-capture account), nor did the introduction of interactions strictly improve performance (as predicted by a grouping account). It seems possible that functional groups do capture attention, but that they also facilitate the processing of the objects they comprise. The drop in performance from the *0i* to *1i* conditions would indicate that the increase in search efficiency resulting from the inclusion of a single interacting pair of distractors did not outweigh the cost incurred by that pair's tendency to capture attention. However, as more interacting distractor pairs were added, the accumulated gains from more efficient processing of interacting objects improved overall performance. Performance in the *3i* condition was only numerically superior to that in the *0i* condition, but if the linear trend across the *1i* to *3i* conditions is extrapolated, then one can imagine that the continued addition of functional interactions among distractor would produce performance reliably exceeding that in the *0i* condition. In fact, in the extreme case, imagine searching for a random object among a disorganized array of distractors versus searching for the same object among distractors organized into a coherent scene. Both intuition and empirical evidence suggest that search will be more efficient in the latter case (Loftus & Mackworth, 1978; Hollingworth & Henderson, 2000).

Finally, localization accuracy data from Experiment 2 suggest that the extraction of spatial information about objects in a stimulus is more efficient when that stimulus includes functional relations between objects. Notably, it was organization of non-target objects that led to this advantage. In Experiment 1, a similar (but unreliable) advantage was observed for localization of target objects that engaged in functional interactions.

Conclusions and Future Directions

The current work was motivated by the idea that natural scenes are mentally represented in terms of the functional groups they comprise (Green & Hummel, 2004). Experiment 1 suggested that objects are more easily detected or identified when they were interacting with related distractors as compared to when they were not interacting with related distractors, or when no related distractors were present. Experiment 2 showed an interesting non-monotonic pattern associated with the introduction of functional interactions to the display. The addition of a single interaction seemed to impair detection, while performance improved with addition of subsequent interactions.

The data from Experiments 1 and 2 suggest that functional interactions may have two effects on visual search: 1) single functional groups may capture attention; 2) objects in functional groups may be processed more efficiently than objects not engaged in interactions.

Existing data from eye movement studies are not consistent with the first claim. A number of studies (De Graef et al, 1990; Henderson et al., 1999; Loftus & Mackworth, 1978) suggest that visual information (e.g., local contrast, spatial frequency, color) is the main determinant of fixations early in natural scene viewing. Evidence indicates that during natural viewing, scene-consistent objects are fixated more rapidly than inconsistent objects, but that this type of semantic information only mediates eye movements after the first several fixations on a scene (De Graef et al., 1990). In short, semantic

information does not seem to influence early fixations, playing a role only later in visual scanning. This suggests that functional groups (which include abstract semantic information) should not capture attention.

Alternatively, the data from these experiments may be explained as an effect of familiarity with canonical arrangements of objects. Because people are routinely exposed to objects arranged in functionally meaningful ways, some other non-attentional influence may be involved. Specifically, the existence mental symbols that represent entire familiar functional groupings may influence the preattentive grouping of a visual stimulus array and lead to faster processing of the objects therein. The existence of perceptual groupings based on functional information (which were not considered in those studies) might affect early attentional guidance in way that is not easily understood if one is looking for effects of semantic consistency only.

Empirical and computational work have been used to study the effects of perceptual grouping on visual search with basic perceptual stimuli (e.g., colored shapes, oriented lines). Some models of search account for effects of perceptual grouping better (and more naturally) than others. For example, the Spatial and Object Search (SOS) model (Grossberg, Mignolla, & Ross, 1994) places perceptual grouping processes at the center of visual search operations. Grouping processes take place pre-attentively in the SOS model, an assumption consistent with a number of empirical results (e.g., Humphreys, et al., 1989).

An important aspect of the SOS model with respect to the functional grouping hypothesis is that its perceptual grouping mechanisms are linked to spatially-invariant representations of objects. SOS allows knowledge about objects to influence perceptual grouping (presumably, so that objects form perceptual units, instead of collections of object parts or features). An extension of SOS might employ representations above the level of objects (e.g. representations of functional groups) to make contact with grouping processes as well. A preattentive grouping mechanism linked to representations of functional groups might yield effects like those observed in Experiments 1 and 2.

Whether functional groups are perceptual groups remains to be established, but the results of Experiments 1 and 2 suggest that familiar functional relations (interactions between objects in a visual scene) may be an important component of visual processing. Current work addresses the possibility that functional groups are in fact perceptual objects.

Acknowledgments

The authors thank Irv Biederman, Steve Engel, Keith Holyoak, Zili Liu, Brian Stankiewicz, members of the LISA lab, and the CogFog group for helpful discussion and comments on this work. Also, thanks to Jerlyn Tolentino and Erica Weiss for many hours of effort in the lab. This work was supported by NIH/NINDS NRSA F31-NS43892-02.

References

Biederman, I., Blickle, T.W., Teitelbaum, R.C. & Klatsky, G.J. (1988). Object search in non-scene displays. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14(3)*, 456-467.

De Graef, P., Christiaens, D., & D'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research/Psychologische Forschung, 52(4)*, 317-329.

Green, C. & Hummel, J.E. (2004). Relational Perception and Cognition: Implications for Cognitive Architecture and the Perceptual-Cognitive Interface. In B.H. Ross (Ed.): *The Psychology of Learning and Motivation, Vol. 44*. San Diego: Academic Press. 201-226.

Grossberg, S., Mignolla, E., & Ross, W.D. (1994). A neural theory of attentive visual search: Interactions of boundary, surface, spatial, and object representations. *Psychological Review, 101(3)*, 470-489.

Henderson, J. M., Weeks, P. A., & Hollingsworth, A. (1999). The Effects of Semantic Consistency on Eye Movements During Complex Scene Viewing. *Journal of Experimental Psychology: Human Perception & Performance, 25(1)*, 210-228.

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition, 7(1/2/3)*, 213-235.

Humphreys, G.W., Quinlan, P.T., & Riddoch, M.J. (1989). Grouping processed in visual search: Effects with single and combined feature targets. *Journal of Experimental Psychology: General, 118*, 258-279.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception & Performance, 4(4)*, 565-572.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience, 6(2)*, 182-189.

Riddoch, M.J., Humphreys, G.W., Edwards, S., Baker, T. & Willson, K. (2003). Seeing the action: Neuropsychological evidence for action-based effects on object selection. *Nature Neuroscience, 6(1)*, 82-89.

Snodgrass, J.G. & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6(2)*, 174-215.

A Multiple-Trace Memory Model Exhibiting Realistic Retrieval Dynamics

Collin Green (cbgreen@ucla.edu)
Aniket Kittur (nkittur@ucla.edu)
Department of Psychology, 1285 Franz Hall
University of California, Los Angeles
Los Angeles, CA 90095

Abstract

A process model of human memory dynamics is proposed as an implementation of Kittur, Green, & Bjork's (2004) mathematical model. Both models are based on an ideal information processing approach, in which an item's accessibility is based on the predicted future need of that item. The proposed model is an adaptation of the multiple-trace architecture of Hintzman's MINERVA2 model (Hintzman 1984; 1986; 1988). We present simulations of complex spacing and practice dynamics encompassing the mechanics of Bjork and Bjork's (1992) New Theory of Disuse, which accounts for diverse phenomena such as massed vs. spaced practice and spontaneous recovery. In addition, we show how the model explains and simulates time-dependent serial position effects (such as the shift from recency to primacy with delay and time-invariant recency effects). The model's potential as a tool for exploring the relationship between the content of items in memory and more general memory dynamics is also discussed.

Memory as a System for Predicting Need

Kittur, Green, & Bjork (2004) described a mathematical model of memory dynamics inspired by Bayesian statistics. The model is driven by the assumption that memory approximates an ideal information processor, keeping memory items accessible to the degree that they are likely to be needed in the future (see Anderson, 1989 for further rationale on this approach). The predicted future need for an item is computed by the model based on the pattern of past retrievals for that item and the time since it was last retrieved. This is best illustrated by analogy.

Imagine that Book A has been checked out of a library once a month for the past year. Book B, on the other hand, has been checked out every week for the last month but never prior to that. If the librarian was forced to choose which of the two books should be kept readily available, the best choice would change over time. Initially, the librarian would probably keep Book B more readily retrievable as it has been needed frequently in the recent past (possibly, an instructor has assigned reading from this book for a class project); however, after a month has passed with neither book being required, the librarian would likely decide that Book A should be more accessible, given its history of being required at regular, if infrequent, intervals.

The Kittur, Green, & Bjork (2004) model functions in a similar way. It calculates the probability that an item will be needed given three key pieces of information: the average interval between past retrievals; the number of times the item was retrieved in the past; and the time since it was last retrieved. The use of these elements allows for a distinction between item accessibility and item storage, a key insight of the New Theory of Disuse (NTD) (Bjork & Bjork, 1992). The model was inspired by and provides a potential algorithmic basis for the NTD and the complex memory dynamics it explains.

The New Theory of Disuse

The NTD accounts for a variety of effects in the human memory literature. The NTD includes the following assumptions about memory (see Bjork & Bjork, 1992):

1)Memory items are associated with two distinct "strengths": a *storage strength* (SS) and a *retrieval strength* (RS). SS indicates how well-learned an item is (that is, the accumulated history of an item is reflected in its SS). RS, on the other hand, indicates how readily accessible an item is for retrieval. RS alone determines the probability that an item will be successfully recalled from memory. SS does not directly influence memory performance, but has important implications for memory dynamics over time[1].

2)SS does not decrease. SS grows during study or retrieval events as a decelerating function of the current SS. That is, all else being equal, items with low SS benefit from study or retrieval events more than items with high SS. The total storage strength across all items in memory is therefore unbounded. Changes in SS are dependent on both RS and SS. An item gains SS as a decelerating function of its current SS, and as a decelerating function of its current RS.

3)RS increases and decreases. As with SS, an item gains RS as a result of study or retrieval events. When the item is not being studied or retrieved, such as when other items are being attended, RS decreases. As a result, gains in RS for one item necessarily result in a loss of RS for the other (unstudied) items in memory, though these are not necessarily changes of the same magnitude. Changes in RS are dependent on both RS and SS: Gain in RS due to a retrieval or study event is a decelerating function of current RS, and an increasing function of current SS. Conversely, RS loss in is faster the larger the current RS is, and slower with larger SS.

4)Generally, retrieval events are more potent than study events. Increments in both SS and RS are larger when an item is retrieved versus when it is studied.

[1] The two-strength theory espoused by NTD and implemented in MNEM is an important difference between it and other related need-based models, such as Anderson's ACT-R (1989). We are currently exploring testable differences between the models.

The postulation of two separate strengths whose magnitudes influence each other is at the core of NTD's account of retrieval and memory dynamics.

The MNEM Model

Many models of human memory employ a strategy that assumes each item stored in memory is represented by a single memory trace. For example, the studied item "horse" would be instantiated as a single mental symbol, and further exposure to "horse" would serve to strengthen or heighten the activation (or gain—i.e. sensitivity to excitation) of that symbol. However, such models struggle with the problem that no two exposures to an item are identical: the spatial, temporal, or subjective context of encoding is variable. Additionally, changes in attention or effort may occur during different exposures to an item and attributes of a stimulus that are important at one point may be more or less important at some point in the future. Multiple trace models of memory are better suited to deal with variable encoding, in that they do not assume that all encodings of an item are linked to a single representation. Such models also do not assume a mechanism for reconciling variable encodings with unitary representation.

MNEM (*M*emory *N*eed *E*xpectation *M*odel), like MINERVA2 and other multiple trace models, works on the assumption that every instance of encoding lays down a new memory trace in the long-term store. If a single stimulus is encoded on multiple occasions (studied and re-studied), then MNEM creates and stores separate traces for each encoding event. Because of random information loss during encoding events (see below), recording new traces for every instance produces variability in the Long Term Memory (LTM) representations of a repeated item. This variability occurs in addition to any variability introduced by context, environment or attention, which may also be introduced to the model.

Representation

The representations upon which MNEM operates are simple, and are adapted from Hintzman's (1984) MINERVA2 model. Each trace in MNEM is an ordered vector of size *n*, with each element taking on the values of -1, 0, or +1. The elements can be thought of as corresponding to specific feature dimensions (e.g. "redness", "roundness", "chair-ness", etc.), with values indicating the absence of a specific feature (-1), the presence of the feature (+1), or a lack of information about the feature (0). The format is open to other interpretations, of course.

Consideration of the history of a memory item depends on the ability to examine past encodings of that item. It is unlikely, however, that any two memory traces are actually identical. That is, identifying instances of trace *T* is simple when literal copies of *T* are stored in several LTM locations, but it is more likely that LTM traces containing the same information are encoded with different contexts, or with different features emphasized. Instead of a single strengthened trace *T*, or many literal copies of *T*, we may store several traces similar to *T*: *T'*, *T"*, etc. As such, it is necessary to resolve some ambiguity about *which* traces in LTM should be counted as instances of a single item.

MNEM uses a specific similarity metric to evaluate the similarity of two memory traces. Borrowing again from the MINERVA2 model, the similarity of an LTM trace *T* to some probe trace *P* is calculated as follows:

$$S(P,T) = \left(\frac{1}{N_R}\right)\sum_{j=1}^{n} P(j)T(j), \qquad (1)$$

where *n* is the number of elements in the trace and N_R indicates the number of relevant features in the pair of traces. Relevant features are defined as features for which at least one of the two traces contains a non-zero value; in other words, if neither trace contains any information about the presence or absence of a feature, then the feature is not counted as relevant. This similarity function approximates a dot product calculated on the feature sets of the two traces, *T* and *P*.

This representational format is admittedly simplistic, though one advantage of this simplicity is that it requires few assumptions. In fact, the MNEM model requires only two key properties of its representations: they must be amenable to *some* systematic similarity metric, and they must be combinable in a systematic way.[2]

Any representational format that meets these requirements is compatible with the MNEM model. This flexibility makes it amenable to incorporation into diverse cognitive architectures, where other components of the system might necessarily place more serious constraints on the representational format. (As an example, as ordered one-dimensional vectors may be too limiting for representing relational structures, an alternative and appropriate format could be used provided it satisfies the above requirements). That human memory traces satisfy these constraints is a common (if sometimes implicit) assumption among cognitive scientists. The ability to judge the degree to which two stimuli are similar is fundamental to human cognition. Schema abstraction, generalization, and conceptual blending are psychological phenomena that may involve the combination of two or more stimuli to form a composite or abstraction.

Architecture

Like MINERVA2, MNEM has two components: a working memory (WM) and a LTM. WM consists of a buffer that holds a single trace. All inputs to and outputs from LTM are buffered by WM. Traces that are in WM may be encoded into LTM, and information retrieved from LTM is brought into WM.[3]

[2] The second requirement is not important for simulating retrieval dynamics, but will be critical in future work when the model is used to generate content from a set of memory traces.

[3] The authors have not attempted to model WM except in the sense that it is a buffer between the world and LTM. In MNEM, multiple traces are not maintained simultaneously, and no attention is required for WM trace maintenance. WM traces may be overwrit-

LTM is simply a collection of memory traces that have been encoded from WM. The current model imposes no (theoretical) limit on the capacity of (the number of traces in) LTM. Each LTM trace is associated with an index. The indices are assigned in the order with traces are encoded into LTM, so that traces encoded earlier have lower indices. The authors consider this equivalent to incorporating spatio-temporal tags on memory traces. Extensions of MNEM may attempt to use a subtler form of spatio-temporal tagging.[4]

Operations

Encoding The encoding operation of the model is relatively simple, and amounts to little more than copying a WM trace into LTM. As discussed, MNEM assumes that variability exists in encoding process (i.e. information is randomly lost during encoding).

The accuracy of encoding depends on a learning rate parameter (L) which indicates the independent probability that any trace feature will be properly encoded (where $0 < L \leq 1$). For example, when $L = 0.7$, seven out of ten features in a trace are accurately copied into the LTM trace (on average). The features that are not properly encoded result in gaps in LTM information (zeros are written into the LTM trace where 1 or -1 existed in the WM trace). During the encoding process, information is only lost, not distorted: a 1 in the WM trace is never erroneously encoded as a -1 in the LTM trace, nor vice versa. This encoding procedure is taken directly from MINERVA2.

Every encoding event yields a new LTM trace, regardless of whether the content of the new trace is redundant with existing LTM traces. The similarity of traces is not considered during the encoding process.

Retrieval Calculating RS for an item is also relatively straightforward. The main complication arises from determining which LTM traces should be considered in the RS calculation when variability exists among different encodings of an item. To address this problem, MNEM "marks" the traces in LTM whose similarity to the probe item exceeds a set criterion. (This criterion similarity, C_s, is a parameter of the model). For example, if C_s is set to 0.75, then only traces for which $S(P,T) \geq 0.75$ will be marked for inclusion in the RS calculation. Once LTM traces are marked, the mean retention interval between them is calculated:

$$\overline{RI(P)} = \frac{1}{(N_m - 1)} \sum_{i=2}^{N_m} [index(M_i) - index(M_{i-1})], \quad (2)$$

where P is the item for which RS is being calculated, M_i is the i^{th} marked LTM trace, and N_m is the total number of marked LTM traces.[5] The *index()* operator simply indicates that the model is using the LTM index for a trace and not the trace itself.

This mean interval is multiplied by the number, or "base rate", of similar instances in LTM. The base rate (*BR(P)*) is equal to the number of marked traces in LTM:

$$BR(P) = N_m. \quad (3)$$

The product of the mean retrieval interval and the base rate[6] is divided by the size of the current retrieval interval, which is the number of time steps that have elapsed since the last marked item was encoded:

$$CI(P) = index(P) - index(M_{\max}), \quad (4)$$

where $index(M_{max})$ indicates the index of the timestep during which a marked trace was most recently encoded. Also, *index(P)*, the time index for the encoding of the current item, is simply set to the index of the current timestep (which is equal to the number of traces in LTM plus one: $N_{ltm} + 1$).

In summary, RS can be characterized thus:

$$RS(P) = \frac{\overline{RI(P)} * BR(P)}{CI(P)}. \quad (5)$$

That is, the accessibility of an item P, is equal to the product of the average retention interval between instances like P in LTM and the number of such instances, divided by the interval that has elapsed since the last instance of P occurred.[7]

In order to compare forgetting curves, it is necessary to normalize *RS(P)*. This is accomplished by finding the ratio of logarithm of *RS(P)* to the maximum value that *RS(P)* obtains for an item (immediate recall).[8] (Because the log may be negative, we add one to both numerator and denominator for convenience). In all simulations, this ratio is reported as RS. That is:

$$RS_{reported}(P) = \frac{\log(RS(P) + 1)}{\log(RS_{\max}(P) + 1)}. \quad (6)$$

---

ten, but this is the only way that information is "lost" from MNEM's WM.

[4] The authors are currently exploring the incorporation of a context vector into encoded representations, or giving individual traces an activation value which would be initialized to some maximum at encoding, and would decay over time. In the latter strategy, the activation value would represent a trace's "age" for purposes of calculating RS. The RS calculation would consider the difference between the activations of two traces. This approach remains untested, but seems promising in that the decay function would likely be non-linear, decelerating as it approaches zero. This being the case, two traces equally displaced in absolute time would become less discriminable with age.

[5] When only a single trace in LTM is marked, the average retention interval is defaulted to a value of 1.

[6] This product is the closest analog to SS in MNEM: $SS(P) = \overline{RI(P)} * BR(P)$. Note that unlike RS, SS is strictly increasing with additional study, and is not subject to decay. SS influences changes in RS, most importantly by retarding the loss of RS over time (see Figure 3).

[7] This definition of RS is at the core of the Kittur, Green, and Bjork (2004) model, which exhibits the same memory dynamics in a single-trace architecture.

[8] See Pavlik & Anderson (2003) for rationale on scaling forgetting using the maximum (current) activation of a trace.

The result of normalization is that immediate recall yields a reported RS of 1, and any delay in recall produces an RS between 0 and 1. This allows for comparison of forgetting curves in terms of probability of recall.

Trace Composition The formation of composite traces from a set of LTM traces is also important in this model. We have specified how one may calculate the RS of a specific item in LTM, but retrieving useful information from the LTM store is another matter entirely. MINERVA2 includes a mechanism that uses similarity to weight traces in LTM, and forms a composite "echo" by averaging these weighted traces. MNEM employs a similar strategy, but instead of *all* traces in LTM, only those that exceed the similarity criterion are weighted and averaged. While this is an important aspect of the model, and may allow simulation of important memory phenomena (e.g. encoding specificity, context effects, etc.) the details of this operation are not directly relevant to the retrieval dynamics discussed here, and we will leave them for another time.

Simulation Results

The NTD was conceived to "post-dict" a number of memory effects. In the previous discussion of that theory, behavioral correlates of RS and SS were noted. MNEM implements the same relationships between RS and SS and its performance is similar to that of humans on a variety of memory tasks.

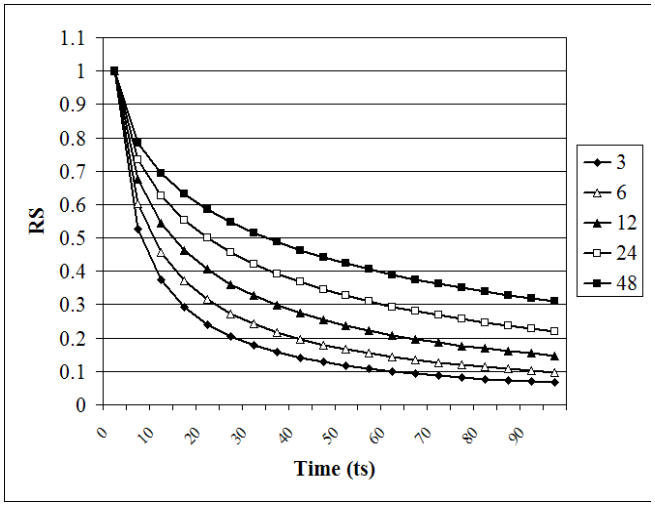


Figure 1: Forgetting curves for items studied three times each, with inter-item intervals of 3, 6, 12, 24, or 48 timesteps.

Forgetting Over Time

MNEM displays forgetting curves that closely resemble those of human subjects. Behavioral data suggest that the probability of recalling a once-studied item declines as a function of the retention interval. More specifically, access to an item declines as a function of intervening experience (Thorndike, 1914; McGeoch, 1932; Bjork & Bjork, 1992).

NTD postulates that probability of recall is linked to RS only, but that changes in RS are mediated by SS. The particular rate of forgetting for an item is influenced by the frequency of exposure to an item (Melton, 1967; Krueger, 1929), as well as the interval between exposures (Peterson, Hillner, & Saltzman, 1962; Whitten & Bjork, 1977). MNEM captures the general shape of forgetting curves, and simulates frequency and spacing effects observed in human data.

In simulation, a single item *A* is studied according to various schedules. At various delays, the RS of *A* is calculated, which indicates the probability that it would be recalled at that interval since last study. To simulate the passage of time without study or retrieval events, a randomly generated memory trace is encoded into LTM on each timestep.[9] Note that in simulation, the calculation of RS does not affect the state of LTM.

The simulated practice schedules vary in the number of exposures of *A*, as well as in the spacing of exposures. Forgetting curves generated by MNEM for items studied with equal frequency, but different inter-item intervals are shown in Figure 1. Figure 2 shows forgetting curves for items studied at equal intervals, but with different frequencies.

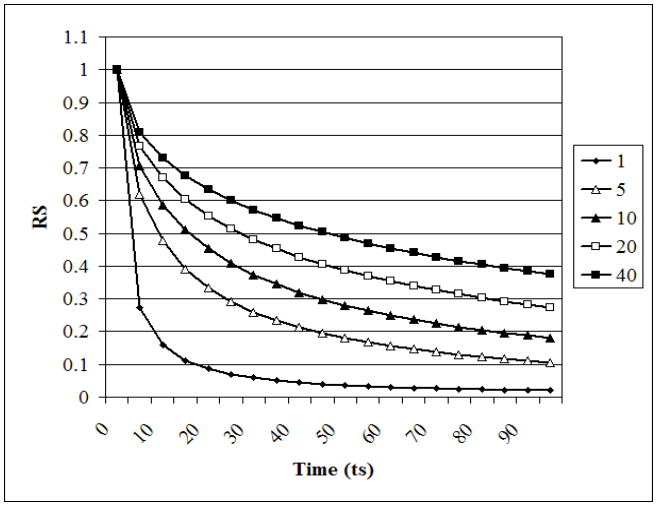


Figure 2: Forgetting curves for items studied with equal spacing (3 timesteps between exposures) and frequencies of 1, 5, 10, 20, or 40 exposures.

Spacing and frequency effects are important aspects of human memory in that they give rise to more complicated phenomena. For example, in some circumstances an old habit may be replaced with a new behavior, only to re-emerge at a later time, a phenomenon known as spontaneous recovery (Estes, 1955; Koppenaal, 1963).

[9] It is worthwhile to note that the noise introduced to the LTM system is relatively unconstrained. In fact, the same method that generates the "studied" trace for these simulations is used to generate the "noise" traces that are interpolated between the study event and the sampling of RS.

NTD and the MNEM model account yield spontaneous recovery as a natural consequence of different forgetting rates. Simulation data in Figure 3 show spontaneous recovery. Item *A* represents an old response that has been learned over a long period of time. Item *B* is a new response intended to replace *A*. As *B* is acquired, *A*'s RS decays substantially. However, we observe that *A* gains an advantage after a certain delay. If *B* is not practiced, the larger SS of item *A* yields a shallower forgetting curve. The decay of RS is slower for trace *A* than for trace *B* and the curves cross over. The older habit will remain more accessible thereafter.

Primacy & Recency

Primacy and recency are well-known memory phenomena. When a list is studied, items that appeared early in the study list are more recallable than items near the middle of the list. Primacy effects have been attributed to covert rehearsal between study presentations (Glenberg et al., 1980). Effectively, subjects create extra study opportunities in the gaps between item exposures.

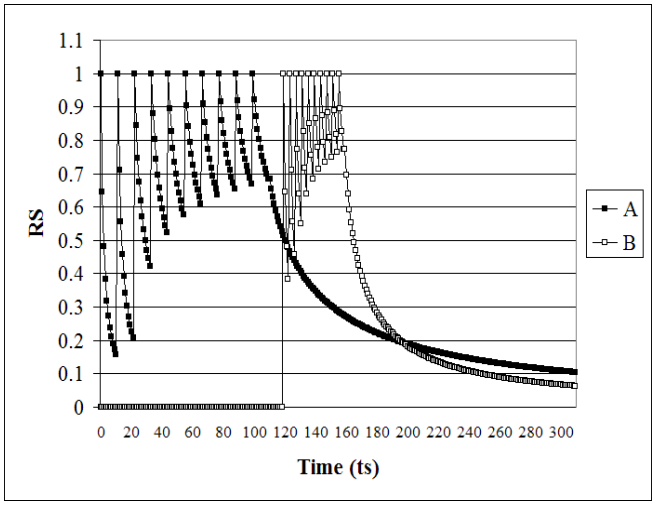


Figure 3: Spontaneous recovery of response *A* occurs after learning response *B*. This is due to a larger increase in response *A*'s SS at reminding, owing to a lower RS at that time.

Similarly, items presented near the end of the list are also recalled better than mid-list items. Recency results from the relatively short retention interval between study and test. Prior work has demonstrated that there is a shift from recency to primacy over increasing retention intervals (Craik, 1970; Knoedler, Hellwig, & Neath, 1999). The MNEM model shows similar behavior.

In simulation, a list of 20 items is studied, with five timesteps between study events. Between trials, the simulated subject is assumed to perform covert rehearsal on some of the items presented so far, in order, for the duration of the interval. This strategy lasts for a limited number of presentations (three, in this simulation), at which point the simulated subject is assumed to become overwhelmed by the number of items and therefore abandons the covert rehearsal strategy. Beyond this point, inter-trial intervals are filled with random traces, as in previous simulations.[10]

At the end of the study phase, the RS for each of the 20 items is calculated at five different retrieval intervals. The serial position curves that result are shown in Figure 4. Three features are notable: the prominence of recency effects in immediate recall; the presence of primacy in all serial position curves; the shift of recency to primacy as the dominant pattern in the data as the retention interval grows.

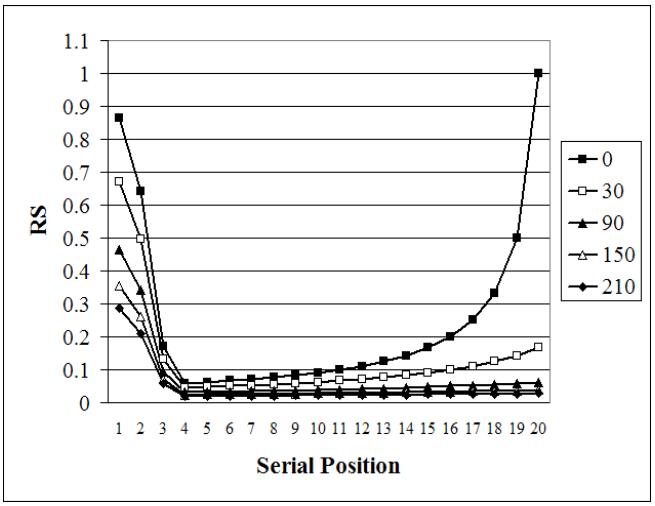


Figure 4: Serial position curves at delays of 0, 30, 90, 150, and 210 timesteps. Note the rapid decay of recency effects relative to the slower decay of primacy.

The recency effects observed in simulation share a subtle property with human behavioral data: time-invariance. Some data from humans suggest that the magnitudes of recency effects follow a ratio rule (Glenberg et al, 1980; 1983; Bjork & Whitten, 1974). This phenomenon was described mathematically by Bjork & Whitten (1974). Specifically, recency effects scale with the log of the ratio of mean presentation interval divided by the current retention interval:

$$recency \propto \log\left(\frac{\overline{RI(P)}}{CI(P)}\right) \quad (7)$$

This behaviorally-derived ratio rule is inherent in the MNEM model (see Kittur, Green & Bjork, 2004). Figure 5 shows serial position curves for various ratios of mean retentional interval to current retention interval.

Conclusions and Future Directions

The model described here shows memory dynamics that are consistent with human behavioral data. Forgetting curves, spacing and frequency effects, and serial position curves are generated in simulation by following the assump-

[10] Glenberg, et al., (1980) observed that primacy effects were eliminated when participants were prevented from performing cumulative rehearsal on early list items.

tions of NTD, and allowing items to accumulate independent SS and RS.

The relative simplicity of this model (and its more general mathematical formulation in Kittur, Green, & Bjork, 2004), makes it a useful tool for exploring subtle issues in memory and generating concrete experimental predictions. There is potential to extend our understanding of retrieval dynamics to a greater diversity of memory phenomena by manipulating the content of the memory traces used in simulation. For example, MNEM provides a natural platform for exploring the influence of inter-item associations on memory dynamics. In addition, MNEM may prove to be informative on issues surrounding schema abstraction, categorization, and other arenas where knowledge content is an issue. Context, encoding specificity, and variability effects may also be amenable to analysis with this model in the future.

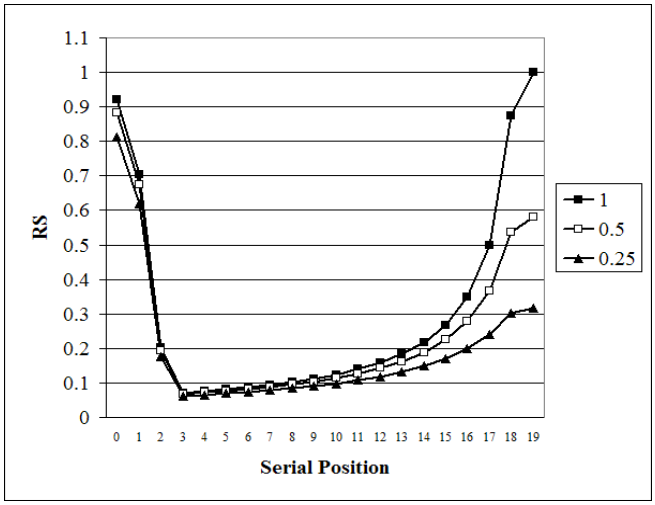


Figure 5: The magnitude of recency effects in MNEM scale with the ratio of mean retention interval to current retention interval. Serial position curves are shown for spacing to retention interval ratios of 1, 0.5, and 0.25.

Acknowledgments

The authors would like to thank Robert Bjork, Tom Wickens, John Hummel, Russ Poldrack, Barbara Knowlton and the CogFog group for comments on this work. Also, thanks to the members of the LISA lab for their feedback and support.

References

Anderson, J. R. (1989). A rational analysis of human memory. In H. L. Roediger, III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 195-210). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Bjork, R.A. & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Schiffrin (Eds.), *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes* (Vol. 2, pp.35-67). Hillsdale, NJ: Erlbaum.

Craik, F.I.M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior, 9*, 143-148.

Estes, W.K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*, 145-154.

Glenberg, A.M., Bradley, M.M., Stevenson, J.A., Kraus, T.A., Gretz, A.L., Fish, J.H., & Turpin, B.N. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning & Memory, 6*, 355-369.

Hintzman, D.L. (1984). MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93(4)*, 411-428.

Hintzman, D.L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95(4)*, 528-551.

Kittur, A., Green, C., & Bjork, R.A. (2004, July). A need-based model of human memory retrieval. Poster presented at the 112th Annual Meeting of the American Psychological Association: Honolulu, HI.

Knoedler, A.J., Hellwig, K.A., & Neath, I. (1999). The shift from recency to primacy with increasing delay. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*, 474-487.

Koppenaal, R.J. (1963). Time changes in strength of A-B, A-C lists: Spontaneous recovery? *Journal of Verbal Learning and Verbal Behavior, 2*, 310-319.

Krueger, W.C.F. (1929). The effects of overlearning on retention. *Journal of Experimental Psychology, 12*, 71-78.

McGeoch, J.A. (1932). Forgetting and the law of disuse. *Psychological Review, 39*, 352-370.

Melton, A.W. (1967). Repetition and retrieval from memory. *Science, 158*, 532.

Pavlik Jr., P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In F. Detje, D. Dorner & H. Schaub, *Proceedings of the Fifth International Conference of Cognitive Modeling* , 177-182.

Peterson, L.R., Hillner, K., & Saltzman, D. (1962). Time between pairings and short-term retention. *Journal of Experimental Psychology, 64*, 550-551.

Thorndike, E.L. (1914). *The psychology of learning*. New York: Teachers College.

Whitten, W.B.H., & Bjork R.A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16*, 465-478.

Using Physical Theories to Infer Hidden Causal Structure

Thomas L. Griffiths
gruffydd@psych.stanford.edu
Department of Psychology
Stanford University

Elizabeth R. Baraff & Joshua B. Tenenbaum
{liz_b,jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

We argue that human judgments about hidden causal structure can be explained as the operation of domain-general statistical inference over causal models constructed using domain knowledge. We present Bayesian models of causal induction in two previous experiments and a new study. Hypothetical causal models are generated by theories expressing two essential aspects of abstract knowledge about causal mechanisms: which causal relations are plausible, and what functional form they take.

Everyday reasoning draws on notions that go far beyond the observable world, just as modern science draws upon theoretical constructs beyond the limits of measurement. The richness of our naive theories is a direct result of our ability to postulate hidden causal structure. This capacity to reason about unobserved causes forms an essential part of cognition from early in life, whether we are reasoning about the forces involved in physical systems (e.g., Shultz, 1982), the mental states of others (e.g., Perner, 1991), or the essential properties of natural kinds (e.g., Gelman & Wellman, 1991).

The central role of hidden causes in naive theories makes the question of how people infer hidden causal structure fundamental to understanding human reasoning. Psychological research has shown that people can infer the existence of hidden causes from otherwise unexplained events (Ahn & Luhmann, 2003), and determine hidden causal structure from very little data (Kushnir, Gopnik, Schulz, & Danks, 2003). This work has parallels in computer science, where the development of a formalism for reasoning about causality – causal graphical models – has led to algorithms that use patterns of dependency to identify causal relationships (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). It has recently been proposed, chiefly by Gopnik, Glymour, and their colleagues (Glymour, 2001; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004), that these algorithms may also explain how people infer causal structure.

A fundamental issue in explaining how people infer causal relationships is accounting for the interaction between abstract causal knowledge and statistical inference. The classic debate between approaches that emphasize cause-effect covariation and those that emphasize mechanism knowledge (e.g., Newsome, 2003) turns on this issue. Causal graphical models provide a language in which the problem of causal induction can be formally expressed. However, conventional algorithms for inducing causal structure (e.g., Pearl, 2000; Spirtes et al., 1993) do not provide a satisfying account of either the roles of causal knowledge or statistical inference, or their interaction. These algorithms use tests of statistical independence to establish constraints that must be satisfied by causal structures consistent with the observed data. No knowledge of how causal mechanisms operate, or the functional form of relationships between cause and effect, enters into the inference process. As we argue below, such knowledge is necessary to explain how people are able to infer causal structure from very small samples, and to infer hidden causes from purely observational data. Constraint-based methods are also unable to explain people's graded sensitivity to the strength of evidence for a causal structure, because they reason deductively from constraints to consistent structures.

We will present a rational account of human inference, Theory-Based Causal Induction, which emphasizes the interaction between causal knowledge and statistical learning. Causal knowledge appears in the form of causal theories, specifying the principles by which causal relationships operate in a given domain. These theories are used to generate hypothesis spaces of causal models – some with hidden causes, some without – that can be evaluated by domain-general statistical inference. We will use this framework to develop models of people's inferences about hidden causes in two physical systems: a mechanical system called the stick-ball machine (Kushnir et al., 2003), and a dynamical system involving an explosive compound called Nitro X.

Theory-based causal induction

Our account of causal induction builds on causal graphical models, extending the formalism to incorporate the abstract knowledge about causal mechanism that plays an essential role in human inferences. We will briefly introduce causal graphical models, consider how prior knowledge influences causal induction, and describe how we formalize the contribution of causal theories.

Causal graphical models

Graphical models represent the dependency structure of a joint probability distribution using a graph in which nodes are variables and edges indicate dependence. The graphical structure supports efficient computation of the probabilities of events involving these variables. In a *causal* graphical model the edges indicate causal dependencies, with the direction of the arrow indicating the direction of causation, and they support inferences about the effects of interventions (Pearl, 2000). An intervention is an event in which a variable is forced to hold a value, independent of any other variables on which it might depend. Intervention on a variable A is denoted do(A). Probabilistic inference on a modifified graph, in which incoming edges to A are removed, can be used to assess the consequences of intervening on A.

The structure of a causal graphical model implies a pattern of dependency among variables under observation and intervention. Conventional algorithms for inferring causal structure use standard statistical tests, such as Pearson's χ^2 test, to find the pattern of dependencies among variables, and then deductively identify the structure(s) consistent with that pattern (e.g., Spirtes et al., 1993). These "constraint-based" algorithms can also exploit the results of interventions, and often require both observations and interventions in order to identify the hidden causal structure. Gopnik, Glymour, and colleagues have suggested that this kind of constraint-based reasoning may underlie human causal induction (Glymour, 2001; Gopnik et al., 2004; Kushnir et al., 2003).

The role of causal theories

Constraint-based algorithms for causal induction make relatively little use of prior knowledge. While particular causal relationships can be ruled out a priori, there is no way to represent the belief that one structure may be more likely than another. Furthermore, the use of statistical tests like χ^2 makes only weak assumptions about the form of causal relationships: these tests simply assess dependency, regardless of whether a relationship is positive or negative, deterministic or probabilistic, strong or weak.

Several researchers (e.g., Shultz, 1982) have argued that knowledge of causal mechanism plays a central role in human causal induction. Mechanism knowledge is usually cited in arguments against statistical causal induction, but we view it as critical to explaining how statistical inferences about causal structure are possible from sparse data. Knowledge about causal mechanisms provides two kinds of restrictions on possible causal models: restrictions on which relationships are plausible, and restrictions on the functional form of those relationships. Restrictions on plausibility might indicate that one causal structure is more likely than another, while restrictions on functional form might indicate that a particular relationship should be positive and strong.

These restrictions have important implications for causal induction algorithms. If all structures are possible, both observations and interventions are typically required to identify hidden causes, and without strong assumptions about the functional form of causal relationships, samples must be relatively large. With limitations on the set of possible causal structures and expectations about functional form, however, it is possible to make causal inferences from just observations and from small samples – important properties of human causal induction.

Using causal theories in causal induction

The causal mechanism knowledge that is relevant for statistical causal inference may be quite abstract, and may also vary across domains. Much of this knowledge may be represented in intuitive domain theories. In contrast to Gopnik et al. (2004), who suggest that causal graphical models are the primary substrate for intuitive theories, we emphasize the role of intuitive theories at a more abstract level, providing restrictions on the set of causal models under consideration. Such restrictions cannot be represented as part of a causal graphical model: causal graphical models express the relations that hold among a finite set of propositions, while causal theories involve statements about all relations that could hold among entities in a given domain.

Formally, we view causal theories as *hypothesis space generators*: a theory is a set of principles that can be used to generate a hypothesis space of causal models, which are compared via Bayesian inference. The principles that comprise a theory specify which relations are plausible and the functional form of those relations. These principles articulate how causal relationships operate in a given domain, but need not identify the mechanisms underlying such relationships: all that is necessary for causal induction is the possibility that some mechanism exists, and expectations about the functional form associated with that mechanism. This vague and abstract mechanism knowledge is consistent with the finding that people's understanding of causal mechanism is surprisingly shallow (Rozenblit & Keil, 2002).

In the remainder of the paper, we will demonstrate how Theory-Based Causal Induction can be used to explain human inferences about hidden causes in physical systems. Different systems require different causal theories. We will examine inferences in a mechanical system, the stick-ball machine (Kushnir et al., 2003), and in a dynamical system, Nitro X, which we explore in a new experiment. When reasoning about these systems, people infer hidden causal structure from very few observations, and are sensitive to graded degrees of evidence.

The stick-ball machine

Kushnir et al. (2003) conducted two experiments in which participants had to infer the causal structure

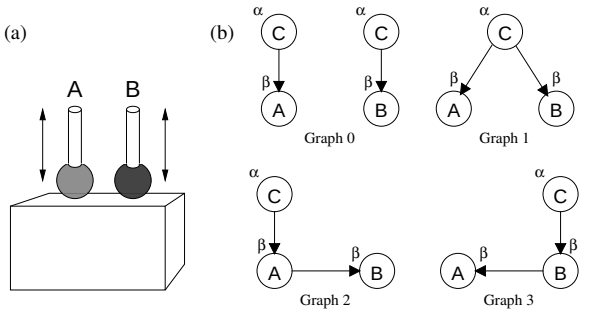


Figure 1: (a) A stick-ball machine. (b) Graphs indicating potential causal structures for the stick-ball machine. Nodes A and B correspond to the two balls, nodes marked C are hidden causes.

of a physical system, the "stick-ball machine", consisting of two colored balls (A and B) mounted on sticks which could move up and down on a box (see Figure 1(a)). The mechanical apparatus moving the balls was concealed, keeping the actual causal relationship unknown. In both experiments, all participants were familiarized with the machine, and told that if one ball caused the other to move it did so "almost always". This probabilistic causal relation was demonstrated by showing the two balls move together four times, an event we denote $4AB$, and A moving alone twice, $2A\bar{B}$. There were three test conditions in Experiment 1, seen by all participants. In the *common unobserved cause* condition, participants saw $4AB$, and four trials in which the experimenter intervened, twice moving A with no effect on B, $2\bar{B}|\text{do}(A)$, and twice moving B with no effect on A, $2\bar{A}|\text{do}(B)$. In the *independent unobserved cause* condition, participants saw $2A\bar{B}$, $2\bar{A}B$, $1AB$, $2\bar{A}|\text{do}(B)$, and $2\bar{B}|\text{do}(A)$. In the *one observed cause* condition, participants saw $4B|\text{do}(A)$ and $2\bar{B}|\text{do}(A)$. Experiment 2 replicated the *common unobserved cause* condition, and compared this with a *pointing control* condition in which interventions were replaced with observations ($4AB$, $2\bar{A}B$, $2A\bar{B}$). The order of conditions and trials within conditions was randomized across participants. In each condition, participants identified the underlying causal structure by indicating graphs similar to those shown in Figure 1(b). The results of both experiments are combined in Figure 2. One causal structure was chosen by the majority of people in each condition – Graph 1 in the *common unobserved cause* condition, Graph 0 in the *independent unobserved causes* condition, Graph 2 in the *one observed cause* condition, and Graph 0 in the *pointing control*.

The results of these experiments provide two challenges to constraint-based accounts. First, people are able to make inferences from small samples – in many cases, far less data than might be required for all relevant χ^2 tests to yield results consistent with the appropriate causal structure. Second, people's

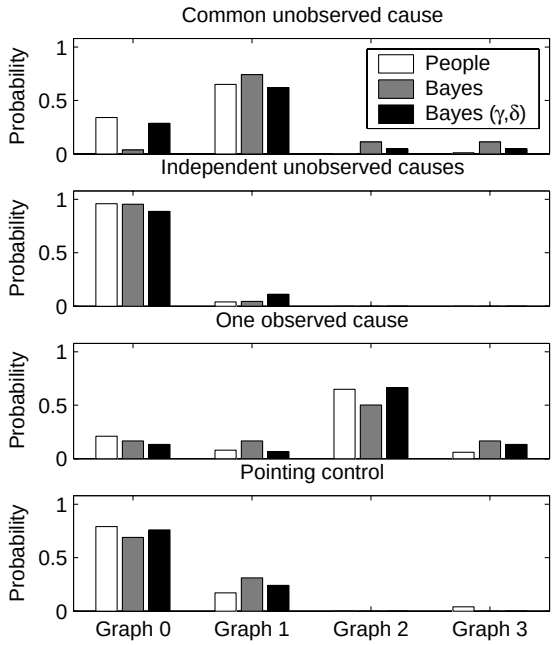


Figure 2: Results of Kushnir et al. (2003), shown with predictions of Bayesian models.

judgments reflect a sensitivity to graded degrees of evidence: in the *independent unobserved causes* condition, over 95% of participants chose Graph 1, while only 60-80% of people chose the most popular structure in the other conditions. This was not simply a consequence of a preference for Graph 0 – the same structure was less popular in the *pointing control* condition, suggesting that there is a difference in the evidence that the data provide for Graph 0 in these two conditions. Constraint-based algorithms are not sensitive to graded degrees of evidence: a causal structure is either consistent or inconsistent with the pattern of dependencies in a dataset.

A theory-based account

Our model of the stick ball machine uses a physical theory that contains three principles:

1. Balls never move without a cause.
2. A hidden cause moves with probability α.
3. A moving cause moves its effect with probability β.

If we add the restrictions that every ball has a single cause and hidden causes never have causes (but can move themselves, per Principle 2), we obtain the four structures shown in Figure 1(b). The principles of the physical theory place strong constraints on the functional form of the causal relationships identified in this structure, allowing us to compute the probability of events involving A and B for each graphical structure, as shown in Table 1.

Given a dataset D, we compute a posterior probability distribution over these structures, $P(\text{Graph } i|D)$, combining prior probabilities,

Table 1: Event probabilities for causal structures

Event	Graph 0	Graph 1	Graph 2
AB	$(\alpha\beta)^2$	$\alpha\beta^2$	$\alpha\beta^2$
$\bar{A}B$	$\alpha\beta(1-\alpha\beta)$	$\alpha\beta(1-\beta)$	0
$A\bar{B}$	$\alpha\beta(1-\alpha\beta)$	$\alpha\beta(1-\beta)$	$\alpha\beta(1-\beta)$
$\bar{A}\bar{B}$	$(1-\alpha\beta)^2$	$1-2\alpha\beta+\alpha\beta^2$	$1-\alpha\beta$
$A\|\text{do}(B)$	$\alpha\beta$	$\alpha\beta$	$\alpha\beta$
$B\|\text{do}(A)$	$\alpha\beta$	$\alpha\beta$	β

Note: Probabilities for Graph 3 are the same as those for Graph 2, exchanging the roles of A and B.

$P(\text{Graph } i)$, with the probability of the observed data under each structure, $P(D|\text{Graph } i)$, using Bayes' rule:

$$P(\text{Graph } i|D) \propto P(D|\text{Graph } i)P(\text{Graph } i)$$

$P(D|\text{Graph } i)$ is the product of the probabilities of the individual events making up D, which can be obtained from Table 1.

If we assume a uniform prior for $P(\text{Graph } i)$, the causal theory leaves two parameters unspecified: α, the probability of a hidden cause moving on a given trial, and β, the probability that a moving cause moves its effect. We set β empirically, via a small experiment. We showed 10 participants a computer simulation of the stick-ball machine, and reproduced the familiarization trials used by Kushnir et al. (2003): participants were told that when A causes B, it makes it move "almost always", and were shown that A moved B on four of six trials. We then asked them how often they expected A would move B. The mean and median response was that A would move B on 75% of trials, so we used $\beta = 0.75$.

Figure 2 shows the predictions of the Bayesian model with $\alpha = 0.47$. The model gave a correlation of $r = 0.94$ with the data, and correctly predicted the most common response in each condition. The model also admits graded degrees of evidence, with the observations and interventions in the *independent unobserved causes* condition providing stronger evidence for Graph 0 than the observations in the *pointing control*. The model departs from people's judgments in one case, failing to predict the minority preference for Graph 0 in the *common unobserved cause* condition. This disparity could have many explanations, such as a default preference for independence between objects, or differences in the salience of different data types and causal structure. For instance, interventions may be weighted higher than observations by a factor of γ, and hidden common causes may receive only a fraction $1/\delta$ of the prior probability accorded to other structures. Figure 2 shows an almost-perfect fit ($r = 0.99$) for such a model, Bayes (γ,δ), with $\gamma = 4, \delta = 2, \alpha = 0.4$. Further experiments will be necessary to determine whether these sorts of psychological variables play a role in the process of causal induction.

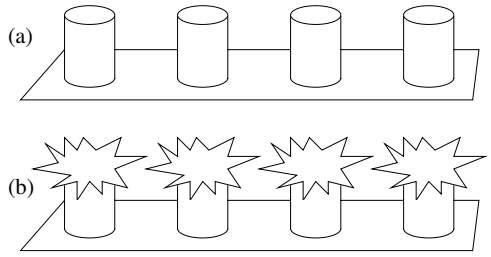


Figure 3: (a) Four cans of the extremely unstable compound Nitro X. (b) A simultaneous explosion.

Nitro X

To provide a further demonstration of the importance of graded degrees of evidence and the ability to infer hidden causes from very little data, we conducted an experiment that tested people's ability to infer the causal structure of a dynamical physical system. Our experiment presents a more severe inductive challenge than the tasks considered by Kushnir et al. (2003), as it requires inferring a hidden common cause from just a single observation, with no verbal cues that such a structure might exist. In the experiment, we introduced people to a novel substance, Nitro X, and illustrated its dynamics: cans of Nitro X could spontaneously explode, and could detonate one another after a time delay that was a linear function of spatial separation, as would be expected from the slow propagation of pressure waves. We then presented them with the *simultaneous* explosion of several cans, without the delays characteristic of pressure waves propagating from one can to the next. We expected that people would see this suspicious coincidence as evidence for some kind of hidden common cause, such as an external force shaking the table. We varied the number of cans, m, to see whether the magnitude of the coincidence had an effect on people's inference to a hidden cause.

Method

Participants Participants were 64 members of the MIT Brain and Cognitive Sciences subject pool, split evenly over four conditions ($m = 2, 3, 4, 6$).

Stimuli The stimuli were pictures of cans sitting on a table, presented on a computer screen. A new set of cans was shown on each trial, and by the end of the trial all cans on the screen had exploded, demonstrated by cartoon explosion graphics like those shown in Figure 3.

Procedure The experiment consisted of three familiarization trials and five test trials. The familiarization trials introduced the participants to Nitro X. In the first trial, participants were told that Nitro X is very unstable, and this was demonstrated by the experimenter tapping a can and the can exploding. In the second trial, participants saw two cans of Nitro X, the experimenter tapped one can,

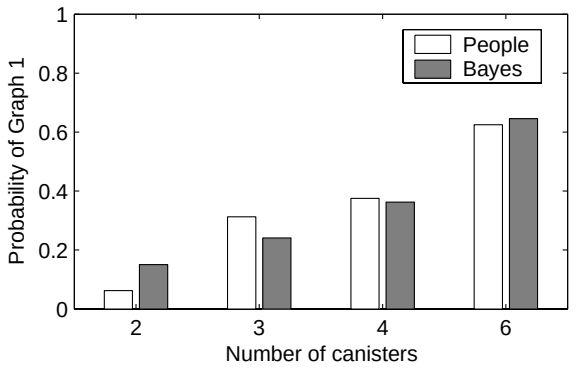


Figure 4: Results of the Nitro X experiment.

which exploded, and the can next to it exploded shortly afterwards. On the third trial, participants were again reminded about the instability of Nitro X, and saw a single can explode without any action by the experimenter, after waiting for a few seconds.

The first two test trials were identical for all four conditions, and both involved four cans exploding in a causal chain, with a delay between successive explosions. In the third test trial, the number of cans in the display was varied, $m = 2, 3, 4$ or 6, depending on condition. After a brief delay, all of the cans exploded simultaneously. The last two test trials allowed the participants to interact with Nitro X by tapping, and will not be discussed further here.

After each test trial, participants were given a sheet of questions for each test trial. These sheets gave three options:

1. The first can exploded spontaneously. That explosion caused the other cans to explode, in a chain reaction.
2. Each can exploded spontaneously, all on its own. There was no causal connection between them.
3. Neither of the above is a likely explanation. Please write a plausible alternative here.

The order of the first two options was counterbalanced, but the third option was always last.

Results and Discussion

For all trials, two rates examined the written responses of participants choosing the third option above, and were in 100% agreement in classifying all such responses as indicating a hidden cause. Over 95% of participants correctly identified the causal chain in the first two trials. The proportion of participants identifying a hidden cause on the third trial, with the simultaneous explosion, is shown in Figure 4. There was a statistically significant effect of m, $\chi^2(3) = 11.36$, $p < 0.01$. The number of cans influenced whether people inferred hidden causal stucture, with most people seeing two cans as independent but six as causally related.

Constraint-based algorithms cannot explain our results. If we imagine that time is broken into discrete intervals, and a can either explodes or does not explode in each interval, then we can construct a contingency table for each pair of cans. Statistical significance tests will identify pairwise dependencies among all cans that explode simultaneously, provided appropriate numbers of non-explosion trials are included. The existence of a hidden common cause is consistent with such a pattern of dependency. However, as a result of reasoning deductively from this pattern, the evidence for such a structure does not increase with m: a hidden common cause is merely consistent with the pattern for all $m > 2$.

This experiment also illustrates that people are willing to infer hidden causal structure from very small samples – just one datapoint – and from observations alone. Constraint-based algorithms cannot solve this problem: while a hidden common cause is consistent with the observed pattern of dependency, causal structures in which the cans influence one another cannot be ruled out without intervention information. People do not consider this possibility because they have learned that the mechanism by which cans influence one another has a time delay. Further situations in which the temporal properties of causal relationships influence causal induction are described by Hagmayer and Waldmann (2002).

A theory-based account

The results of the Nitro X experiment are easy to model: any increasing function of the number of cans would be sufficient. Our goal in modeling these data is to illustrate how Theory-Based Causal Induction extends to a system with non-trivial dynamics and different causal mechanisms, and to show that inferences to hidden causes from the smallest possible sample – a single observation – can have a physically plausible and statistically rational explanation.

We model the explosion times of cans by assuming that at each infinitesimal moment, there is a certain probability that the can will explode. This assumption means that the explosion time of each can follows a Poisson process, with a "rate parameter" determining the probability of explosion at each moment. We set the rates using the following principles:

1. A can explodes spontaneously at rate α.
2. A hidden cause becomes active at rate γ.
3. At the moment a hidden cause is active, a can influenced by that cause explodes at rate $\alpha + \beta$.

A complete theory of Nitro X would need to include further principles stating the functional form of the causal relationship between cans, encoding the fact that this relationship involves a time delay. We have omitted these principles because they do not directly affect the inference to a hidden cause when all explosions are simultaneous.

This theory generates a large number of possible causal structures, with hidden causes influencing various subsets of the cans. We will focus on the two structures shown in Figure 5: Graph 0, in which all

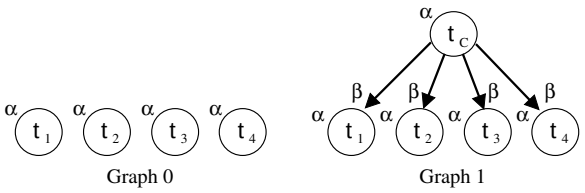


Figure 5: Graphs indicating potential causal structures for the Nitro X experiment.

cans explode spontaneously, is the "null hypothesis" for any inference concerning hidden causes, while Graph 1, in which all cans are also influenced by a hidden cause, gives the highest probability to a simultaneous explosion. These structures are defined on variables representing the time at which cans explode, $t_1, \ldots, t_m$, and the time the hidden cause becomes active, t_C. The inference to a hidden common cause is modeled by computing the posterior probability $P(\text{Graph } 1|T)$, where $T = \{t_1, \ldots, t_m\}$. In a simultaneous explosion, all t_i take the same value, t.

It follows from the theory outlined above that for Graph 0, each t_i is an independent Poisson process with rate α, which gives $P(T|\text{Graph } 0) = \alpha^m \exp\{-m\alpha t\}$. For Graph 1, t_C follows a Poisson process with rate γ. Conditioned on t_C, each t_i is a Poisson process with rate α, except at the moment when the hidden cause becomes active, at which point the rate is $\alpha + \beta$. Computing $P(T|\text{Graph } 1)$ requires integrating over all values of t_C, which we approximate by choosing t_C to maximize $P(T|t_C)$:

$$\begin{aligned} P(T|\text{Graph } 1) &= \int_0^\infty P(T|t_C)P(t_C)\, dt_C \\ &\approx \gamma(\alpha+\beta)^m \exp\{-m\alpha t - \gamma t\} \end{aligned}$$

Applying Bayes' rule, it follows[1] that $P(\text{Graph } 1|T)$ is a sigmoid function of m,

$$P(\text{Graph } 1|T) = \frac{1}{1+\exp\{-gm-b\}}$$

for $g = \log\frac{\alpha+\beta}{\alpha}$ and $b = \log\frac{P(\text{Graph } 1)}{P(\text{Graph } 0)} + \log\gamma - \gamma t$.

The model predicts that increasing m should increase $P(\text{Graph } 1|T)$ for any positive values of α and β, as this results in a positive gain, g. The theory involves four parameters: α, β, γ, and $P(\text{Graph } 0)$. Since these four parameters are not identifiable – multiple sets of parameter values are consistent with the same sigmoid function – we set the parameters of the sigmoid g and b. Using $g = 0.58$ and $b = -2.90$ gives $r = 0.958$, and the predictions shown in Figure 4. These parameters indicate $\beta = 0.79\alpha$ and an initial preference for Graph 0.

Our theory-based approach explains why the number of cans involved in a simultaneous explosion should influence the evidence for a hidden cause, but is clearly not the only model compatible with these data. However, our analysis exposes the rational basis for human judgments, and makes further intuitive predictions that we are in the process of testing. For example, the $-\gamma t$ term in the expression for b indicates that, all other things being equal, decreasing the time before a simultaneous explosion increases the evidence for a hidden cause.

Conclusion

Explaining human causal induction requires supplementing the formal methods developed in computer science with the causal domain knowledge that people possess. We have shown that using physical theories to inform rational statistical inference makes it possible to explain how people infer hidden causal structure from such limited data. We anticipate that the same framework, using appropriately modified causal theories, can shed light on inferences about hidden causes in other domains.

Acknowledgments We thank T. Kushnir and L. Schulz for helpful discussions. TLG was supported by a Stanford Graduate Fellowship, JBT by the P.E.Newton chair.

References

Gelman, S. A. and Wellman, J. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition*, 38:213–244.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. MIT Press, Cambridge, MA.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111:1–31.

Hagmayer, Y. and Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory and Cognition*, 30:1128–1137.

Kushnir, T., Gopnik, A., Schulz, L., and Danks, D. (2003). Inferring hidden causes. In *Proceedings of the 25th Conference of the Cognitive Science Society*.

Luhmann, C. C. and Ahn, W.-K. (2003). Evaluating the causal role of unobserved variables. In *Proceedings of the 25th Conference of the Cognitive Science Society*.

Newsome, G. L. (2003). The debate between current versions of the covariation and mechanism approaches to causal inference. *Philosophical Psychology*, 16:87–107.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press, Cambridge, UK.

Perner, J. (1991). *Understanding the representational mind*. MIT Press, Cambridge, MA.

Rozenblit, L. R. and Keil, F. C. (2002). The missunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26:521–562.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(Serial no. 194).

Spirtes, P., Glymour, C., and Schienes, R. (1993). *Causation prediction and search*. Springer-Verlag, NY.

[1] A full derivation of this result is available at http://www-psych.stanford.edu/~gruffydd/reports/nitrox.pdf

Evidence of Muddy Knowledge in Reaching for the Stars: Creating Novel Endings for Event Sequences

Rebecca Grimes-Maguire (rebecca.grimes@ucd.ie)
Mark T. Keane (mark.keane@ucd.ie)
Department of Computer Science, University College Dublin
Belfield, Dublin 4, Ireland

Abstract

This experiment examines people's ability to invent creative outcomes to simple event sequences. We report a study where participants are given everyday event descriptions and asked to describe either a predictable outcome (Predictable group) or a creative outcome (Creative group). Following the Creative Cognition approach (Finke, Ward & Smith, 1992), we expected that though those instructed to be creative might generate novel and interesting outcomes, they would also be bound by their knowledge of the outcomes that typically occur. The results support this prediction, in that while the Creative group manifested more inventive variability in their outcomes relative to the Predictable group, their proposed outcomes still overlapped in part with those of the Predictable group. These results show that although creativity may take people beyond their knowledge, they can never fully break free from that knowledge.

Introduction

Creativity is like reaching for the stars with your feet firmly stuck in the mud of everyday life. While the creative act takes us beyond what an individual or a society has thought before, it seems to be inextricably constrained by what is known already (see e.g., Boden, 1995; Finke, Ward & Smith, 1992; Perkins, 1981; Sternberg, 1999). In science, new theories come from reactions to old paradigms, but still work from the same methodologies and findings of previous decades. In the arts, similar reactions to the conventions of a previous age occur, though often themes and materials remain the same. In this paper, we examine this interplay between creativity and the constraints placed on it by prior knowledge by studying people's generation of novel outcomes to conventional event sequences. We often need to imagine unconventional or novel outcomes to typical happenings (e.g., in launching a new product or assessing the impact of new technologies). Yet, we know of little work which examines people's creativity in such situations.

The idea that creativity is often constrained by prior knowledge has been strongly and convincingly argued for in Finke et al.'s (1992) Creative Cognition Approach and their 'Geneplore model'. This approach has been supported by several appropriate empirical demonstrations. For example, Ward (1994) asked participants to imagine and then draw creatures that live on a distant planet very unlike earth. This simple creative task revealed that converse to the instructions, almost all the participants produced animals very like the ones living on this planet, in that they exhibited features such as bilateral symmetry, external organs (e.g. eyes, ears) and appendages (e.g. legs, tails). Ward concluded that the participants were constrained by their experience of real world animals and could not deviate easily from such prototypes. This type of influence from background knowledge has been subsequently demonstrated in studies on the generation of novel product names (Rubin, Stoltzfus & Wall, 1991), ideas for traffic improvement (Marsh, Landau & Hicks, 1997), and even the generation of non-words (Marsh, Ward & Landau, 1999). Haught & Johnson-Laird (2003) have reported similar findings in a task where people were asked to come up with a creative sentence incorporating two or three specific nouns. They found that people were quite restricted in their output; for example, the words 'lion' 'strawberry' and 'harp' tended to result in similar sentences, such as "The lion was playing the harp while eating the strawberry". In Ward's (1994) view, these "structured imagination" effects occur because, when faced with a problem whose solution requires creativity, people tend to take the path of least resistance by retrieving domain specific information or an existing solution (whether this is an experimenter provided example or self-generated from previous knowledge) and then attempt to modify the old construct in some novel way.

In the present study, we look at the constraint placed by background knowledge in a task that deals with novel sentence generation involving script-like scenarios (Bower, Black & Turner, 1979; Schank & Abelson, 1977). In our task, people are presented with typical event sequences that have incomplete, but predictable, endings (see Appendix). These scenarios either involve conventional events that proceed uninterrupted (e.g. "Matthew had wanted to quit his job for months. One day he walked into his boss's office…"), or events that are interrupted by some surprising event or state (e.g. "The cup of coffee was balanced on the arm of the chair. Suddenly, Richard sneezed….). In both cases, the main manipulation was to ask people to come up with a creative outcome to the sequence. In the remainder of this paper, we detail this study and sketch the properties of a computational model that might capture the effects found.

Proposing Creative Outcomes to Events

Following the Creative Cognition Approach, we expected that our creative-ending generation task would manifest the constraining influence of prior, background knowledge in our participants. In the experiment, there were two main groups, the Predictable and Creative groups. Both groups were given the same set of event sequence materials

(divided into Unfolding and Surprise scenarios). However, the Predictable group was asked to "think of a typical ending to the scenario…", whereas the Creative group was asked to "think of a creative ending to the scenario…". Thus the design was a 2 x 2 one, with Group being a between-participants variable (Predictable or Creative) and Scenario being a within-participants variable (Unfolding or Surprise).

The main prediction was that the Creative group would generate many of the same outcomes as the Predictable group, as they would be constrained by their background knowledge of the typical endings of these events. However, we thought that something additional would also be included in these endings, giving them an added novel twist. So, in the specific measures we used, we expected more elaborate endings in the Creative group (i.e., more propositions generated), but we also expected that some of these propositions would overlap with those produced by the Predictable group (i.e., propositions reflecting a common ending). To put it another way, the Predictable group's endings should strongly overlap with those generated by the Creative group.

We had no apriori grounds for expecting a difference between the Unfolding and Surprise scenarios, though they do appear to be distinct categories. In the Unfolding scenarios the sequence of actions proceeds unchecked in a predictable way. In the Surprise scenarios one state or sequence of actions is cut across by another sequence of actions. Interestingly though, in the Surprise scenarios the interrupting sequence is also predictable, it's a "typical surprise" (e.g., a poorly balanced object being knocked).

Experiment

Method

Participants Thirty native English-speaking undergraduate psychology students from University College Dublin volunteered for this experiment.

Materials Twenty-four scenarios involving typical everyday event sequences (see Appendix). All scenarios had two sentences and required a third to complete the sequence. The 24 materials consisted of three types of sequences: 8 Unfolding items, 8 Surprise items and 8 filler items. The Unfolding items described two events/states in a typical sequence with a predictable outcome (e.g. "Cathy saw the cake in the window. She hadn't had lunch that day…"). The Surprise items described one event/state that was interrupted by another event/state leading to a predictable outcome (e.g., "The little boy played by the edge of the pond. Suddenly he slipped on some moss…").

Design In the 2 x 2 (Group x Scenario) design, participants were randomly assigned to one of two between-participant groups, Predictable (N=15) and Creative (N=15). All participants received the same 24 scenarios, which were presented in a different random order to each participant.

Procedure Each participant was given a booklet containing all the materials, the first page of which included instructions. The items were presented so that only one scenario appeared on each page (one sentence per line with a prompt stating 'your ending:' in the space below each scenario). Participants in the Predictable group were given the instructions to "think of a typical ending to the scenario", whilst those in the Creative group were asked to "think of a creative ending" to be written as a concluding sentence. In the Creative group participants were also asked to describe a "creative turn of events, not just the use of creative language" so as to avoid a misinterpretation of the instructions. An earlier pilot showed that without this instruction, some people just produced purple-prose versions of typical endings rather than truly novel endings.

Scoring Participants' responses in completing the presented sentences were firstly rated for level of *creativity*. Then the responses were analysed into propositions. As a further measure of creativity, we wished to examine the *diversity* and *richness* of responses made, but we also examined the *commonalities* between responses to determine if there was any overlap across the different conditions.

To measure *creativity*, following Haught & Johnson-Laird's (2003) procedure, two judges independently rated each sentence (blind to condition) on a 7-point scale, with a score of 1 denoting a highly uncreative sentence and a score of 7 denoting an extremely creative sentence.

To measure *diversity*, for each item we categorised the distinct propositions used in people's endings. So for example, for the "Cathy looked at the cake in the shop window. She hadn't had lunch that day…" item, there were three distinct classes of responses given as endings:

(1) Cathy gets the cake.
(2) Cathy decides not to get the cake for some reason (e.g., diet).
(3) Cathy was hungry.

To measure the *richness* of the responses, we scored the endings produced for their word length and the number of different events mentioned in them. This measure was used because, even though people were asked to provide just one sentence, in many cases multiple events/states were included in the responses. So for example, in the cake-seeing scenario, the response "But she knew she was on a diet so decided to wait until she got back to her office, and then ate something less fattening", was classed as having *three* events/states:

(1) Attribute of being on a diet (a state).
(2) Cathy went back to office (event 1).
(3) She ate something less fattening (event 2).

To measure *commonality*, we noted the most common response, i.e. the frequency of occurrence of a given response across a given group. In the cake-seeing scenario the most common event was "Cathy gets the cake" which received 11 counts in the Predictable group and 9 in the Creative group.

For each of these measures, two raters independently scored the materials. The inter-rater reliability was uniformly high on each; for example, in the diversity measure a random sample of ratings showed 94.99% inter-rater reliability in categorising the different responses.

Table 1: Sample responses from two scenarios

"Katie searched everywhere for her little kitten. Then she heard a miaow from the bin"

Predictable	*N*	*Creative*	*N*
Katie opened bin	11	Katie opened bin	4
An explanation that kitten was in bin	3	An explanation that kitten was in bin	4
Kitten walked out of bin	1	Katie found whole cluster of kittens	1
		Katie found wrong kitten	1
		Kitten was being carried away	1
		Katie couldn't understand how kitten was in bin	1
		Bin collector came	1
		Katie was relieved	1
		Katie was disappointed	1

"The cup of coffee was balanced on the arm of the chair. Suddenly Richard sneezed"

Predictable	*N*	*Creative*	*N*
Cup of coffee fell	13	Cup of coffee fell	7
Richard saved cup from falling	1	Richard saved cup from falling	1
Richard's snot went into coffee	1	Richard's snot went into coffee	3
		Chair propelled backwards	1
		Friend got shock and dropped her cup of hot chocolate	1
		Spaceship flew out of nose	1
		A gust of wind went through the window	1

Results & Discussion

To summarise, analysis of the results showed that, though the Creative group produced more creative, diverse and richer responses than the Predictable group, they also could not avoid the commonly occurring events that were invited by the scenario. These results demonstrate that in generating novel outcomes, people are restricted by their background knowledge. Table 1 illustrates samples of the responses made by participants in two scenarios in the experiment.

Creativity of Responses All of the responses were rated blind-to-condition by two judges independently on a 7-point scale. The judges' ratings were reliably correlated (Pearson's $r = 0.748$, $p < 0.01$). A 2 x 2 ANOVA on these ratings for the Group (between-participants) and Scenario (within-participants) variables revealed a main effect of Group, Materials and a reliable interaction, $F(1,478) = 4.65$, $p < 0.05$, $MSe = 6.01$. As expected, responses from the Creative group ($M = 3.523$) were rated as being more creative than those of the Predictable group ($M = 2.245$), $F(1,478) = 188.82$, $p < 0.01$, $MSe = 397.84$. It was also found that the Unfolding materials were rated as more creative ($M = 2.96$) than the Surprise materials ($M = 2.80$), $F(1, 478) = 4.897$, $p < 0.05$, $MSe = 6.338$. This finding suggests that such unfolding events promote greater creative products. This was an unexpected result. It may indicate that it is harder to break the inevitability of the outcome to a surprise scenario because its outcome is much more highly determined. However, we should exercise some care in making sweeping conclusions about this difference, as it is not reflected in any of the other measures. The reliable interaction showed that the most creative condition was the Creative-Unfolding one ($M = 3.68$), followed by the Creative-Surprise ($M = 3.36$), Predictable-Unfolding ($M = 2.24$) and Predictable-Surprise ($M = 2.23$) conditions respectively.

Diversity of Responses Overall, one would expect a greater diversity of responses in the Creative group versus the Predictable group. This is exactly what we found (see Figure 1). A 2 x 2 ANOVA on the diversity scores revealed a reliable main effect of Group, with the Creative group generating more classes of responses ($M = 6.37$) than the Predictable group ($M = 3.5$), $F(1, 14) = 34.6$, $p < 0.01$, $MSe = 66.13$. There was no reliable effect of Scenario and no reliable interaction. An indication of the greater diversity in Creative responses can be seen in Table 1 where both scenarios show more diversity in the Creative condition.

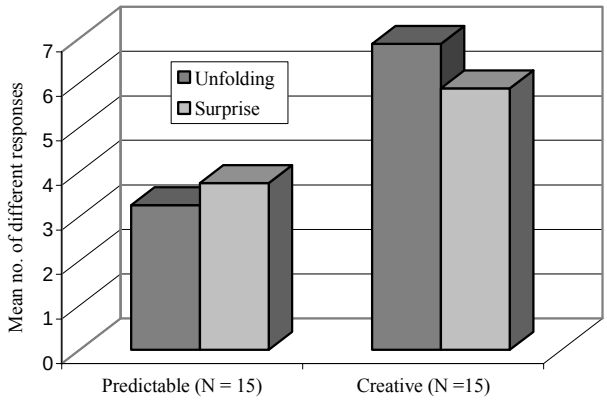


Figure 1: Diversity - the mean number of different responses generated for each condition

Richness of Responses Another index of creativity is the elaborateness or richness of the endings generated. In general, one would expect a greater richness in the responses made by the Creative group than by the Predictable group. Like Haught & Johnson-Laird (2003) we tapped this dimension by examining the average sentence length of people's endings. In addition to this we calculated the mean number of different events in each response.

A 2 x 2 ANOVA again revealed a main effect of Group but no other reliable effects. The Creative group was more likely to provide longer responses ($M = 12.5$ words) than the Predictable group ($M = 9.24$ words), $F(1, 233) = 35.071$, $p < 0.01$, $MSe = 1220.29$. An example of a Predictable response for the first scenario in Table 1 was "She reached in and pulled the kitten out" (word count = 8), a creative response for the same scenario was "She pulled a white kitten from the bin, her kitten was black so she put the white kitten back and carried on looking" (word count = 23).

A 2 x 2 ANOVA on the mean number of different events in the ending showed a comparable pattern; a reliable main effect of Group, but no other effects. The Creative group was more likely to include additional events per ending (M = 1.8) than the Predictable group (M = 1.46), $F(1,233)$ = 19.583, $p < 0.01$, MSe =13.86. Using the example above, the predictable response was classed as having one event and the creative was classed as having three events.

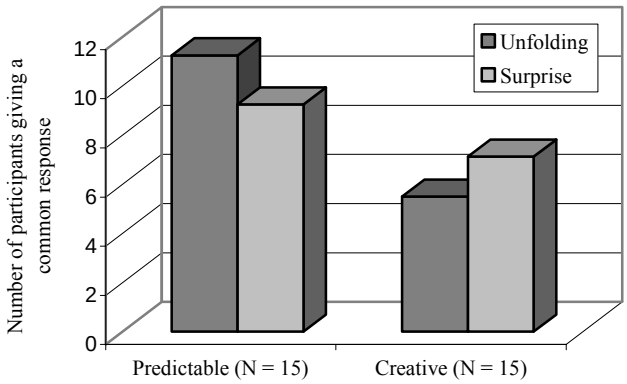


Figure 2: Commonality – count of most common response for each scenario across conditions

Commonality of Responses The above measures show the generativity of the Creative group at work relative to the Predictable one, but they do not reveal the constraints we expected from background knowledge based on Ward's (1994) proposals. If this constraint is in evidence we should see that, in spite of the clear differences in the creativity of responses, there should be certain commonalities between the Creative and Predictable groups too. Specifically, we should see many of the Creative group using the same, inevitable events as some part of their endings. As Figure 2 illustrates, this is exactly what we found. In an analysis of the most commonly produced response, we observed that while those in the Predictable condition (M = 10.25) were more likely to produce the common event, those in the Creative condition (M = 6.313) also produced this same event to a high degree, $F(1, 14) = 31.042$, $p < 0.01$, MSe = 124.03). Again an example of this can be seen in the second scenario in Table 1: the most common response for all participants is that the "cup of coffee fell", which received high counts in both conditions.

General Discussion

The aim of this study was to explore how creative instructions would influence an individual's completion of a common event sequence. The Creative Cognition approach argues that background knowledge can play a constraining role in the creative process, a proposal that is confirmed by the present results. More specifically this experiment has shown that the expectations we have about certain events in the world have a profound influence on our thought processes. We found that although responses of the Creative group were more creative, rich and diverse than those in the Predictable group, certain elements of the endings provided by both groups overlapped considerably. Thus, it appears that whilst creativity in essence involves some degree of variability and unpredictability, it is firmly rooted in our background knowledge of events. In the remainder of this section, we discuss the relationship of these results to the literature on comprehension, and how they might be modeled computationally.

Consistency With Theories of Comprehension

Graesser Singer & Trabasso (1994) stress that knowledge of goals, actions and events are deeply embedded in our perceptual and social experience. As we interact with the environment, we have a strong tendency to interpret event sequences as causal seqeunces, and a similar process occurs in comprehension by means of *inferences* (Kintsch, 1998; Zwaan, 1999). For example, Duffy (1986) observed that when reading, we continually form expectations about future events, so as to develop a causal chain of narrative. The 'Situation Model' of comprehension holds that we construct a detailed mental representation of people, objects, locations, events and actions described in a text (e.g. Zwaan, 1999). Consequently, when reading the scenarios of the present experiment, it was difficult for participants to avoid making rapid, almost automatic inferences about the mental states of the characters and/or the events that would typically occur. In the scenario where Cathy sees the cake for example, it can easily be inferred that Cathy is hungry and that she would like to eat the cake. It could be hypothesised that in this task, the participants naturally link the two sentences together, and in order to provide a coherent ending, they must fit their response with the depicted events so that it is easily understandable and 'makes sense' when read. It is this fundamental knowledge constraint that often overrides instructions to be creative.

Possible Computational Models

Connell & Keane (2002, 2003, in press) have developed a computational model of plausibility judgements for event sequences that is consistent with the above general theory of comprehension. At present, this model takes some event description and finds alternative possible inferential paths to link the events described, this elaborated representation then being scored to assess the plausibility of the description. As such, this Plausibility Analysis Model (PAM) is one possible candidate model that could be extended to deal with the present findings. Such an extension would have to rely on two significant changes: (i) the generation of further possible events to a given sequence based on background knowledge and then (ii) the application of some set of selection heuristics to rank order these possible outcomes for their novelty. As a first pass, such rules could just favour less likely outcomes; that is, outcomes that could possibly occur but that are not strongly supported by prior experience. Obviously such an extension would invite new predictions about the relationship between plausibility and creativity too.

Concluding Comments

The present paper reports a novel study of people's ability to generate creative endings to sentences describing commonplace event sequences. This work connects several areas that have previously been quite separate; namely creativity, sentence comprehension and plausibility. The convergence of these three areas presents a real opportunity for understanding this type of creativity in a new and computationally well-specified way. In short, we should be able to characterise the mud of everyday knowledge, exactly how it glues us to the ground and, yet, the exact nature of the way in which we reach for the stars.

Acknowledgments

This research was funded by a grant to the first author from the Irish Research Council for Science, Engineering and Technology and supported by Science Foundation Ireland under Grant No.03/IN.3/I361 to the second author. Thanks to Phil Maguire and three anonymous reviewers who provided some helpful comments.

References

Boden, M. A. (1995). Creativity and unpredictability. *Stanford Education and Humanities Review, 4* (2).

Bower, G. H., Black, J. B., & Turner, T. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177-220.

Connell, L., & Keane, M.T. (2002). The roots of plausibility: The role of coherence and distributional knowledge in plausibility judgments. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Connell, L., & Keane, M.T. (2003). PAM: A cognitive model of plausibility. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Connell, L., & Keane, M.T. (in press). What plausibly affects plausibility: Concept coherence and distributional word coherence as factors influencing plausibility judgments. To appear in *Memory and Cognition.*

Duffy, S. A. (1986). Role of expectations in sentence integration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 208-219.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative Cognition: Theory, Research, and Applications*. Cambridge. MA: Bradford-MIT Press.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Haught, C., & Johnson-Laird, P. N. (2003). Creativity and constraints: The creation of novel sentences. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition.* Cambridge: Cambridge University Press.

Marsh, R. L., Landau, J. D., & Hicks, J. L. (1997). Contributions of inadequate source monitoring to unconscious plagiarism during idea generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 886-897.

Marsh, R. L., Ward, T. B., & Landau, J. L. (1999). The inadvertent use of prior knowledge in a generative cognitive task. *Memory & Cognition, 27,* 94-105.

Perkins, D.N. (1981). *The Mind's Best Work.* Cambridge: Harvard University Press.

Rubin, D. C., Stoltzfus, E. R., & Wall, K L. (1991). The abstraction of form in semantic categories. *Memory and Cognition, 19*, 1-7.

Schank, R. C., & R. P. Abelson. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (Ed.). (1999). *Handbook of Creativity*. Cambridge, MA: Cambridge University Press.

Ward, T. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*, 1-40.

Zwaan, R. (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science, 8,* 15-18.

Appendix: Materials Used in the Study

Unfolding Scenarios

1 Cathy looked at the cake in the shop window. She hadn't had lunch that day.
2 The dog saw the bone in kitchen bin. He wagged his tail in anticipation.
3 Thomas the cat felt bored. He noticed the dangling tablecloth.
4 Katie searched everywhere for her little kitten. Then she heard a miaow from the bin.
5 James wanted to read the paper. He stopped at the shop on his way home from work.
6 Robert hated his old car. He decided to call the bank.
7 Matthew had wanted to quit his job for months. One day he walked into his boss's office.
8 Jim felt very cold. He got some coal and firelighters.

Surprise Scenarios:

1 John and Pat were kicking a football on the street. A speedy car sharply turned the corner.
2 Michael's shopping bags were bursting with groceries. He felt one of the handles begin to break.
3 The cup of coffee was balanced on the arm of the chair. Suddenly, Richard sneezed.
4 The yacht sailed on as the crew slept. A rocky reef lay directly ahead.
5 Paul was crossing a busy road. Unexpectedly his mobile phone rang.
6 Peter and Sally ate lunch in the small restaurant. They didn't realise that the meat wasn't properly defrosted.
7 The little boy played at the edge of the pond. Suddenly, he slipped on some moss.
8 The sheep were grazing in the field. Suddenly, a wolf approached the flock.

Task Complexity and Difficulty in Two Computer-Simulated Problems: Cross-cultural Similarities and Differences

C. Dominik Güss (dguess@unf.edu)
Emma Glencross (eglencross@juno.com)
Ma. Teresa Tuason (ttuason@unf.edu)
Lauren Summerlin (lauren233axo@aol.com)
F. Dan Richard (drichard@unf.edu)
Department of Psychology, 4567 St. Johns Bluff Road, South
Jacksonville, FL 32224 USA

Abstract

Complex problems have often been described along certain dimensions, e.g. complexity, transparency, and dynamics. However, problem descriptions of the researcher and problem-characteristics perceived by the participant might differ. This study investigates subjective task complexity and its relationship to complex problem solving performance. Research questions are: Do problem perceptions differ a) between different complex problems? b) between cultures? and c) between participants' performance? Two hundred eighty three students from the US, Brazil, and India participated in this study. Participants played the two computer simulations, Fire and Coldstore, and filled out a problem-characteristics questionnaire after each simulation. Factor analysis revealed two factors; one labeled "Task Complexity", the other "Task Difficulty". Results indicate a) that Fire was perceived as more complex and more difficult than Coldstore in the Brazilian and US sample. The Indian sample perceived both problems as equally complex and difficult; b) a significant main effect of culture was found in Fire and Coldstore regarding Complexity; c) a significant main effect of performance was found for Task Difficulty in Fire and Coldstore, but not for Task Complexity. Cultural variables that could explain the results, such as uncertainty avoidance and differences in computer experience, are presented. Results are further discussed under a theoretical and applied perspective.

Complex Problem Solving and Culture

The study of complex problem solving has increased in the last decades especially in Europe (Frensch & Funke, 1995). Computer simulations of complex problems have been widely used to study human problem solving behavior (Brehmer & Dörner, 1993). The researchers were motivated to incorporate into their simulations characteristics common to real life situations, e.g. complexity, transparency, and dynamics (Dörner, 1996). A problems' complexity is derived from the inclusion of many interdependent variables. Complex problems are nontransparent in that the problem solver initially does not know or understand the nature of the hidden variables in the problem situation. The situations change dynamically with and without the actions of the problem solver. One might derive that the more complex, the more non-transparent, and the more dynamic a problem objectively is, the more difficult it is. However, the described problem characteristics are not objective descriptions of complex problems, but are dependent on the knowledge and experience of the problem solver. Individuals differ in experiences regarding problem-related knowledge and strategic knowledge. The complexity, transparency, and dynamics of a situation interpreted subjectively might be completely different. Therefore, both the problem's specific characteristics and the experience of the problem solver will influence the subjective interpretation of the problem. This interpretation is a crucial aspect of the problem solving process and thus, one might expect individual differences.

Knowledge and experience of the problem solver is strongly influenced by one's cultural environment and several studies have shown how problem solving differs between cultures (Cole, Gay, Glick, & Sharp, 1971; Güss, 2002; Strohschneider & Güss, 1999). Culture is a broad term that can be defined in many ways (Kroeber & Kluckhon, 1963). Under a psychological perspective it can refer to implicit and explicit shared knowledge that is transmitted from generation to generation (Smith & Bond, 1998). This knowledge is helpful for a specific group to adapt to specific conditions of the environment. Cross cultural differences in problem solving strategies validate that the people's knowledge base is strongly influenced and shaped by their cultural environment. In essence, the more interesting questions is why and how problem solving strategies are influenced by culture.

One aspect of this implicit knowledge are values that direct behavior, one such value is called *uncertainty avoidance*. A problem is by definition an uncertain situation as the problem solver does not know how to reach a goal state. Uncertainty avoidance refers to "the extent to which the members of a culture feel threatened by uncertain or unknown situations" (Hofstede, 2001, p. 161). Our expectation is that values of uncertainty avoidance influence the initial perception of a problem. For example Hofstede (2001) studied uncertainty avoidance in 53 countries. In his study, India and the United States showed weak uncertainty avoidance, whereas Brazil showed high uncertainty avoidance.

This study investigates the following questions: Do participants perceive different problems in different ways, i.e. is one problem regarded as more complex than another problem? Do perceived problem-characteristics differ between cultural groups? If so, can differences in uncertainty avoidance explain differences in problem perception? Might it be that those problems are created and described with a bias from a western point of view?

Method

Participants

Participants in this study were 283 students from three countries. In India (n=96), participants came from the University of Kerala in Thiruvananthapuram, and from Loyola College, Kerala. In Brazil (n=86), participants were from the universities of Gama Filho, Rio de Janeiro and Faculdade Roy Barbosa, Salvador da Bahia. In the United States (n=101), students were from Northern Illinois University and University of North Florida. Participants received either course credit or were paid for their participation. Students were from the schools of arts and sciences, social sciences, and business. None of the participants had taken part in other complex problem experiments prior to this study. Seventy percent of the participants were majors in psychology. One hundred seventy-six participants were female, and 107 participants were male. Their ages ranged between 18 and 38 years. The average age in the US sample was 22.6 years, in the Indian sample 23.8 years, and in the Brazilian sample 22.0. The mean ages in the three cultural samples were not significantly different. Samples were comparable according to course or major and gender. Data were collected in group sessions (2 hours) and individual sessions where Fire and Coldstore simulations were administered (2 hours).

Materials: Fire, Coldstore, and Problem-characteristics-questionnaire

Participants played two computer simulations, called "Fire" (Gerdes, Dörner, & Pfeiffer, 1993) and "Coldstore" (Reichert & Dörner, 1988). Instructions to each simulation were provided and test games were played before the actual simulation started. After each simulation, a questionnaire regarding simulation characteristics was completed. Instructions and questionnaires were translated from English in translation-backtranslation procedures into Brazilian Portuguese with the help of bilingual Brazilians. The material was presented in Brazilian Portuguese in Brazil and in English in the US and in India. Indian participants, as bilinguals, had no difficulties in answering the Likert-scale questions. The questionnaires consist of identical items and are labeled Fire-characteristics-questionnaire (FCQ) and Coldstore-characteristics-questionnaire (CCQ). This questionnaire is a modification of the one originally developed by Schaub (2001). It consists of 24 items which participants rate on a 7-point Likert scale regarding complexity, transparency, and dynamics. Examples of questions related to complexity are "I find the game complex" and "There are many variables in this simulation". Examples of questions related to transparency are "Not everything is visible that you would like to see" and "The game developments were quite surprising". Examples of questions related to dynamics are "The simulation is dynamic with many changes", "Changes occur often without my intervention".

In the Fire simulation, the participant assumes the role of a fire fighting commander and has to try to protect three towns and the forest from approaching fires. In Coldstore, the participant takes the role of a supervisor in a grocery store with a coldstorage unit. The automatic temperature device has broken down and the participant has to manually control the temperature until the cooling trucks arrive. The goal is to keep the temperature stable at an optimal temperature in order to keep products from freezing or spoiling. Each simulation lasts about 12 minutes.

A comparison of the simulation characteristics of Fire and Coldstore shows that Fire is more complex and more dynamic, yet both are similarly non-transparent. On the screen of the Fire simulation, the participant sees the forest, three cities, fire fighting trucks, helicopters, water dikes, and stone area. In Coldstore, the participant sees the control wheel, the actual temperature, and the target temperature on the screen. Fire is not just more complex regarding stimuli on the screen, but also regarding possible actions. Fire offers 4 main (and a few other) command options for 3 helicopters and 9 fire fighting trucks at any time. These commands can be given to individual units or several units at the same time. At any given time, a person has the choice of a minimum of 4 x (12!) = 312 alternatives. (In fact, the participant has still some additional options). On the other hand, the participant can also just wait and watch what happens. Coldstore offers the participant only one option, a control wheel which the participant clicks with the cursor to regulate the temperature.

Although both simulations can be described as highly dynamic and changing in a non-linear way, Fire is significantly more dynamic than Coldstore. Fires break out at certain times in the simulation, in different locations. Wind strength, wind direction, and interventions of the participant all influence how the fire spreads. Subjectively, participants experience time pressure. In Coldstore, the development of the temperature changes but with delayed effects.

Both simulations can be described as moderately non-transparent. In Fire, participants see the fire, and see wind direction and strength. However, many participants have problems operating the commands. Even if participants have played a test game and have read the instructions, many don't understand the impact of the consequences and long-term effects of some of the commands. In Coldstore, transparency is related to delayed feedback. Participants have the impression, and are often surprised, that the temperature does not immediately react to changes on the control wheel. For many participants, the reasons behind the temperature fluctuations are hard to understand. Analysis of participants' questionnaire responses will show if Fire is indeed described as more complex, more dynamic, and similarly non-transparent as Coldstore. Data analysis will

reveal a universal or a culture-relative perception of problem characteristics.

Initially, the reliability of the two game characteristics-questionnaires (FCQ, CCQ) was studied. Exploratory factor analysis was conducted to investigate the underlying theoretical structure of the instruments, and measurement equivalence was studied with item analysis. Data regarding participants' subjective evaluation of simulation characteristics was then compared between the three cultural groups and related to performance in the two simulations.

Results

In cross-cultural comparisons, data must be analyzed for equivalence before any meaningful cross-cultural comparisons can be made. Cronbach's alpha was used to assess the instrument's reliability. The Cronbach's alpha coefficients of 13 complexity items were .632, .552, and .802 for the Indian, Brazilian, and US samples respectively. The internal consistency for complexity items is adequate for the US sample, but not for the Indian and Brazilian samples. The Cronbach's alpha coefficients of the 7 transparency items and 4 dynamics items in all three cultural samples were relatively low, i.e. between .242 and .551. This was the case for the Fire- and the Coldstore-characteristics-questionnaire (FCQ and CCQ). The Cronbach's alpha coefficients for all 24 items in Fire were .762, .691, and .827, for India, Brazil, and the US, and indicate a one-dimensional construct. Further investigation of the measurement's structure using exploratory factor analysis was conducted for the overall sample and for each cultural group. Results indicated two main factors (overall and in each cultural sample). The first factor "Task Complexity" refers to items indicating complexity, transparency, and dynamics. The second factor "Task Difficulty" refers to the subjective impression of the participant on the situation. To compare the factor structure and loadings between the different cultural groups, the coefficient of congruence (MacCallum, Widaman, Zhang, & Hong, 1999) was calculated. Results indicate a similar factor structure in all the cultural samples in both simulations. In a next step, item analysis was conducted following van de Vijver and Leung (1997) approach. Items showing cultural bias were excluded from further analysis.

Items of Factor 1 and Factor 2 are significantly correlated in both simulations (Fire-complexity and Fire-difficulty, r = .189**; Coldstore-complexity and Coldstore-difficulty, r = .178**. Moreover, items of Factor 1 in both simulations are significantly correlated (Fire-complexity and Coldstore-complexity, r = .361**) and items of Factor 2 in both simulations are significantly correlated (Fire-difficulty and Coldstore-difficulty, r = .130*). Factor 1 in one simulation and Factor 2 in the other simulation are not significantly correlated. Cronbach's alpha coefficients for the new scales "Task Complexity" and "Task Difficulty" were calculated for the cultural subgroups and for the overall sample. The Cronbach's alpha coefficients ranged between .606 and .833 in the two Fire item subscales and between .648 and .795 in the two Coldstore item subscales. These reliability measures can be considered satisfactory considering the small number of items in each scale.

Task Complexity and Task Difficulty in the US, Brazilian, and Indian Samples

In a next step, Fire and Coldstore were compared regarding Complexity and Difficulty. Results showed that Fire was perceived as more complex and more difficult than Coldstore (see Figure 1). Lower mean scores indicate high complexity and high difficulty. Dividing the overall score of 18 by 7 (7 items), gives an average of 2.6 on a scale from 1 to 7. This means that overall, the Fire simulation was considered quite complex and quite difficult [Complexity: M_F = 17.59, SD_F = 6.43; M_C = 22.97, SD_C = 8.80; $F(1, 551)$ = -67.36, $p < .001$, $\eta^2 = .11$] and [Difficulty: M_F = 18.57, SD_F = 5.98; M_C = 21.74, SD_C = 6.96; $F(1, 557) = 33.25$, $p < .001$, $\eta^2 = .056$].

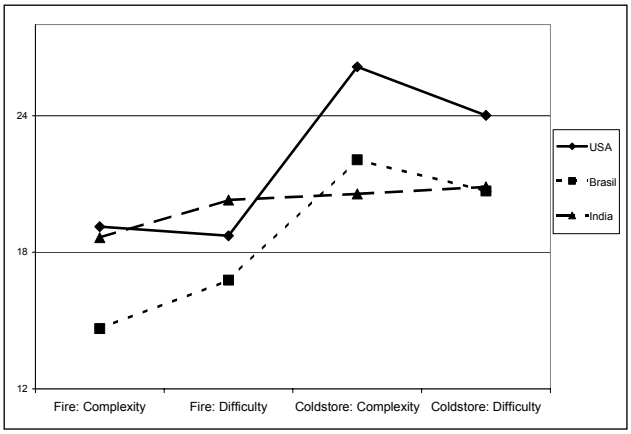


Figure 1: Mean values of US, Brazilian, and Indian participants in Complexity and Difficulty in Fire and Coldstore (Complexity and Difficulty scores are inverted).

In a next step, Complexity and Difficulty in Fire and Coldstore were compared among the three cultural groups using ANOVAs. Scheffe post-hoc tests were calculated to compare differences in the mean values between the three cultures. Comparisons of mean values between cultures in Fire showed that the US and Indian samples had significantly higher average scores regarding Complexity and Difficulty than the Brazilian sample. Brazilian participants perceived the Fire simulation as more complex and difficult (low scores indicate higher complexity). US participants perceived Coldstore as significantly less complex and less difficult. Interestingly, Indian participants' ratings in Coldstore did not differ significantly from their ratings in Fire. These findings show how participants from different cultures view problems quite differently. In the following parts, we will analyze Complexity and Difficulty in relation to task performance.

Fire: Task Complexity and Difficulty in Relation to Actual Task Performance

The performance variable in Fire was the percentage of protected forest at the end of the simulation. Among all the participants, the percentage of protected forest ranged from 41.97% to 97.45% (M = 63.91%, SD = 19.36). The overall scores were distributed in percentile ranks: 25th percentile at 46.61%, 50th percentile at 52.92%, and 75th percentile at

83.01%. Every participant was assigned a score from 1 to 4 according to his or her performance, with higher scores indicating higher performance. Separate two-way ANOVAs were calculated with the two independent variables culture (3 levels) and performance score (4 levels). Dependent variables were Complexity and Difficulty.

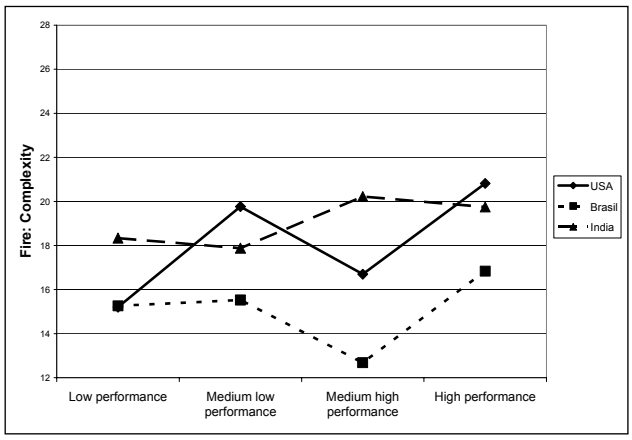


Figure 2: Mean Complexity values of US, Brazilian, and Indian participants according to performance in Fire (Complexity scores are inverted).

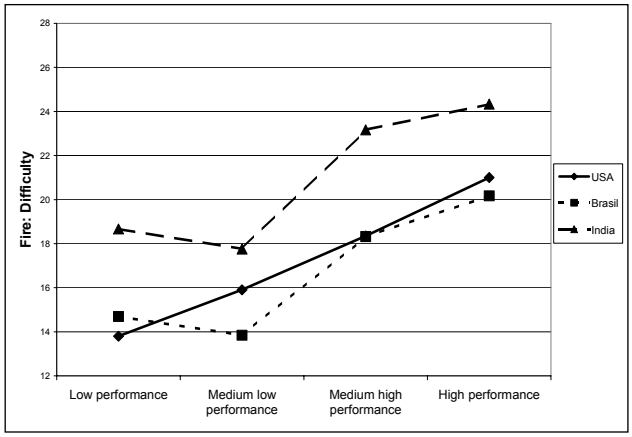


Figure 3: Mean Difficulty values of US, Brazilian, and Indian participants according to performance in Fire (Difficulty scores are inverted).

Regarding the Fire simulation, a significant main effect of culture was found in both Complexity, $F(2, 260) = 8.66$, $p < .001$, $\eta^2 = .065$, and Difficulty, $F(2, 260) = 14.11$, $p < .001$, $\eta^2 = .10$ (see Figures 2 and 3). Post-hoc Scheffe tests showed, that Brazilian participants rated the Fire simulation as significantly more complex and more difficult than US and Indian participants. A significant main effect of performance was found for Difficulty, $F(3, 259) = 17.69$, $p < .001$, $\eta^2 = .175$, but not Complexity. Those who performed better tended to rate the Fire simulation as less difficult. However, regardless of their actual performance on the Fire task, the Fire simulation was viewed with similar complexity. Regarding complexity, interaction effects between performance and culture were not significant.

Coldstore: Task Complexity and Difficulty in Relation to Actual Task Performance

In Coldstore, the sum of the absolute deviations (SAD) from the target temperature was the performance criterion. The minimum of SAD was 134.82, the maximum 1360.85 (M = 632.99, SD = 263.71, N = 288). The overall SAD scores were distributed in percentile ranks: 25th percentile at 434.80, 50th percentile at 667.41, and 75th percentile at 826.29. The SAD was recoded into values of 1 to 4 according to performance, with higher scores indicating the least deviations, and thereby better performance.

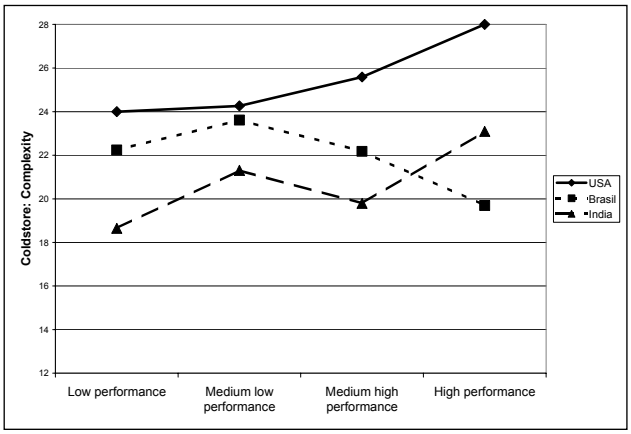


Figure 4: Mean Complexity values of US, Brazilian, and Indian participants according to performance in Coldstore (Complexity scores are inverted).

Separate ANOVAs were calculated with the independent variables, culture (3 levels) and performance score (4 levels). Dependent variables were Complexity and Difficulty. Again, a high score stands for low complexity and low difficulty (inverted, e.g. "I find the game complex" 1-Yes, 7 -No).

In the Coldstore simulation, a significant main effect of culture was found for Complexity, $F(2, 258) = 6.64$, $p = .002$, $\eta^2 = .051$, but not for Difficulty (see Figures 4 and 5). US participants found the Coldstore simulation less complex compared to Brazilian and Indian participants. A significant main effect of performance was only found for Difficulty, $F(3, 257) = 3.391$, $p = .019$, $\eta^2 = .039$ but not for Complexity. As post hoc Scheffe tests reveal, the American participants rated the simulation's complexity as well as the simulation's difficulty significantly lower than the Brazilian and Indian participants. No significant differences between the Brazilian and Indian samples were found.

As evident in Figure 5, the Difficulty scores in Coldstore are relatively similar among the low, medium low, and medium high performance groups with an unusual pattern in the American sample. Only high performing participants viewed the simulation as less difficult. This means that participants who performed well in Coldstore (group 4), rated the simulation as less difficult than those who did not

perform as well. We found this trend in the American and Brazilian samples, but not in the Indian sample. The interaction effect between performance and culture was significant for Difficulty, $F(6, 254) = 2.585$, $p = .019$, $\eta^2 = .059$.

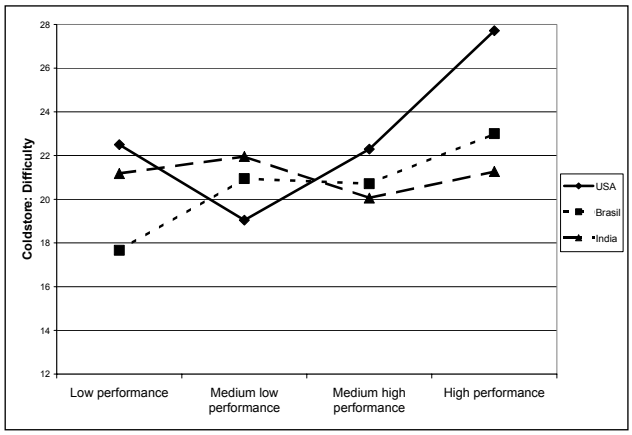


Figure 5: Mean Difficulty values of US, Brazilian, and Indian participants according to performance in Coldstore (Difficulty scores are inverted).

Discussion

In this study, we asked whether problem-characteristics-perceptions differ between complex problems, between cultures, and between performance levels. These questions were studied by administering the simulations Fire and Coldstore and the problem-characteristics-questionnaires to US, Brazilian, and Indian participants. The questionnaires assessed Task Complexity and Task Difficulty. Overall and as expected, Fire was perceived as more complex and more difficult than Coldstore.

In both simulations, no differences were found regarding participants' perception of Task Complexity in relation to performance. Regardless of whether a participant performed in the low, medium or high level, the perception of the simulations' complexity was relatively similar. However, Task Difficulty in both simulations was dependent on performance levels. Participants who performed better regarded the simulations as less difficult compared to those who performed less well. The participant's perception of the task, whether difficult or easy, is related to their actual performance.

This study showed that task complexity and difficulty assessment is an essential step if one wants to compare performance in specific problems. If these simulations are administered in an applied setting or in training programs, it is important to know how the characteristics of these simulations are perceived.

Data analysis also revealed interesting cross-cultural differences. Brazilian participants, compared to Indian and US participants, found the Fire simulation more complex. US Americans, compared to Brazilian and Indian participants, found Coldstore less complex and less difficult. Brazilian and US participants found Fire more complex and difficult compared to Coldstore. However, Indian participants found Fire and Coldstore equally complex and difficult.

There are several possible explanations for these cross-cultural differences. One most plausible reason for the differences among the three cultures is uncertainty avoidance. We expect that low scores in uncertainty avoidance will result in low ratings of complexity and difficulty. In our study, we assessed uncertainty avoidance with the same three questions Hofstede used, but applied them to the school context instead of the work context. In Hofstede's study (2001), India and the United States showed weak uncertainty avoidance scores, whereas Brazil showed high uncertainty avoidance. Surprisingly in our samples, India showed the strongest uncertainty avoidance, Brazil the least uncertainty avoidance, and the US scores were between the Indian and Brazilian ones. The differences between the countries were statistically significant. The different results of our study and Hofstede's study might be related to the samples. Whereas Hofstede's samples consisted of IBM managers, our samples consisted of students. Our participants were also significantly younger and Hofstede has shown significant correlations between age and uncertainty avoidance. A final reason for the different results might be related to changes in cultures. Most of Hofstede's data were collected between 1967 and 1972, i.e. more than 30 years ago and having undergone significant political, economic, and societal development. India, for example, underwent many economic and political changes, especially since the opening of its borders in the 1990s to the world market.

Data analysis revealed that differences in uncertainty avoidance cannot explain cultural differences in problem perception. Brazilian participants, for example, had the lowest uncertainty avoidance scores but the highest Complexity and Difficulty scores in Fire.

A seemingly obvious influence on Complexity and Difficulty scores may be attributed to differences in computer experiences, familiarity with such computer simulations, and motivation to play and succeed in the simulations. However, current results show that although these variables were assessed, none of them can explain the cross-cultural differences in Complexity and Difficulty. The correlations between these variables and task Complexity and Difficulty were not statistically significant.

Why do Brazilian participants compared to US and Indian participants find Fire more complex and difficult? Brazilians perception of Fire as more complex and difficult might be related to the Brazilian time-orientation (Milosevic, 2002). Cross-cultural studies show lower punctuality in Brazil compared to the US (Levine, West, & Reis, 1980) and a more impulsive present-orientation in Brazil (Strohschneider & Güss, 1998). Dealing with a highly dynamic situation like Fire puts the participant under time pressure. Thus, Brazilians with a culture of less strict time orientation may regard the situation as complex and difficult, since their focus is mostly on the immediate, current situation and not on actions in a course of time.

Another question related to the results is why Indian participants view Complexity and Difficulty of Fire and Coldstore similar and Brazilian and US participants view

Fire more complex and difficult than Coldstore. Some studies describe on a general level Western thought as analytic and Eastern thought as holistic (Nisbett, Peng, Choi, & Norenzayan, 2001). A more detailed look at this dichotomy in empirical studies about thinking patterns might reveal more detailed intra- and cross-cultural variations on this general theme. It might be that Brazilian and US participants pay more attention to the details and characteristics of the problems. In a current study (Glencross & Güss, 2004) we find that Indian participants inquire less about problem situations and show more optimism regarding successful planning than US participants. Indians seem to accept the situations as they are, without asking many questions. We find this acceptance of the problem also in other cross-cultural studies between India and Germany (Güss, 2000; Strohschneider & Güss, 1999). These findings could explain why Fire and Coldstore are perceived similarly by the Indian participants.

To summarize, this study stresses the importance of cross-cultural research in the field of cognition. This study showed how perception of task complexity and difficulty can differ between participants of different cultures. For further research, these problem-characteristics-perceptions could be related to strategies that participants use to deal with the complexity and difficulty of the task. It is likely that individuals use different strategies to deal with different perceptions of task complexity and difficulty. The perception of problem characteristics is often the start of the problem solving process, and one can expect and be amazed by the interesting and relevant cross-cultural differences during the next problem solving stages.

Acknowledgements

This research would not have been possible without the support of friends and colleagues in Germany, the US, Brazil, and India. We would like to thank especially Prof. Dietrich Dörner in Germany for providing the two simulations; Prof. Charles Miller in Illinois; Prof. Cristina Ferreira, Prof. Cilio Ziviani, Prof. Nadia, and Dr. Miguel Cal in Brazil; Prof. Krishna Prasaad Sreedhar, Dr. S. Raju, Dr. Ajay Kesavan, and Mr. Ibrahim Syed in India, and the many students who participated.

This research would also not be possible without funding. This study is based on work supported by the National Science Foundation under Grant 0218203 from 2002 to 2005 with the title "Cultural influences on dynamic decision making".

References

Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior, 9,* 171-184.

Cole, M., Gay, J., Glick, J., & Sharp, D.W. (1971). *The cultural context of learning and thinking.* New York: Basic Books.

Dörner, D. (1996): *The logic of failure.* New York: Holt.

Frensch, P., & Funke, J. (Eds.) (1995). *Complex problem solving: The European perspective*. Hillsdale, NJ: Erlbaum.

Gerdes, J., Dörner, D. & Pfeiffer E. (1993). *Interaktive Computersimulation "Winfire"* [The interactive computer simulation "Winfire"]. Otto-Friedrich-Universität Bamberg: Lehrstuhl Psychologie II.

Glencross, E., & Güss, D. (2004). *Values, problem solving, and planning strategies in India and the United States*. Manuscript under preparation.

Güss, D. (2000). *Planen und Kultur*? [Planning and culture?] Lengerich, Germany: Pabst.

Güss, D. (2002). Decision making in individualistic and collectivist cultures. In W. J. Lonner, D. L. Dinnel, S. A. Hayes, & D. N. Sattler (Eds.), *OnLine Readings in Psychology and Culture, Western Washington University, Department of Psychology, Center for Cross-Cultural Research*. Web site: http://www.wwu.edu/~culture

Hofstede, G. (2001). *Culture's consequences* (2nd ed.). Thousand Oaks: Sage.

Kroeber, A. A. & Kluckhohn, C. (1963): *Culture. A critical review of concepts and definitions.* New York: Vintage Books.

Levine, R. V., West, L. J., & Reis, H. T. (1980). Perceptions of time and punctuality in the United States and Brazil. *Journal of Personality and Social Psychology, 38*(4), 541-550.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 49*, 615-618.

Milosevic, D. Z. (2002). Selecting a culturally responsive project management strategy. *Technovation 22*(8), 493-508.

Nisbett, R. E., Peng, K., Choi, I, & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291-310.

Reichert, U., & Dörner, D. (1988). Heurismen beim Umgang mit einem „einfachen" dynamischen System [Heurisms in handling a „simple" dynamic system]. *Sprache & Kognition, 7*(1), 12-24

Schaub, H. (2001). *Persönlichkeit und Problemlösen* [Personality and problem solving]. Hernsbach. Beltz.

Smith, P. B., & Bond, M. H. (1998). *Social Psychology across cultures* (2nd ed.). London: Prentice Hall..

Strohschneider, S., & Güss, D. (1998). Planning and problem solving: Differences between Brazilian and German students. *Journal of Cross-Cultural Psychology, 29*(6), 695-716.

Strohschneider, S. & Güss, D. (1999). The fate of the Moros: A cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology, 34*(4), 235-252.

van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology. Volume 1: Theory and method* (pp. 257-300). Needham Heights, MA: Allyn & Bacon.

Spatial Orientation Using Map Displays: A Model of the Influence of Target Location

Glenn Gunzelmann (glenn.gunzelmann@mesa.afmc.af.mil)
National Research Council Research Associate
Air Force Research Laboratory
6030 South Kent St.
Mesa, AZ 85212-6561

John R. Anderson (ja+@cmu.edu)
Department of Psychology, Baker Hall 342-C
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This paper presents a model of human spatial orientation, using a task that involves locating targets on a map of the space. The model uses a hierarchical solution process that was reported by many of the participants in an empirical study. It encodes the location of the target in the visual scene by identifying a cluster of objects that contains the target and then encoding the position of the target within that cluster. By applying this description of the target's location to the representation on the map of the space, the model is able to correctly identify the target. Using this general strategy, it reproduces all of the major trends in the empirical data.

Introduction

The relative ease with which people are able to navigate through familiar and unfamiliar environments is a human ability that is not well understood. This process requires the integration of multiple sources of information, since immediate visual perception rarely provides a complete representation of a space. To make informed decisions, generally additional information is necessary. When the space is familiar, this information may be available in memory (e.g., a cognitive map). In other cases, however, people often use external maps of a space to facilitate their decision-making.

When external maps are used in conjunction with visual perception to make spatial judgments, one source of difficulty is the difference in how spatial information is represented in the two views of the space. In visual perception, spatial information is available in egocentric terms (e.g., Klatzky, 1998). That is, the locations of objects in the space are encoded in terms of their distance from the viewer and their bearing relative to the viewer. So, the viewer serves as the origin and the direction the viewer is facing defines the orientation. In contrast, external maps identify the orientation and origin within an allocentric frame of reference. These representations are commonly oriented according to cardinal directions, with north at the top.

When the frames of reference in two representations of a space are different, they must be brought into correspondence before they can be used together to facilitate decision-making (Levine, Jankovic, and Palij, 1982). This process requires the ability to identify a common point in both views of the space, along with another piece of information (a second point or a reference direction) to align the orientations. Once this is done, information can be shared between the two views to provide more complete information about the space. Orientation tasks require individuals to establish correspondence between two views of a space. Often, participants are shown a target in one view of a space and are asked to locate it in the other view. Research has shown that the difficulty of this kind of task depends on a number of factors, including the location of the target object and the difference in the orientations (misalignment) between the two views of the space (e.g., Easton and Sholl, 1995; Hintzman, O'Dell, and Arndt, 1981; Rieser, 1989).

The cognitive model that is presented here illustrates a perspective for understanding how the results obtained in studies of orientation tasks arise. The model was developed in the context of the ACT-R architecture. The remainder of this paper presents a brief description of the empirical work on which the model is based, followed by a description of the model and its performance.

Experiment

Participants were shown 2 views of a space containing 10 objects. On the left was a visual scene showing the 10 objects, one of which was highlighted to identify it as the target. On the right side was a map of the space, indicating the locations of the 10 objects as well as the viewer's position. Participants were asked to click on the object on the map that corresponded to the target that was indicated in the visual scene. Figure 1 shows a sample trial.

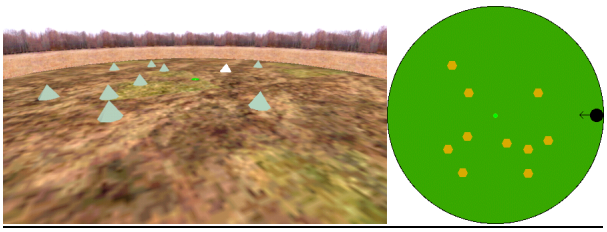


Figure 1: Sample trial for the orientation task.

Method

The spaces were created using the Unreal Tournament game engine (2001), which allows users to create their own 3-D worlds. The 10 objects in the space for each trial were placed in clusters, which were centered around one of 8 positions in the space. In each trial there were four clusters, containing one, two, three, and four objects. The positioning of the clusters was such that on some trials, there were two clusters directly in front of the viewer (one nearby and one farther away), a cluster on the left, and a cluster on the right. In the other trials, there were two clusters on each side of the space relative to the viewer, one nearby and the other farther away. The sample trial in Figure 1 shows the latter case. The configurations and locations of the clusters (in terms of the number of objects in each one) relative to the viewer were counterbalanced. This design resulted in spaces where the objects were not arranged in a well-defined pattern, making it less likely that participants would use strategies that have been described for similar tasks in the past (Gunzelmann and Anderson, 2002).

This experiment varied several factors to closely examine how they impact human performance on orientation tasks. First, the target was located in one of the clusters on each trial. As a result, the target could have been positioned in any of eight general locations relative to the viewer on a given trial. In addition, the target was located in a cluster that contained from one to four objects. So, the target was located in the vicinity of zero to three nearby distractors. Finally, this experiment involved a manipulation of the degree of misalignment between the two views of the space. The two perspectives were either aligned, misaligned by 90 degrees (clockwise or counterclockwise), or misaligned by 180 degrees (maximally misaligned).

There were 20 participants in the experiment. Ten participants completed one half of the possible trials in the experiment, while the other ten completed the other half. When they finished the experiment, participants completed a version of the Vandenberg and Kuse Mental Rotations Test, an assessment that measures spatial ability (Vandenberg and Kuse, 1978). Participants were ranked based upon their scores on this task. Using these rankings, participants were matched between the two conditions and their data were combined to create "meta-participants". The data from these meta-subjects were used in the analyses described here, although the general conclusions remain the same if the data from the two conditions are analyzed independently. Finally, the experiment utilized a drop-out procedure, such that if a participant made an error, the trial was repeated later in the experiment. More complete details concerning the methodology are available in Gunzelmann and Anderson (submitted).

Results

Response times and accuracy were recorded for each trial in the experiment. Overall, accuracy was quite high (96%), and the pattern of errors was quite similar to the response time data (r=.83). This suggests that the results were not due to a speed-accuracy trade-off. As a result of the high degree of accuracy, the data presented here consider the response times for correct responses that participants produced as they performed the experiment.

There are a number of important findings in this experiment, which are summarized in Figures 2, 3, and 4 below. First, in terms of misalignment, the data correspond well with previous research (Figures 2 and 4). As the misalignment between the two views increases, response times increase as well, $F_1(3,27)=38.62$, $p<.001$. Next, the local distractors placed around the target also had an impact on performance (Figures 2 and 3). These data show that as more local distractors were present, participants took longer to identify the correct object on the map, $F_1(3,27)=60.67$, $p<.001$. The magnitude of this effect, however, depended on the degree of misalignment between the two views (Figure 2). Specifically, the impact of the number of local distractors increased as misalignment between the two views increased, $F_1(9,81)=8.79$, $p<.001$.

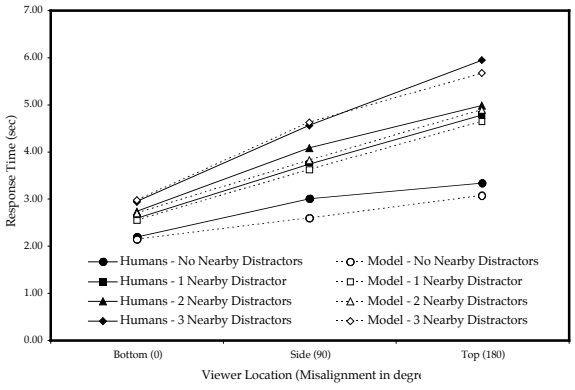


Figure 2: Response times (sec) as a function of misalignment and the number of distractors nearby to the target.

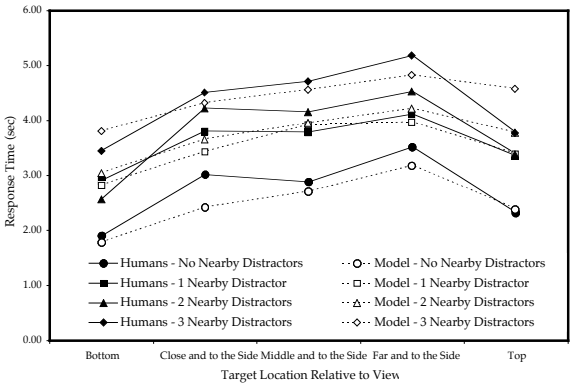


Figure 3: Response times (sec) as a function of the target's location relative to the viewer and the number of distractors nearby to the target.

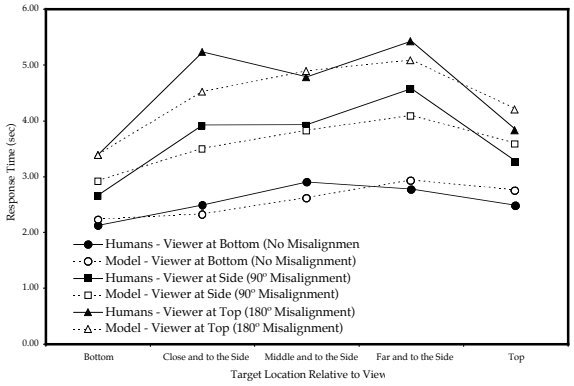


Figure 4: Response times (sec) as a function of the target's location relative to the viewer and the misalignment between the two views.

The impact of the target's location relative to the viewer is shown in Figures 3 and 4. The overall effect was significant, $F_1(7,63)=11.39$, $p<.002$. This result contains two major components. Firstly, response times were fastest when the target was located more directly in front of the viewer (the first and last points on each line in Figures 3 and 4). Secondly, difficulty tended to increase as the target was located farther from the viewer on one side of the visual scene or the other (the remaining points in Figures 3 and 4). Importantly, these trends were produced regardless of the number of distractors that were located near to the target (Figure 3), and this interaction was not significant, $F1(21,189)=1.79$, $p>.15$. In contrast the target's location did have an influence on the impact of misalignment (Figure 4). This result shows that the impact of misalignment was diminished when the target was in line with the viewpoint (the first and last points on each line in Figure 4). This interaction was significant in these data, $F1(21,189)=3.78$, $p<.02$.

In addition to the response time data, retrospective verbal reports from participants provided evidence about how they did the task. In general, participants indicated that they engaged in a two-step process to find the answer. The first step involved identifying a cluster of objects that contained the target so that the cluster could be found on the map. Once the cluster was identified, participants determined which of the objects in the cluster was the correct response.

Discussion

The results of this experiment highlighted several factors that contribute to difficulty in orientation tasks. First, misalignment between the two views of the space impacted difficulty similarly to the results of previous studies (e.g., Gunzelmann and Anderson, 2002; Hintzman, et al., 1981; Rieser, 1989; Shepard and Hurwitz, 1984). Also, the findings show that the location of the target relative to the viewer within a space influences how difficult it will be to locate it on a map. Unlike previous research, this result is demonstrated without using a highly organized configuration of objects in the space. As the target was positioned farther from the viewer, and when the target was less directly in front of the viewer, difficulty increased. These findings suggest that participants were using the viewer's location in the space as a key reference feature to help them determine the location of the target. Response times were faster when the target was in a location that could be encoded more easily with respect to the viewer's position in the space.

This experiment also showed that difficulty increased as more objects were located in the vicinity of the target, a factor that previous research has not addressed. This result suggests that participants were considering only a portion of the space when trying to locate the target, since the total number of objects was the same for all trials. In addition, this effect did not vary as a function of the particular location of the target. This outcome suggests that the location of the cluster does not impact how the target's location within that cluster was encoded.

The hierarchical solution process reported by participants illustrates how they were able to limit their search to a subset of the items in the space and shows why more local distractors would result in longer response times. The presence of more objects near to the target requires more, or more complex, transformations to bring the information in the two views into correspondence, which should take more time (e.g., Bethell-Fox and Shepard, 1988). It appears that one can view the process of solving these tasks as developing a description of the target's location, which then has to be transformed to apply to the map. This description could be verbal, or could involve the creation of a mental image.

ACT-R Model

ACT-R is a general theory of cognition that has been implemented as a running simulation (Anderson and Lebiere, 1998). It operates as a production system with several core assumptions related to its operation. First, there is a division between declarative and procedural knowledge. Declarative memory contains information in the form of chunks, while procedural knowledge is composed of productions, which contain information about transforming one state into another. The latest version of ACT-R is composed of a set of modules for perceptual, motor, and cognitive aspects of human performance. Information is processed independently within these modules, allowing them to operate in parallel. There are buffers associated with each of the modules that essentially represent working memory. The contents of these buffers are what drive the production system. Productions match against the contents of the buffers, and it is only the contents of those buffers that can be directly accessed. It is at the level of production selection and execution that the system operates in a serial manner.

Because ACT-R includes perceptual and motor modules, it is able to interact with experimental software under realistic constraints. Although the perceptual module currently is not sophisticated enough to parse the visual scene shown in Figure 1, it does contribute important timing information to the model's performance. The motor module adds additional constraints to the mouse movements and clicks that the model executes. The parameters that control these aspects of ACT-R's behavior are based on largely on the EPIC theory (Kieras and Meyer, 1997). At a general level, the model was implemented within this architecture to perform the task based on the two-step process described by the participants in their verbal reports. There are, however, a number of details that are important to the model's performance as it goes through this general process. These are described next.

Model Design

The model begins each trial by locating the target in the visual scene. Once this location has been identified, the model finds other objects that are in the vicinity of the target. It counts those objects, and encodes the overall location of the cluster as being in the left, right, or central portion of the visual scene. Then, to encode the location of the target in the cluster, the model revisits the items in the cluster, and encodes the target's position relative to the near-far and left-right axes. So, the model develops a representation of the target's location in the visual scene that would be something like "the leftmost object in the cluster of two on the right of the visual scene." This is the target's position in the sample trial in Figure 1.

With a representation of the target's location in the visual scene, the model shifts its attention to the map of the space, beginning by locating the viewer's position, which is indicated. The next step is to find the correct cluster. If the cluster was encoded as being in the middle of the visual scene, the model searches straight out from the viewer's location to find it. However, when the cluster was positioned to one side or the other, spatial updating is required when the two views are misaligned so that the correct portion of the map is searched. This updating consists of a remapping of "left" and "right" to the corresponding directions on the map relative to the viewer's orientation. For instance, in the sample trial in Figure 1, the right portion of the visual scene corresponds to the top half of the map.

The updating process is a source of difficulty which requires extra time. In addition, when the two views of space are maximally misaligned (the viewer is at the top of the map), the updated values for the map directly conflict with the egocentric values. This adds a second source of difficulty to the updating process, which adds additional time to its execution. Once the values are updated, the model is able to search the appropriate portion of the map for the correct cluster. To perform this search, the model begins near to the viewer's location and searches outward until it finds an object that is in a cluster of the appropriate size.

Once the first step of finding the appropriate cluster is completed, the model needs to determine which of the objects within the cluster on the map is the target. Like the cluster location, the encoding of the target's position is based in the egocentric coordinate system from the visual scene. So, when the two views are misaligned, updates to this information are needed to match the rotated coordinate system of the map. These updates are similar to those described above. Misalignment is one source of difficulty, while direct conflict between the two reference frames is another.

One detail of this process requires some explanation. In the model, the amount of updating done in the second step depends on the number of objects in the cluster. When there are no nearby distractors, this step is skipped. In this case, when the "cluster of one" is found, the model is immediately able to respond by clicking on that object. In cases where there are 2 or 3 objects in the cluster, there is a simple encoding of the target's position within the cluster that requires only one axis. With 2 objects, the target is always on the left or on the right, and is also always the closest or farthest object in the cluster. When the cluster has three objects, the target can also be the one in the middle on each of the axes. In the model, this possibility is represented by having the model update only one of the axes in order to locate the target within the cluster.

When the cluster has 4 objects, the encoding necessarily becomes more complex. This is represented

by having the model update both axes when the cluster has four objects in it. Once again, there is the potential for direct conflict between the two frames of reference in these updates, which adds to the difficulty of this operation. The basic idea is that the complexity of the description needed to encode the target's location increases as the number of objects in the cluster increases. As a result, the difficulty of transforming this description so that it applies appropriately to the map increases as well. This notion is supported by past research, which has demonstrated that more complex figures take longer for individuals to mentally rotate (Bethell-Fox and Shepard, 1988).

The model's performance is modulated by several parameters. First, as noted above ACT-R's perceptual mechanisms are not currently sophisticated enough to process a raw image like the one shown in Figure 1. Thus, as a simplification, the model is presented with a 2-D, egocentrially-oriented representation of the visual scene, which essentially is another map. So, the model implicitly embodies the assumption that participants extract a 2-D representation from the visual scene as they encode the information from it. A constant of .25 seconds was added to the model on each trial to represent the cost of extracting such information from the visual scene.

The second parameter in the model was the retrieval time, which was set to .11 seconds. As the model does the task, it requires some declarative information (mostly related to directional information). Each time a chunk is retrieved from declarative memory, it takes .11 seconds. However, most of the model's performance is driven by the information on the screen, so this parameter does not play a large role in determining the model's predictions.

The only other parameter that was manipulated in the model controls how long it takes to perform the operations needed to update the directions (left, right, up, and down) that it uses to locate the target on the map. The parameter was set so that each of the updates requires .60 seconds. This value applies to each operation that is necessary, and is also applied when direct conflict arises between the allocentric frame of reference and the original egocentric reference frame. This means that if one axis needs to be updated, it will take .60 seconds. If two axes need to be updated, it will take 1.20 seconds. However, if in addition to the update there is direct conflict between the two frames of reference, these updates take 1.20 and 1.80 seconds respectively. These costs apply to the updates needed to locate the cluster and to identify the target within the cluster. When an update is needed to identify the portion of the map where the cluster is located, it involves updating a single axis (left-right). When the cluster has been located and the search begins for the target, the update involves one axis when there are one or two nearby distractors, and two axes when there are three nearby distractors. Each of these updates may or may not involve direct conflict that needs to be resolved in addition to the update. All of the other parameters in the model were given their default ACT-R values.

Model Performance

The model captures all of the major trends in the data. First, the model reproduces the misalignment effect (Figures 2 and 4). As the misalignment between the two views increases, the model takes longer to respond. In the model, this effect comes from the costs of updating the frame of reference to find the cluster on the map and to find the target within the cluster. In addition, the costs associated with the second update depend on the size of the cluster, producing the interaction between misalignment and the number of nearby distractors shown in Figure 2. As the number of nearby distractors increases, the impact of misalignment increases. This illustrates the idea that it is more difficult to update the descriptions of the target locations when those descriptions are more complex. In the model, it is the extra cost associated with the spatial updating process as more nearby distractors are present that produces this interaction. The mechanisms in the model capture the effect of both misalignment and the number of distractors well, with an overall correlation of .992 for the data shown in Figure 2 (RMSD=.187 seconds).

The model makes predictions about the difficulty of the task based upon the target's location in the visual scene as well. These data are shown in Figures 3 and 4, along with the empirical results. The model produces a good qualitative fit to the data, although the particular values are a little off in some instances. The model's predictions arise because it begins its search for the cluster on the map from the viewer's position, moving outward until it locates an object in the cluster. Thus, when the target is farther from the viewer, it takes the model longer to locate an object in the cluster. In addition, the model produces an interaction between target location and misalignment (Figure 4). The effect of misalignment is smaller when the target is directly in front of the viewer (bottom and top target locations) because no spatial updating is necessary to find the cluster. As noted above, the same effect appears in the empirical data, and the model captures the effect well (r=.954, RMSD=.325 seconds for the data in Figure 4).

Finally, there is no interaction in the model between the number of nearby distractors and the location of the target in the visual scene (Figure 3). This corresponds to the empirical results as well (r=.910, RMSD=.351). The result is because of the two-step process, reported by participants, that the model uses to do the task. The target's location within the cluster is encoded without regard to the location of the cluster. So, the impact of the target's location results from the search for the

cluster. Similarly, the number of nearby distractors only impacts the solution process after the cluster has been located, when the correct target must be identified.

Conclusions

Overall, the model produces data that are in line with the performance of the human participants, which lends support to the conclusion that they were using the strategy they reported to do the task. The model produces all of the major trends, in most cases with data that are very close to the data from the human participants. In the model, misalignment impacts both the search for the cluster and the search for the target within the cluster. In contrast, the location of the target only influences the search for the cluster, both in terms of its distance from the viewer and whether or not it is in line with the viewer's position. The interaction of target location with misalignment in the model arises because no spatial updating is needed when the target is directly in front of the viewer. Finally, the number of nearby distractors only impacts the search for the target within the cluster, interacting with misalignment because of the different amounts of spatial updating required based on the number of objects in the cluster. Note that the location of the target does not interact with the number of nearby distractors, suggesting that they affect different aspects of the solution process. The performance of the model supports the conclusion that similar processes are being used by the participants.

In conclusion, this model provides a framework for understanding human performance on spatial tasks. It's most important characteristics relate to the hierarchical encoding of the target's location in the visual scene. This encoding allows the model to limit its search to a portion of the map, ignoring many of the objects in the space. In addition, the model's performance assumes that the two steps in the solution process are independent. As a result spatial updating that is performed for step 1 does not carry over to the execution of step 2. This contributes to the large effect of misalignment on the model's performance. Finally, the model also indicates that perceptual-motor aspects of performance are important factors in this kind of task. The time needed to execute shifts of visual attention contribute to many of the effects described here, especially the impact of the target's location and the impact of the number of local distractors. These issues deserve careful attention in future research.

Acknowledgements

This research was supported by AFOSR grant #F49620-99-1-0086 to Dr. John R. Anderson at Carnegie Mellon University and by NRSA training grant #5-T32-MH19983 from NIMH to Dr. Lynne Reder at Carnegie Mellon University.

References

Anderson, J. R., & Lebiere, C. L. (1998). *The atomic components of thought.* Hillsdale, NJ: Lawrence Erlbaum.

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 12-23.

Easton, R. D., & Sholl, M. J. (1995). Object-array structure, frames of reference, and retrieval of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 483-500.

Gunzelmann, G., & Anderson, J. R. (2002). Strategic differences in the coordination of different views of space. In W. D. Gray and C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 387-392). Mahwah, NJ: Lawrence Erlbaum.

Gunzelmann, G., & Anderson, J. R. (submitted). Location matters: Why target location impacts performance in orientation tasks.

Hintzman, D. L., O'Dell, C. S., & Arndt, D. R. (1981). Orientation in cognitive maps. *Cognitive Psychology, 13*, 149-206.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*, 391-438.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freksa, C. Habel, and K. F. Wender (Eds.). *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge* (pp. 1-17). New York: Springer-Verlag.

Levine, M., Jankovic, I. N., & Palij, M. (1982). Principles of spatial problem solving. *Journal of Experimental Psychology: General, 111*, 157-175.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1157-1165.

Shepard, R. N., & Hurwitz, S. (1984). Upward direction, mental rotation, and discrimination of left and right turns in maps. *Cognition, 18*, 161-193.

Unreal Tournament [Computer software]. (2001). Raleight, NC: Epic Games, Inc.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills, 47*, 599-604.

Seeing the Unobservable – Inferring the Probability and Impact of Hidden Causes

York Hagmayer **(york.hagmayer@bio.uni-goettingen.de)**
Department of Psychology, University of Göttingen, Gosslerstr. 14
37073 Göttingen, Germany

Michael R. Waldmann **(michael.waldmann@bio.uni-goettingen.de)**
Department of Psychology, University of Göttingen, Gosslerstr. 14
37073 Göttingen, Germany

Abstract

The causal impact of an observable cause can only be estimated if assumptions are made about the presence and impact of possible additional unobservable causes. Current theories of causal reasoning make different assumptions about hidden causes. Some views assume that hidden causes are always present, others that they are independent of the observed causes. In two experiments we assessed people's assumptions about the occurrence and statistical relations involving a hidden cause. In the experiments, participants either only observed a cause or actively manipulated it. We assessed participants' assumption online after each learning trial and at the end of the learning phase. The results show an interesting dissociation. Whereas there was a tendency to assume negative correlation in the online judgments, the final judgments tended more in the direction of an independence assumption. It could also be shown that the judgments were generally coherent with the learning data. These results are consistent with normative theories that drop independence as the default assumption.

Introduction

Most events are causally influenced by more than a single cause. Unfortunately, very often these other causes are unknown or cannot be easily observed. Therefore we often have to rely on the observed statistical relationship between cause and effect when assessing causal strength. For example, whenever a new influenza virus invades East Asia, health representatives try to estimate its health risks, well aware of the fact that many other factors determine whether a patient will die or not. The question of how to assess causal strength when there are hidden causes has challenged normative theories of causality and psychological theories of causal reasoning for some time. A number of different accounts have been proposed analyzing how the causal impact of an observed factor can be accurately estimated if certain assumptions are made about potential hidden causes. In this report we will first give a brief overview of how two current theories of causal reasoning handle hidden causes. In the second part of the report we will present two experiments in which we assessed the assumptions of learners about the impact and probability of hidden causes. In the final section we will discuss potential theoretical implications of these findings.

Theoretical Accounts of Hidden Causes

We are going to focus on a simple causal structure consisting of a single observable cause C and one possible hidden cause A both influencing a joint observable effect E. The two observable events C and E are statistically related. Cause C is neither sufficient nor necessary for the effect, $P(e|c)<1$ and $P(e|\sim c)>0$. How can the causal impact of the observed and - if possible - the impact of the hidden cause be assessed in such a situation?

Associative Theories and the Constant-Background Assumption

Associative theories, such as the Rescorla-Wagner theory (Rescorla & Wagner, 1972), would model this task as learning about the association between a cue representing the observed cause and an outcome representing the effect. Along with the cause cue a second background (or context) cue would be part of the model. This background cue is assumed to be always present and to represent all other factors that might also generate the outcome. Thus, the background cue would play the role of representing the hidden cause A in the outlined causal model. According to the Rescorla-Wagner rule, only weights of cues that are present in a certain trial are being updated. Therefore the permanently present background cue will generally compete with the cause cue in cases in which the cause cue is present. If the outcome is also present the associative weights of both cues will be raised, if the outcome is absent, the weights will be lowered. However, in cases in which the cause cue is absent, only the weight of the background cue will be altered. At the asymptote of learning the associative weight of the observed cause will equal the contingency (i.e., $\Delta P=P(e|c)-P(e|\sim c)$) of the cause cue and the outcome. The associative weight of the background cue will correspond to the probability of the outcome in the absence of the cause cue. Thus, the more often the outcome (=effect) occurs on its own, the higher the associative weight of the background cue will be.

Power PC Theory and the Independence Assumption

Cheng's (1997) Power PC analysis of the causal impact of a single cause can be viewed as a special case of a causal Bayes net in which two causes independently influence a joint common effect (Glymour, 2001, Tenenbaum & Griffiths, 2003). The theory states that the occurrence of the effect E is a consequence of the causal powers of the observed cause C and a hidden cause A (p_c and p_a), and of their base rates $P(c)$ and $P(a)$. Formally the probability of the effect equals the sum of the base rates of the two causes multiplied by their causal power minus the intersection of the causes multiplied by both causal powers:

$P(e) = P(c) \cdot p_c + P(a) \cdot p_a - P(c) \cdot P(a) \cdot p_c \cdot p_a$.

Therefore the probability of the effect E given that the observed cause C has occurred is

$P(e|c) = p_c + P(a|c) \cdot p_a - P(a|c) \cdot p_c \cdot p_a$ [1],

and the probability of the effect given that the observed cause is absent is

$P(e|\sim c) = P(a|\sim c) \cdot p_a$ [2].

Equations [1] and [2] offer an account for hidden causes irrespective of whether they are dependent or independent of the observed cause C. However, if they happen to covary, the power of the causes cannot directly be estimated by the observable data because there are four unknown parameters to be estimated by two observable conditional probabilities. Therefore Power PC theory makes the assumption that the observed and the hidden causes are independent, $P(a|c) = P(a|\sim c) = P(a)$. Based on this assumption the causal power of the observable cause can be calculated by

$p_i = (P(e|c) - P(e|\sim c)) / (1-P(e|\sim c))$.

The independence assumption of Power PC theory implies that the probability of the hidden cause stays the same regardless of whether the observed cause has occurred or not. If this assumption holds, Equation [2] defines lower boundaries for the base rate and the causal strength of the hidden cause. The causal power of the hidden cause and its base rate have to be at least as big as the probability of the effect in the absence of the observed cause, $p_a \geq P(e|\sim c)$ and $P(a) \geq P(e|\sim c)$. Equation [2] also defines a coherence criterion for estimates about hidden causes. In order to be compatible with the observed data, the estimates must honor this equation.

It is important to note that even if independence is not assumed, Equations [1] and [2] still hold and have implications for the unobservable cause. The power of the hidden cause and its probability in the absence of the observed cause are still determined by Equation [2]. Therefore estimates for both values should be constrained by $P(e|\sim c)$. Moreover, Equation [1] provides constraints for the admissible probabilities of the hidden cause in the presence of the observable one. However, this constraint is fairly complex and does not provide the same straightforward implications as Equation [2].

Summary

Both theories consider hidden causes. Associative theories assume that a hidden cause (i.e., the constant background) is always present. In contrast, Power PC and other causal Bayes net theories assume that the hidden cause is independent of the observed cause and that its probability is constrained by the data. The probability of the effect in the absence of the cause marks its lower boundary. These theories also permit to model statistical dependence between the observed and the hidden causes.

Both theoretical accounts agree that $P(e|\sim c)$ is to a certain degree indicative of the causal strength of the hidden cause. But whereas associative theories generally regard this probability as a valid indicator, Power PC and other causal Bayes net theories view this conditional probability as a lower boundary of the causal impact of the hidden cause.

Experiments

The following two experiments explore what assumptions participants make about the presence and impact of a hidden cause in a trial-by-trial learning task, and whether these assumptions conform to the predictions of any of the discussed theoretical models. Thus far very little research has been conducted about naïve participants' assumptions about hidden causes. An exception is a study by Luhmann and Ahn (2003). They found that participants judged the impact of a hidden cause to be higher if $P(e|\sim c)$ was 0.5 than if it was zero. The experiments presented in this report will go beyond these findings. In addition to causal strength estimates, we collected assessments of the probability of the hidden cause using different kinds of measures. We also varied the learning conditions.

In both experiments participants learned about the causal relation between an observable cause and a single effect. Additionally participants were told that there was one other possible but unobservable cause of the effect. The statistical relation between the observable cause and the effect was manipulated in the two experiments while either keeping the contingency (Experiment 1) or the causal power (Experiment 2) constant. In Experiment 1 participants could only passively observe the cause, which occurred at its natural base rate, in Experiment 2 participants were allowed to manipulate the cause. A number of dependent variables were collected to assess participants' estimates of the probability of the hidden cause and the impact of both the observed and the unobserved causes. Participants were asked to guess the presence of the hidden cause on each trial during learning, and they were asked to give summary estimates after learning was completed. In one condition ("prediction before effect") participants were first informed about the presence or absence of the cause in each trial, and then they were asked to guess the presence of the hidden cause without receiving feedback about this alternative cause. Finally they were informed whether the effect has occurred at this particular trial or not. Predictions of the hidden cause prior to effect information can only be guesses based on observed frequencies of the effect in past trials. Based on normative theories (e.g., Power PC theory) we expected participants to generate independence between the causes. In the second condition ("prediction after effect") participants received information about the presence of both the cause and the effect and then had to predict the hidden cause. As before no feedback was provided about the hidden cause. In this situation participants had complete information about the cause

and the effect which should allow them to make more informed guesses about the hidden cause, especially if the observed cause is absent: If in this case the effect is present, participants should conclude that the hidden cause is also present. However if the effect is absent, they should have the intuition that the hidden cause is absent. Predictions based on the presence of the observed cause are more difficult. If in this case the effect is absent, participants should infer that the hidden cause is more likely to be absent than present; if the effect is present the hidden cause should also be given a higher probability of being absent. Based on the theories outlined above, we expected that participants in both conditions would generate independence between the causes in their trial-by-trial predictions. A third control condition left out the trial-by-trial predictions. In this condition participants rated the causal strength of the observed and the hidden cause as well as the probability of the hidden cause in the presence and in the absence of the observed cause after the learning phase. Again we expected participants to rate the causes to be independent. We also expected that the strength ratings for the observed cause would be based on causal power, and that the ratings for the hidden cause would be influenced by $P(e|\sim c)$.

Experiment 1

With Experiment 1 we pursued two goals. The first was to investigate whether participants would assume independence between the observable and unobservable cause. The second goal was to find out whether the power estimates for the unobservable cause would be influenced by the probability of the effect in the absence of the observed cause. Participants were given the task to assess the causal relation between a fictitious microbe ("colorophages") and the discoloration of certain flowers. In addition they were told that there was only one other possible cause of the effect, an infection with another fictitious microbe ("mamococcus"), which was currently not observable. Participants were then directed to a stack of index cards providing information about individual flowers. The front side of each index card showed whether the flower was infected by colorophages or not, and the backside informed about whether the flower was discolored or not. Then participants were instructed about the specific learning procedure in their condition. The learning conditions were manipulated as a between-subjects factor. In Condition 1 ("prediction before effect") participants were first shown the front side of the card, then they had to guess whether the flower was also infected by the other microbe, and finally the card was turned around by the experimenter revealing whether the flower was in fact discolored or not. In contrast, in Condition 2 ("prediction after effect") the card was first turned around and then the participant made her guess about the hidden cause. Guesses were recorded without giving feedback. In the third, control condition cards were simply shown and turned around by the experimenter.

As a second factor the statistical relation between the observed microbe and discoloration was manipulated. Three different data sets consisting of 20 cases each were constructed. Table 1 summarizes the statistical properties of the three data sets. As the table shows, the contingency ΔP was constant across the data sets, whereas both $P(e|\sim c)$ and causal power were rising. All three data sets were shown to every participant in a within-subjects design. Different data sets were introduced as data from different species of flowers. It was pointed out to participants that the effectiveness of the microbes might vary depending on the species. The order of the presented data sets was counterbalanced.

Table 1: Statistical properties of data sets shown in Experiment 1

	Data Set 1	Data Set 2	Data Set 3
$P(c)$	0.50	0.50	0.50
$P(e\|c)$	0.60	0.80	1.00
$P(e\|\sim c)$	0.10	0.30	0.50
ΔP	0.50	0.50	0.50
Power p_c	0.56	0.71	1.00

After each learning phase participants were asked to rate the causal influence of the observed and the hidden cause on a scale ranging from 0 ("no impact") to 100 ("deterministic impact"). Participants were also asked to estimate how many of ten flowers that were infected with the observed microbe were also infected with the other microbe, and how many of ten flowers that were not infected with the observed microbe were instead infected with the other microbe. No feedback was provided about these assessments.

36 students from the University of Göttingen were randomly assigned to one of the learning conditions. Figure 1 shows the mean ratings of the impact of the observed and the hidden causes.

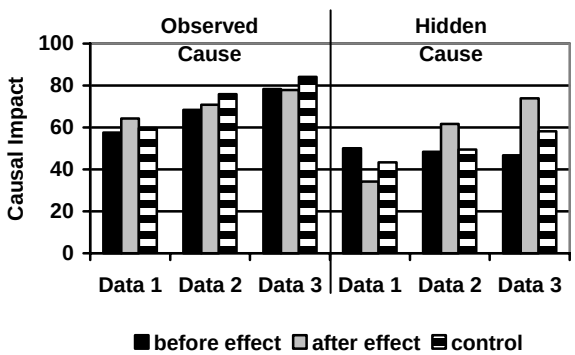


Figure 1: Mean ratings of causal impact for the observed cause (left) and the unobserved cause (right) in Exp. 1.

An analysis of variance revealed a significant increase in impact ratings for the observed cause, $F(2,66)=12.7$, $MSE=296.6$, $p<.01$, supporting the predictions of Power PC theory. The same analysis for the hidden cause resulted also in a significant main effect of the factor data set, $F(2,66)=4.92$, $MSE=408.1$, $p<.05$, which indicates that with increasing $P(e|\sim c)$ participants tended to assume a stronger impact of the hidden cause. This result is in accordance with the predictions of all theoretical accounts. However, the interaction between data sets and learning condition also

turned out to be significant, $F(4,66)=4.55$, $MSE=408.1$, $p<.05$. The observed increase was strongest in the 'prediction after effect' condition followed by the control condition. This interaction might be due to the learning procedure. In the 'prediction after effect' condition participants were sensitized to the possible presence and impact of the hidden cause more than in the other two conditions. Being informed about the occurrence of the effect in the absence of the observable cause is a strong cue pointing to the presence of the hidden cause.

Table 2 shows the results concerning participants' assumptions about the dependence between the causes. The online predictions of the hidden cause in the presence and absence of the target cause were transformed into conditional frequencies, and combined into subjective contingencies, $\Delta P = P(a|i) - P(a|\sim i)$. On the left side of Table 2 the generated contingencies underlying online predictions are listed, the right hand side shows the corresponding contingencies based on the final probability ratings.

Table 2: Mean estimates of dependence between observed and unobserved cause. Numbers indicate contingencies (possible range:-100 to +100).

	Generated Dependence			Estimated Dependence		
	Data Set 1	Data Set 2	Data Set 3	Data Set 1	Data Set 2	Data Set 3
Before Eff.	33.3	8.3	-21.7	30.0	5.8	-3.3
After Eff.	17.0	17.5	-28.8	0.8	0.0	-22.6
Control	-	-	-	14.4	5.8	-4.2

An analysis of variance of the generated contingencies yielded a significant trend from positive to negative assessments which proved independent of learning condition, $F(2,44)=22.1$, $MSE=749.8$, $p<.01$. The estimated contingencies showed a similar trend, $F(2,66)=4.2$, $MSE=753.9$, $p<.05$. The mean contingencies in the different data sets varied slightly across learning conditions, $F(2,33)=2.8$, $MSE=1646.4$, $p<.10$. The generated contingencies were significantly above zero if $P(e|\sim c)$ was zero, and significantly below zero if $P(e|\sim c)$ was 0.5. The estimated contingencies showed a similar but only marginally significant pattern. Thus, there was a hint of a dissociation between online and post hoc assessments which will be followed up in Experiment 2.

These results do not conform to the theoretical assumptions of the discussed theories. Participants did not assume that the hidden cause was always present or that the two causes were independent.

A closer analysis of the conditional probabilities revealed that the negative trend was due to an increase in the generated and estimated probability of the hidden cause in the absence of the observed cause, whereas the subjective probability of the hidden cause in the presence of the observed cause remained relatively stable. This pattern is in part consistent with the analysis outlined in the introduction, $P(a|\sim c)$ seems directly constrained by $P(e|\sim c)$. In contrast, the constraint for $P(a|c)$ is more complex, which may be the reason why participants had more difficulties honoring it.

Even if participants' answers did not conform to the independence assumption, their answers still might be coherent with the observed data. Both Power PC theory and Bayesian models can model dependence between observed and hidden causes. Although precise power estimates might be impossible, the data still yields constraints on plausible estimates. The most important constraint is that the product of the causal power (or strength) of the hidden cause and the probability of the hidden cause in the absence of the observed cause must equal the probability of the effect in the absence of the observed cause. To find out whether participants honor this constraint we used their ratings to recalculate the probability of the effect when cause C was absent:

$P(e|\sim c)_{rec}$ = Rating Impact $A \cdot$ Rating $P(a|\sim c)$.

The results are shown in Figure 2. It can be seen that the recalculated probabilities in the 'prediction after effect' condition were surprisingly close to the actually observed probabilities. In contrast, the recalculated probabilities in the other two conditions were inaccurate. Apparently, participants had to be sensitized by the learning procedure to the presence and impact of the hidden cause to be able to derive coherent estimates. Learning that the effect is present in the absence of the target cause apparently provided the necessary information to make educated guesses about the hidden cause. Without this information the guesses showed some systematicity but did not conform very well to the observed data.

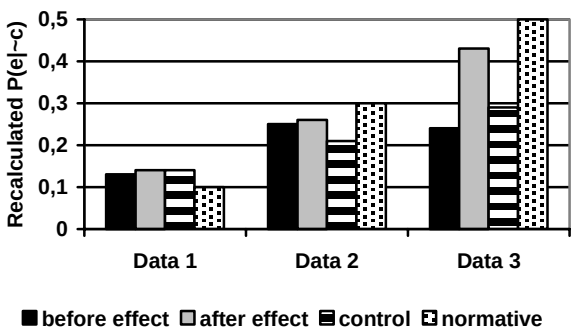


Figure 2: Mean recalculated probabilities of the effect in the absence of the observed cause (Experiment 1).

Experiment 2

In Experiment 1 we used a scenario in which the observable cause could only be passively observed. Therefore a dependence of the observed and unobserved cause was possible and maybe for some participants plausible. In Experiment 2 we allowed participants to arbitrarily manipulate the observable cause. Since these random interventions cannot be based on the presence or absence of the hidden cause, they should make the independence between the alternative causes more salient than in the observation context. Thus, we expected that participants would now assume the causes to be independent in all conditions of the present experiment.

Participants were instructed to imagine being a captain on a pirate ship firing his battery at a fortress. A second ship, which cannot be seen, was also firing at the fortress. Participants had a certain number of shells available and had to decide on each trial whether to fire or not. This procedure ensured that all participants saw the same data despite the fact that they set the cause themselves. The three learning conditions of Experiment 1 were used again. Participants had to guess whether the other ship currently fires either before they were informed about the occurrence of an explosion in the fortress ("prediction before effect"), or they had to predict the other ship's action after they had learned whether the fortress was hit ("prediction after effect"). In a third, control condition no predictions were requested.

Three data sets consisting of 60 cases each were constructed. Table 3 shows the statistical properties of the data. In contrast to Experiment 1 the contingency between the observed cause and the effect decreased across the data sets, whereas the causal power remained stable. Participants learned about all the three data sets with order being counterbalanced.

Table 3: Data shown in Experiment 2

	Data Set 1	Data Set 2	Data Set 3
$P(c)$	0.50	0.50	0.50
$P(e\|c)$	0.70	0.80	0.90
$P(e\|\sim c)$	0.00	0.33	0.67
ΔP	0.70	0.47	0.23
Power p_c	0.70	0.70	0.70

60 students from the University of Göttingen were randomly assigned to one of the three learning conditions. The same dependent variables as in Experiment 1 were collected.

Figure 3 shows the results for the estimates of the causal impact of the observed and the hidden cause.

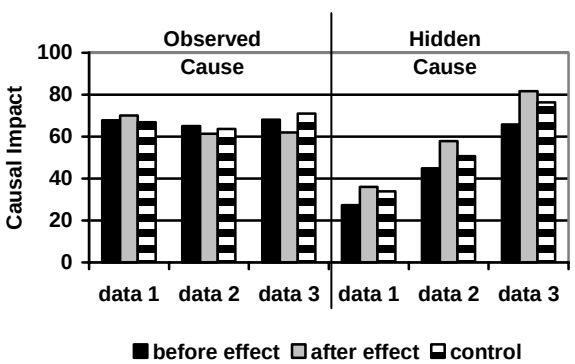


Figure 3: Mean ratings of causal impact for the observed cause (left) and the unobserved cause (right) in Experiment 2.

An analysis of variance of the impact ratings for the observed cause yielded no significant effects, which is in line with the predictions of Power PC theory. As in Experiment 1 the estimated impact of the hidden cause rose significantly across the data sets, $F(2,114)=65.7$, $MSE=408.2$, $p<.01$. This finding is consistent with the predictions of all discussed theories. There was also a significant difference between learning conditions, $F(2,57)=4.06$, $MSE=591.8$, $p<.05$. Participants in the 'prediction after effect' condition rated the impact of the hidden cause to be higher than in the other two conditions. This results points in the same direction as the results of Experiment 1 indicating that predictions with effect information may have sensitized participants to the role of the hidden cause.

Table 4: Mean estimates of dependence between observed and unobserved causes. The numbers express contingencies.

	Generated Dependence			Estimated Dependence		
	Data Set 1	Data Set 2	Data Set 3	Data Set 1	Data Set 2	Data Set 3
Before Eff.	-23.8	-33.2	-29.2	-4.1	8.2	-8.2
After Eff.	-3.9	-20.4	-35.4	12.8	-10.8	-1.0
Control	-	-	-	9.5	9.2	-6.0

Table 4 shows the results concerning the assumed dependence between the two causes. Although the random interventions were expected to increase the salience of independence, participants generated a negative dependence between the two causes which rose across the data sets, $F(2,76)=6.97$, $MSE=510.1$, $p<.01$. The interaction also turned out to be significant, $F(2,76)=3.57$, $MSE=510.1$, $p<.05$. The negative ratings decreased more strongly when participants had received effect information than in the contrasting condition ("prediction before effect"). As in Experiment 1 this trend can be traced to an increase in the generated probability of the hidden cause in the absence of the observed cause. In contrast, the estimated dependencies did not statistically differ. The results are consistent with an independence assumption. Thus, in this experiment there was a clear dissociation between online and posthoc judgments.

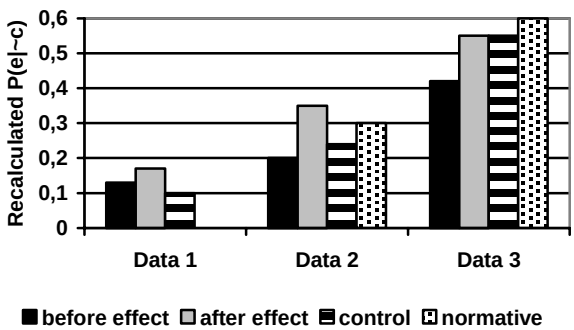


Figure 4: Mean recalculated probabilities of the effect in the absence of the observed cause (Experiment 2).

Figure 4 is based on an analysis of the coherence of the estimates with the constraints from the learning data (using the same method as in Experiment1). It can again be seen that participants honored the normative constraints.

Conclusions

All current theories of causal reasoning consider hidden causes that may also influence observable effects. In most theories independence between the observable cause and the hidden cause is the default assumption, which is a precondition for giving precise estimates for the causal strength of the observed cause-effect relation. Whereas associative theories create independence by assuming constant presence of alternative causes, Power PC theory and Bayesian models are more flexible. Typically these models assume a varying independent hidden cause. However, these theories can also model situations violating the independence assumption by providing bounds for consistent estimates. All theoretical accounts agree that the impact of the hidden cause has to be at least as high as $P(e|\sim c)$. We found evidence in both experiments that participants honored this constraint. Moreover, our analyses showed that participants' judgments about the probability and impact of the hidden cause were in most conditions coherent with the data.

Furthermore, we assessed participants' assumptions about the statistical relation between the observed and the hidden causes. In Experiment 1 learners passively observed the causal relations. In this experiment participants expressed that the causes were positively correlated when $P(e|\sim c)$ was low but they assumed a negative correlation when $P(e|\sim c)$ was high. The generated and estimated probabilities suggest that participants may have assumed that $P(e|\sim c)$ is an indicator of the probability of the hidden cause in the absence of the observed cause ($P(a|\sim c)$) and an indicator of the impact of the hidden cause (p_a) but that $P(e|c)$ conveys little information about the probability of the hidden cause conditional upon the presence of the observed cause ($P(a|c)$). As a consequence participants only adapted their guesses about $P(a|\sim c)$ to the observed $P(e|\sim c)$ while sticking with the initial assumption about $P(a|c)$ irrespective of $P(e|c)$.

In Experiment 2 we increased the salience and plausibility of independence between the alternative causes by letting participants randomly manipulate the observable cause. And indeed the final estimates expressed the assumption of independence. However, surprisingly participants generated a negative correlation in their trial-by-trial predictions. Using the explanation we gave for Experiment 1, this pattern implies that the initial assumption of $P(a|c)$ was at a relatively low level. People may find it unlikely that two independent actions are performed simultaneously by coincidence. In addition participants may erroneously overapply the 'principle of explaining away' (Pearl, 1988) in this task. This principle states that it is generally true that alternative independent causes are less likely in the subset of events in which the cause and effect are present as compared to the whole set of events in which only the effect has occurred. However, in the overall set of events causes should still exhibit independence regardless of the order in which the causal events are experienced. Another related possible explanation might be that people are reluctant to consider overdetermination of effects. Since one cause suffices to explain the effect, assuming a second hidden cause is not necessary. Intuition tells us that one cause is enough for the presence of an effect. It is interesting to see that this intuition seems particularly strong when participants consider single trials of cause-effect patterns. In this situation learners have to decide whether one or two causes generated the effect. Looking back at the learning set at the end of the learning phase seems to decrease the salience of these possible cases of overdetermination, which may be the reason for the interesting dissociation between the tendency to assume negative correlations in online judgments but independence in the summary judgments at the end.

Theoretical Implications

Our results contradict the assumption of associative theories that learners assume constant presence of alternative, hidden causes. The results also indicate that independence of varying causes is not the general default assumption. The online predictions revealed a tendency to assume correlations between alternative causes. Both Power PC theory and causal Bayes nets allow modeling this assumption. Although causal power may in these cases not always be numerically identifiable, these theories can provide constraints for plausible estimates. Future research will have to explore the boundary conditions and the generality of people's assumption across different tasks. The observed dissociations in the present studies indicate that a simple account may be unlikely.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Glymour, C. (2001). *The mind's arrow. Bayes nets and graphical causal models in psychology.* Cambridge, MA: MIT Press.

Luhmann, C. C., & Ahn, W.-K. (2003). Evaluating the causal role of unobserved variables. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Erlbaum.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II. Current research and theory* (pp. 64-99) New York: Appleton-Century-Crofts.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in Neural Information Processing Systems, 15*, 35-42.

Strategy Shifts in Mixed Density Search

Tim Halverson (thalvers@cs.uoregon.edu)
Department of Computer and Information Science, 1202 University of Oregon
Eugene, OR 97403-1202 USA

Anthony J. Hornof (hornof@cs.uoregon.edu)
Department of Computer and Information Science, 1202 University of Oregon
Eugene, OR 97403-1202 USA

Abstract

Visual search in an important aspect of many tasks, but it is not well understood how many aspects of layout design affect visual search. This research investigates, with reaction time and eye movement data, the effect of local density on the visual search of structured layouts of words. Layouts were all-sparse, all-dense, or mixed. Participants found targets in sparse groups faster even after numerosity effects were factored out, and searched sparse groups before dense groups. Participants made slightly more fixations per word in sparse groups, but these were much shorter fixations. Perhaps most interesting, roughly halfway through searching each mixed layout, participants appeared to switch search strategies with respect to the number of fixations per group of words and fixation duration. When dense groups were searched early in a trial, search strategies were more similar to search strategies in the all-sparse layouts. When searched later in a trial, search strategies were more similar to search strategies of all-dense groups. When combining densities in a layout, it may be beneficial to place important information in sparse groups.

Introduction

It is through visual search that people locate the content and controls for many tasks. Yet, it is not well understood how many layout design practices affect visual search. A large body of basic research on visual search exists in psychology (for example, Greene & Rayner, 2001; Hayhoe, Lachter, & Moeller, 1992; Shen, Reingold, & Pomplun, 2000; Treisman, 1998). Many phenomena have been observed and many theories have been proposed to explain them. However, there has been comparably little research on how to apply basic psychological phenomena in a practical setting. A good applied theory of visual search is needed.

Previous research has investigated the extent to which theories from basic research apply to more ecologically valid tasks in human-computer interaction. One such line of research investigated the visual search of hierarchical layouts with experimentation and cognitive modeling (Hornof, 2001, in press; Hornof & Halverson, 2003). In these experiments, participants searched for a precued target item in labeled or unlabeled layouts. In the labeled layouts, groups had headings and the participant was precued with the target group heading as well as the target item. In the unlabeled layouts, the groups had no headings. It was found that a useful visual hierarchy motivated fundamentally different search strategies. That is, when useful group headings are present, people will first search the headings and then the group content.

The current study extends the work in Hornof (2001) and Hornof and Halverson (2003) by investigating the visual search of more complex non-hierarchical layouts. The purpose of this research is to (a) further inform the development of a predictive tool for evaluation of visual layouts, and (b) contribute to the theories of applied visual search in human-computer interaction.

Varying the density of text and objects is one common design practice used to establish grouping and hierarchy in visual displays (Mullet & Sano, 1995). This paper reports a study that investigates the effect of varying local density on visual search strategies of two-dimensional menu-like lists of words.

The density of items in a display is one factor that has been shown in effective field of view (EFV) studies to affect the number of items that can be perceived in a single fixation. EFV, also referred to as the useful field of view or perceptual span, is the region from which the visual perceptual system processes information in a single fixation. There have been many studies on EFV for various tasks (for example, Bertera & Rayner, 2000; Mackworth, 1976; Rayner & Fisher, 1987; Reingold, Charness, Pomplun, & Stampe, 2001). These studies have found a limited region in the visual field that is sufficient for normal perception of static scenes. This region can be centered on the point of fixation or can be asymmetric with respect to the point of fixation. In addition, these studies have found that the EFV varies in size by type of stimuli, type of task, and task difficulty.

Bertera and Rayner (2000) varied the spacing (density) between a fixed number of randomly placed characters in a search task. They found that search time decreased and the estimated number of letters processed per fixation increased as the density increased. Mackworth (1976) showed similar results in a study in which participants searched for a square among uniformly distributed circles on a scrolling vertical strip. Ojanpää, Näsänen, and Kojo (2002) studied the effect of spacing on the visual search of word lists, and found that as the vertical spacing between words increased (i.e. as density decreased), search time also increased. In general, research examining the interaction between EFV and density has found that the visual search of more dense stimuli is faster per object, with the decrease in the number

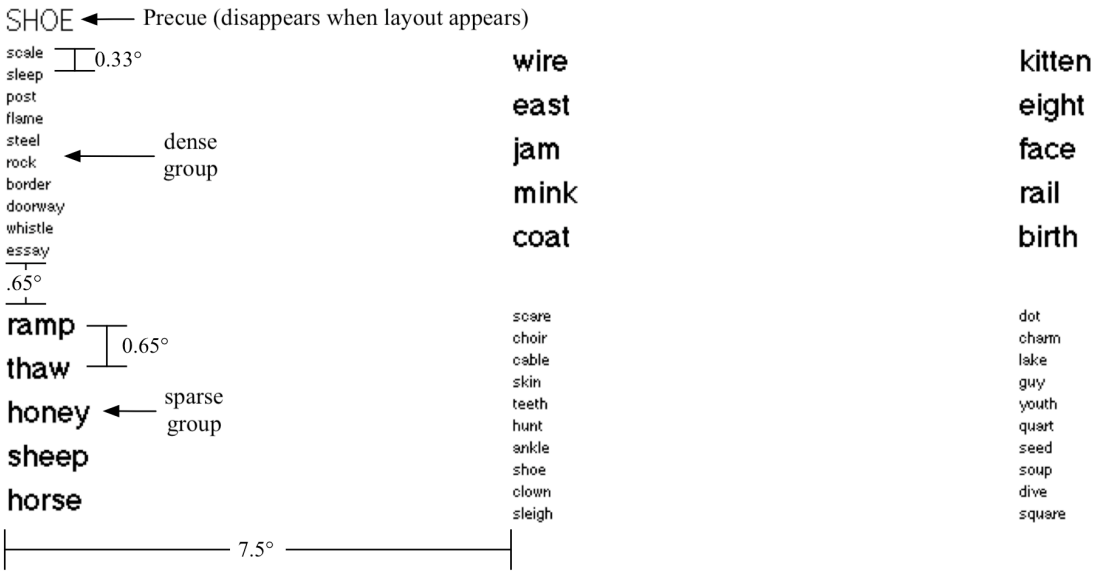


Figure 1: A mixed-density layout. All angle measurements are in degrees of visual angle.

of fixations required to find the target being the largest factor.

Density may be measured as *overall density* or *local density*. Overall density is the number of items per degree of visual angle over an entire layout. Local density is the number of items per degree of visual angle within a visually distinct group.

Besides affecting the search time and number of items inspected per fixation, local density may also affect the *order* of inspection. Several studies have found that visual attention is drawn to "more informative" stimuli (for example, Berlyne, 1958; Mackworth & Morandi, 1967). "More informative" is often defined as regions having greater contour in pictorial stimuli. For example, with geometric shapes, angles are considered more informative than straight lines. Yet, it is not readily known how to predict *a priori* which of two stimuli are more informative. One plausible factor of "informativeness" is local density. It may be that regions with a higher local density are more informative since they are more likely to contain more angles.

The following hypotheses were tested in this study:

H1: The search time per word is greater in sparse layouts than in dense layouts.

H2: Dense regions will be searched before sparse regions.

The following experiment builds on previous research by investigating the extent to which previous findings hold in tasks that are more ecologically valid than those used in Bertera and Rayner (2000) and Mackworth (1976). While these previous studies are very informative, the stimuli are single characters or simple shapes. It is unclear whether the same phenomena will be seen with stimuli in which the items are more complex, such as words, or when density changes within a visual layout. One spatial property – local density – was manipulated in this study.

Method

Participants

Twenty-four people, 10 female and 14 male, ranging in age from 18 to 55 years of age (mean = 24.5) from the University of Oregon and surrounding communities participated in the experiment. The participants were screened as follows: 18 years of age and older; experienced using graphical user interfaces (such as Microsoft Windows or Macintosh); no learning disability; normal use of both hands; and normal or corrected-to-normal vision. Participants were paid $10, plus a bonus that ranged from $0 to $4.54 based on their performance.

Apparatus

Visual stimuli were presented on a ViewSonic VE170 LCD display set to 1280 by 1024 resolution at a distance of 61 cm that resulted in 40 pixels per degree of visual angle. The experimental software ran on a 733Mhz Apple Power Macintosh G4 running OS X 10.2.6. The mouse was an Apple optical Pro Mouse, and the mouse tracking speed was set to the fourth highest in the mouse control panel.

Eye movements were recorded using an LC Technologies Eyegaze System, a 60 Hz pupil-center/corneal-reflection eye tracker. A chinrest was used to maintain a consistent eye-to-screen distance.

Stimuli

Figure 1 shows a sample layout from one mixed-density trial. All trials contained six groups of left-justified,

vertically-listed black words on a white background. The groups were arranged in three columns and two rows. Columns were 7.5 degrees of visual angle from left edge to left edge. Rows were separated by 0.65 degrees of visual angle.

There were two types of groups with different local densities: *Sparse* groups contained five words of 18 point Helvetica font with 0.65 degrees of vertical angle between the centers of adjacent words (0.45° for word height, and 0.2° for blank space). *Dense* groups contained 10 words of 9 point Helvetica font with 0.33 degrees of vertical angle between the centers of adjacent words (0.23° for word height, and 0.1° for blank space). Both types of groups subtended the same vertical visual angle.

There were three types of layouts: *sparse*, *dense*, and *mixed-density*. Sparse layouts contained six sparse groups. Dense layouts contained six dense groups. Mixed-density layouts contained three sparse groups and three dense groups. The arrangement of the groups in the mixed-density layouts was randomly determined for each trial. Sparse and dense layouts were identical to the mixed-density layout, with the exception of group densities.

This experiment was designed, in part, to determine the effect of combining multiple local densities in a single layout. Combining multiple local densities necessitated maintaining the number, size (in degrees of visual angle), and spacing of groups between layouts. Therefore, text size and number of words per group were varied to produce different local densities. Text size often covaries with local density in real-world tasks.

The words used in each trial were selected randomly from a list of 765 nouns generated from the MRC Psycholinguistic Database (Wilson, 1988). No word appeared more than once per trial. The words in the list were selected as follows: three to eight letters, two to four phonemes, above-average printed familiarity, and above-average imagability. Five names of colors and thirteen emotionally charged words were removed.

The target word was randomly chosen from the list of words used in each trial. The participant was precued with the target word before each layout appeared. The precue appeared at the same location every time, directly above the top left word in the layout, in 14 point Geneva font.

Procedure

Each trial proceeded as follows: The participant studied the precue; clicked on the precue to make the precue disappear and the layout appear; found the target word; moved the cursor to the target word; and clicked on it.

The trials were blocked by layout type. Each block contained 30 trials, preceded by five practice trials. The blocks were fully counterbalanced.

At the start of each experiment, the eye tracker was calibrated to the user. The calibration procedure required the participant to fixate a series of nine points until the average error between the predicted point of gaze and the actual location of the points fell below an error threshold (approximately 6.35 mm). During the execution of the experiment, an objective measure of the eye tracker's calibration was taken during each trial as described in Hornof and Halverson (2002). In short, if the calibration had deteriorated below a threshold (2.13 cm), a calibration was automatically initiated before the next trial. In addition, the trial in which the error was found was not analyzed, and a new trial was added to the block.

To separate visual search time from mouse pointing time, the point completion deadline was used (Hornof, 2001). In short, participants were instructed to not move the mouse until the target was found. Once the mouse was moved more than five pixels in any direction, they had a small amount of time (determined by Fitts' law) to click on the target. If this time was exceeded, a buzzer sounded and the trial was recorded as an error. The trial in which the error occurred was not analyzed, and a new trial was added to the block.

Results

Since dense groups contained more words, the following analyses were conducted after normalizing for the number of words per layout. This was accomplished by dividing the search time and number of fixations per trial by half of the number of words in the layout.[1] Table 1 shows the mean search time per word, the mean number of fixations per word, and the mean fixation duration for each layout type. The mean search time per word, mean fixations per word, and mean fixation duration for each of the twenty-four

[1] Measures were divided by half based on the assumption that participants, on average, searched half of the items. This assumption is not consequential for analysis purposes.

Table 1: Search time per word, fixations per word, and fixation duration for sparse, mixed-density, and dense layouts.

	Search Time per Word (ms)		Fixations per Word		Fixation Duration (ms)	
Layout	Mean	SD	Mean	SD	Mean	SD
Sparse	208.25	49.10	.69	.16	250.44	33.21
Mixed	253.58	61.78	.70	.14	306.97	48.81
Dense	265.11	54.52	.62	.14	369.65	67.89

n=24

participants were analyzed using repeated-measures ANOVAs. Eye movements that started before the precue was clicked and after the target was clicked are excluded from all eye movement analysis. An alpha level of .05 was used for all statistical tests.

Participants spent, on average, less time per word in layouts with fewer dense groups, $F(2,46) = 13.94$, $p < .01$. Post-hoc analysis showed that the search time was faster in the sparse than in the mixed layouts ($p < .05$) or dense layouts ($p < .05$); but not different between the mixed and dense layouts ($p > .05$). Participants made slightly fewer fixations per word in layouts with more dense groups, $F(2,46) = 3.25$, $p = .05$. Post-hoc analysis showed that participants used fewer fixations per word in the dense layouts than in the mixed layouts ($p = .01$). Conversely, participants' fixations were longer in layouts with more dense groups, $F(2.46) = 61.82$, $p < .01$. Post-hoc analysis showed that participants made longer fixations in the dense layouts than in the mixed layouts ($p < .05$) and longer fixations in the mixed layouts than in the sparse layouts ($p < .05$).

The search time per trial was analyzed by layout uniformity (all one density vs. mixed density) and target group density. Figure 2 shows the results. Locating a target in dense groups took longer than sparse groups, $F(1, 23) = 83.87$, $p < .01$. The mean search time for all-sparse and all-dense was no different than the mean search time for mixed-density layouts, $F(1,23) = 1.03$, $p = .32$. However, there was an interaction between layout uniformity and target group density, $F(1,23) = 16.87$, $p < .01$. In other words, when the target was in a sparse group, participants found the target faster in all-sparse layouts than in mixed layouts; when the target was in a dense group, participants found the target faster in mixed-density layouts than in all-dense layouts. Further, in mixed density layouts, participants found the target faster when it was in a sparse group, ($p < .01$).

Group visitation data were also analyzed. A group was visited if one or more contiguous fixations fell within 1 degree of visual angle of the group. Group revisits were not included in the analysis presented here. The order of group visitation in mixed density layouts was tested by comparing the percentage of visitations to sparse or dense groups for the first through sixth group visit, regardless of the position of each group in the layout. The results are shown in Figure 3. The data show that participants tended to visit sparse groups before dense groups, $\chi^2(5, N = 24) = 500.04$, $p < .01$.

The mean number of fixations per group and mean fixation duration per group were analyzed. Group revisits were not included in the analysis presented here because it was assumed that the participants' behavior may differ within groups already visited. Additionally, the final groups visited were not included because it was assumed that the participants' behavior may differ within the group in which the target was found. Repeated-measures ANOVAs were conducted to test the effects of group density, layout type (all one density or mixed density), and order of group visit. Figure 4 shows the number of fixations per group as a function of the order in which groups were visited, regardless of the group position in the layout. Each layout type is plotted separately. Mixed layouts are further separated by the visits to dense versus sparse groups. Figure 5 is similar to Figure 4, but shows the mean fixation duration.

The overall number of fixations in all-dense and all-sparse layouts was no different than in mixed-density layouts, $F(1,9) = 2.69$, $p = .14$. The fixations in mixed density layouts are longer than in other layouts, $F(1,9) = 11.22$, $p < .01$. Participants used more fixations per group in dense groups than in sparse groups, $F(1,9) = 112.30$, $p < .01$. Fixation durations were longer in dense groups than in sparse groups, $F(1,9) = 139.36$, $p < .01$.

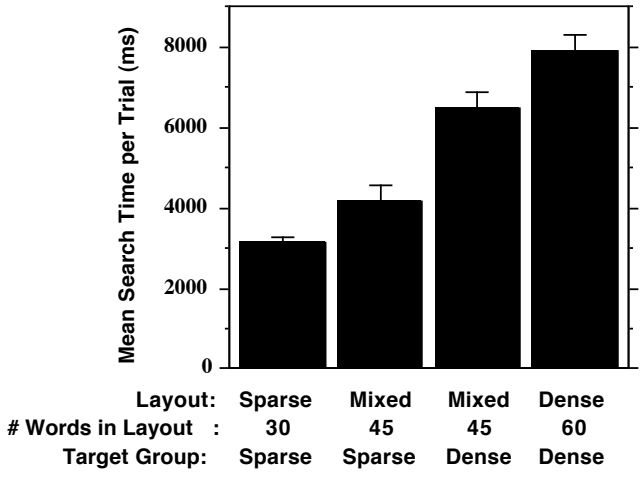


Figure 2: Search time for trials in which the layout was sparse, mixed-density, or dense, and the target was in either a sparse or dense group. Error bars indicate ±1 standard error.

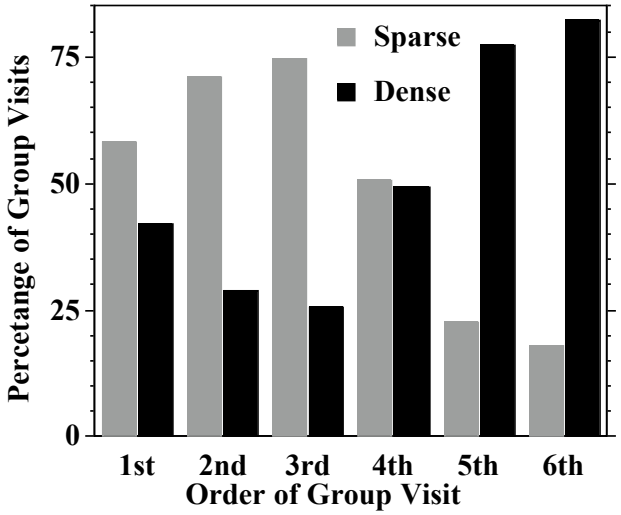


Figure 3: The percentage of visits in mixed density layouts that were to sparse or dense groups, as a function of the order in which groups were visited.

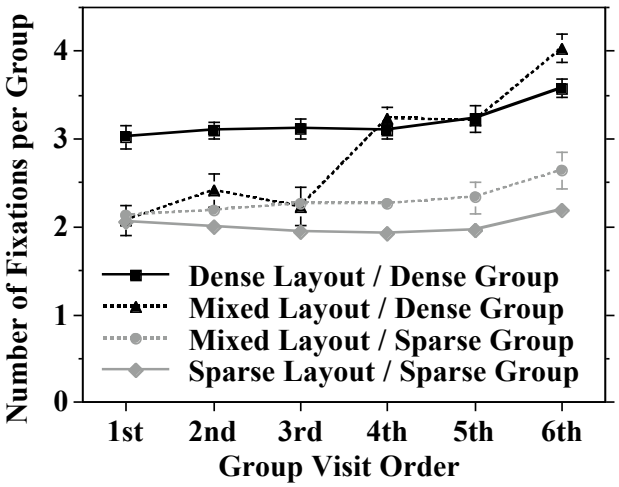


Figure 4: Mean number of fixations per group as a function of layout, the density of the group currently visited, and order of the visit. Error bars indicate ±1 standard error.

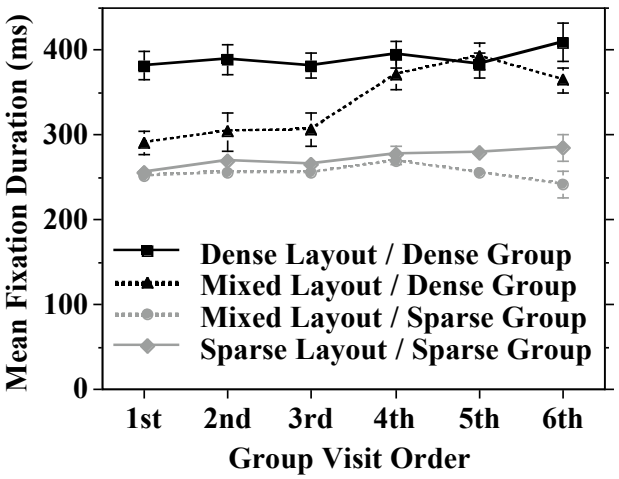


Figure 5: Mean fixation duration by group as a function of layout, the density of the group currently visited, and the order of visit. Error bars indicate ±1 standard error.

Participants used more fixations per group as search progressed, $F(5,45) = 8.14$, $p < .01$. Contrast analysis revealed that the sixth group visited received more fixations than all other groups (all p's < .05), but there were no differences between any other orderings (all p's > .05). Fixation durations tended to be longer for groups visited later than for groups visited earlier, $F(5,45) = 4.89$, $p < .01$. The following interactions were also found in the fixations per group data: The difference between the number of fixations in sparse and dense groups was greater in uniform density layouts than in mixed density layouts, $F(1,9) = 5.20$, $p = .05$. As search progressed (i.e. from left to right in Figure 4), the number of fixations increased faster in mixed-density layouts than in all-dense and all-sparse layouts, $F(5,45) = 6.7$, $p < .01$. The number of fixations increased faster in dense groups than in sparse groups, $F(5,45) = 5.05$, $p < .01$.

Discussion

The data counter the study's first hypothesis – that the search time per word is greater for sparse layouts than for dense layouts. The search time data reported here demonstrate that people actually spent *less* time per word searching sparse layouts. Participants adopted a more efficient eye movement strategy that used slightly more, but much shorter, fixations in the sparse groups. This result is contrary to the search time results found by Bertera and Rayner (2000) and Ojanpää, et al. (2002) in which the search time decreased as the density increased. This discrepancy may be due to the way in which density is manipulated. In the previous studies, the spacing between items was varied. This could result in a need for more saccades, as both Rayner (2000) and Ojanpää, et al. (2002) found, to move the EFV over the next group of unprocessed stimuli. In the current study, the size of words (i.e. font size) was varied. The smaller words were more tightly packed, which could have made it more difficult for people to fixate directly on the desired words, requiring more saccades as found in this study. It may be that although various factors affect local density, they do not all affect visual search of those densities in the same way.

The data also counter this study's second hypothesis – that participants will search dense groups first. A preference for search order as a function of group density was found. However, it was in the opposite direction than predicted. The search time data show that when the target was in a sparse group, the mean search time was much closer to that of the sparse layouts, and that when the target was in a dense group, the mean search time was much closer to that of the dense layouts. If one density were consistently searched before the other, then we would expect the search time for targets located in groups of a preferred density to be lower than the search time for targets located in the other groups, which is what we observed. The data suggest that the participants tended to search the sparse groups first. This preference was confirmed with analysis of the eye movements in the mixed layouts. As is seen in Figure 3, participants tended to look at sparse groups first.

While the first group visited was quite often a dense group, as seen in Figure 3, this is expected as the top-left group in the layout was equally likely to be either sparse or dense and 89% of all initial fixations were to that group. These are likely *anticipatory fixations*, as predicted and observed by Hornof and Halverson (2003).

A trend that emerged from the data analysis is evidence of a shift in search strategy between the third and fourth group visited in mixed layouts, right around the time that participants tended to switch from sparse groups to dense groups. When a dense group was one of the first three groups visited, the participants tended to search the dense groups in the same manner as sparse groups, with fewer and shorter fixations. Yet, when the participants searched

through the all-dense layouts, all-sparse layouts, or sparse groups in the mixed-density layouts, no significant changes in oculomotor programming were found at any point during the search. This suggests that the participants started searching mixed-density layouts with a more eager approach, adopting the search strategy used for the preferred sparse-density groups; then, as the search progressed and the target had not been found, participants reverted to a different strategy for dense groups.

Conclusion

This research investigates the effect of local density on visual search of structured, two-dimensional layouts. It is shown that sparse groups of words are searched faster and, when presented with dense groups, sparse groups are searched earlier than dense groups. This lends support to the practice of displaying important information in less dense groups.

Further, at least in the mixed density task, people appear to apply local search strategies used for sparse groups to all groups, regardless of density, early in the task. At some point in the unfolding of their visual search, approximately halfway through, the participants made a global strategy shift towards a more thorough search of dense groups. This suggests that care should be taken when combining densities in a visual layout. Performance in a mixed density task cannot be predicted by assuming people will search regions of a given density the same as they will in a layout of uniform density. Additional research will determine how these findings generalize to a variety of mixed-density layouts.

Acknowledgments

This research is supported by the Office of Naval Research grant N00014-02-10440 and the National Science Foundation grant IIS-0308244. Both grants are to the University of Oregon with Anthony Hornof as the principal investigator.

References

Berlyne, D. E. (1958). The Influence of Complexity and Novelty in Visual Figures on Orienting Responses. *Journal of Experimental Psychology, 55*, 289-296.

Bertera, J. H., & Rayner, K. (2000). Eye movements and the span of effective stimulus in visual search. *Perception & Psychophysics, 62*(3), 576-585.

Greene, H. H., & Rayner, K. (2001). Eye movements and familiarity effects in visual search. *Vision Research, 41*, 3769-3773.

Hayhoe, M. M., Lachter, J., & Moeller, P. (1992). Spatial Memory and Integration Across Saccadic Eye Movements. In K. Rayner (Ed.), *Eye Movements and Visual Cognition: Scene Perception and Reading*. New York: Springer-Verlag.

Hornof, A. J. (2001). Visual search and mouse pointing in labeled versus unlabeled two-dimensional visual hierarchies. *ACM Transactions on Computer-Human Interaction, 8*(3), 171-197.

Hornof, A. J. (in press). Cognitive Strategies for the Visual Search of Hierarchical Computer Displays. *Human-Computer Interaction*.

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye tracking data by using required fixation locations. *Behavior Research Methods, Instruments, and Computers, 34*(4), 592-604.

Hornof, A. J., & Halverson, T. (2003). *Cognitive strategies and eye movements for searching hierarchical computer displays.* Proceedings of the Conference on Human Factors in Computing Systems, Ft. Lauderdale, FL.

Mackworth, N. H. (1976). Stimulus Density Limits the Useful Field of View. In R. A. Monty & J. W. Senders (Eds.), *Eye Movements and Psychological Processes*. Hillsdale, NJ: Lawrence Erlbaum.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2*(11), 547-552.

Mullet, K., & Sano, D. (1995). *Designing Visual Interfaces: Communication Oriented Techniques*. Englewood Cliffs, New Jersey: Prentice Hall PTR.

Ojanpää, H., Näsänen, R., & Kojo, I. (2002). Eye movements in the visual search of word lists. *Vision Research, 42*(12), 1499-1512.

Rayner, K., & Fisher, D. L. (1987). Eye movements and the perceptual span during visual search. In J. K. O'Regan & A. Levy-Schoen (Eds.), *Eye Movements: From Physiology to Cognition*. Amsterdam: North-Holland.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science, 12*(1), 48-55.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Distractor Ratio Influences Patterns of Eye Movements During Visual Search. *Perception, 29*, 241-250.

Treisman, A. (1998). The Perception of Features and Objects. In R. D. Wright (Ed.), *Visual Attention* (Vol. 8). New York: Oxford University Press.

Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine Usable Dictionary, Version 2. *Behavior Research Methods, Instruments, and Computers, 20*, 6-11.

Lies in Conversation:
An Examination of Deception Using Automated Linguistic Analysis

Jeffrey T. Hancock (jeff.hancock@cornell.edu)
Department of Communication, Cornell University
320 Kennedy Hall, Ithaca, NY 14850 USA

Lauren E. Curry (lec26@cornell.edu)
Department of Communication, Cornell University
320 Kennedy Hall, Ithaca, NY 14850 USA

Saurabh Goorha (sg278@cornell.edu)
Department of Communication, Cornell University
320 Kennedy Hall, Ithaca, NY 14850 USA

Michael T. Woodworth (mwoodwor@dal.ca)
Department of Psychology, Dalhousie University
Life Sciences Center, Halifax, NS, B3K 1L4, Canada

Abstract

The present study investigated changes in both the sender's and the receiver's linguistic style across truthful and deceptive dyadic communication. A computer-based analysis of 242 transcripts revealed that senders used more words overall, increased references to others, and used more sense-based descriptions (e.g., seeing, touching) when lying as compared to telling the truth. Receivers naïve to the deception manipulation produced more words and sense terms, and asked more questions with shorter sentences when they were being lied to than when they were being told the truth. These findings are discussed in terms of their implications for linguistic style matching.

Introduction

Maxims such as "honesty is the best policy" and "let the truth be told" reinforce the notion that telling the truth is the best way to communicate. When telling everyday lies, then, deceivers must be careful to assume a position of sincerity in order to make their partners believe them and avoid being viewed in a negative light. In fact, this feat might not be very difficult to accomplish. Previous research suggests that it is quite difficult to catch a liar as deception detection rates in many experiments are not much better than chance (Vrij, 2000).

In general, there are three methods for trying to detect deceit. The first method focuses on vocalic and physical nonverbal behaviors (e.g., movements, smiles, voice pitch, speech rate, stuttering, and eye gaze) (Vrij, 2000). The second method involves measuring physiological responses with various technologies, such as polygraph machines (Vrij, Edward, Roberts, & Bull, 2000).

The third method is concerned with the content of what is said (e.g., verbal behavior, as well as a study of linguistic properties of liars' texts). For example, previous research suggests that liars tend to make less sense and tell less plausible stories (e.g., making discrepant and ambivalent statements), among other verbal characteristics (for review, see DePaulo, Lindsay, Malone, Mulenbruck, Charlton, & Cooper, 2003).

The present study employs automated linguistic analysis, in which a computer program is used to analyze the linguistic properties of texts, to examine the verbal content of deceptive and truthful conversations. As Pennebaker, Mehl, and Niederhoffer (2003) note, words used in daily interactions reveal both psychological and social aspects of peoples' worlds. Certain words and parts of speech can be markers of emotional, psychological, and cognitive states. Given that deceiving others likely involves changes in emotional or psychological states, linguistic cues detected using automated techniques may indicate lying in conversation.

Linguistic Indicators of Deception

A review of the relatively small literature concerned with automated linguistic analyses of deception indicates that, to date, at least four main types of linguistic cues have been associated with deception: 1) word counts 2) pronoun usage, 3) words pertaining to feelings and the senses, and 4) exclusive terms (Burgoon, Buller, Floyd, & Grandpre, 1996; Burgoon, Bliar, Qin, & Nunamaker, 2003; Newman, Pennebaker, Berry, & Richards, 2003; Pennebaker et al., 2003).

Consider first differences in word counts across deceitful and truthful messages. Previous studies have found that senders offer fewer details when lying than when telling the truth (Burgoon et al., 2003; DePaulo et al., 2003; Vrij, 2000). Senders may offer fewer details because they are less familiar with what they are discussing, or because they are trying to avoid providing details that may be inconsistent

with their fabrication. As such, senders may be expected that deceptive interactions would be characterized by fewer words on the part of the sender.

With regard to pronoun usage, Newman et al. (2003) observed that individuals consistently used first person singular pronouns less frequently when lying than when telling the truth. Using first person pronoun words such as "I," "me,"or "my" involves taking ownership of a statement, and deceivers may refrain from using these first person pronouns due to either a lack of personal experience or a desire to dissociate themselves from the lie being told. The findings regarding the use of second and third person pronouns are less consistent. Some studies have found that liars are less likely to use second and third person pronouns (Newman et al., 2003) while other studies have found that liars in fact use more second and third person pronouns (Ickes, Reidhead, & Patterson, 1986). According to Ickes et al. (1986), senders who are careful about constructing deceptive messages will exhibit an increased other-focus and therefore a higher use of second and third person pronouns. Finally, DePaulo et al. (2003) also found that liars are more likely to use third person pronouns in their deceptive interactions.

Research examining verbal cues associated with feelings and sense terms (e.g., see, touch, listen, etc.) suggests that deceivers tended to use more expressiveness, which includes both negative and positive forms of emotion, compared to truth-tellers (Burgoon et al., 2003). In addition, senders may be more likely to use sense words in an effort to create a detailed story to avoid eliciting skepticism from the deceiver (Burgoon et al., 2000).

Finally, previous research also suggests that liars use fewer exclusive words than truth-tellers (Newman et al., 2003). Exclusive words include prepositions and conjunctions such as "but," "except," "without," and "exclude." These words require a deceiver to discuss what is in a category and what is not. As such, deceivers may find it a more complex task to invent what was done versus what was not done (Newman et al., 2003).

Deception, Conversation and the Receiver

Although the literature on automated approaches to linguistic analysis of deception suggests that word counts, pronouns, feeling words, and exclusion words may predict deception, previous research is limited in two important ways. First, previous research has been limited primarily to analyses of deception in the context of monologues rather than in conversational contexts. For example, Newman et al. (2003) conducted five studies in which participants discussed a given topic by writing about it, talking about it to a video camera, or by typing their views on it. In these cases, only the liar's behavior was analyzed because there was no target of those lies present during the studies. Given that most lies tend to occur during conversations with others (DePaulo, Kashy, Kirkendol, & Epstein, 1996), and given the fact that language use in conversation differs in important ways from language use in monologues (Clark, 1996; Schober & Clark, 1989), the focus of previous research on monologue-based deception may limit its applicability to everyday conversation.

A second, and related, weakness is that previous research on linguistic predictors of deception has focused almost exclusively on the sender (i.e., the teller of the deception or the truth). For example, Newman et al. (2003) examined only a sender's handwriting, videotapes, and typed transcripts. In no case were the reactions of receivers (i.e., the targets of deceptive messages) studied. However, in conversations there is a reciprocal exchange between senders and receivers that can have important effects on deceptive behavior (Burgoon et al., 1996; Burgoon, Buller, & Floyd, 2001). As such, it may be important to look at both parties when examining interactions. If senders alter their behavior in systematic ways when lying versus when they are telling the truth, as previous research suggests, then an important question that remains to be addressed is whether receivers will also behave differently when lied to than when they are told the truth.

One possible outcome is that receivers will engage in linguistic style matching, which refers to the degree to which two people in conversation adjust their own speaking behavior, or style, to match their partners' behavior (Niederhoffer & Pennebaker, 2002). The observation that people vary their words on a turn-by-turn level when in conversations with others is assumed to reflect the coordination processes inherent in natural conversations (Grice, 1989; Niederhoffer & Pennebaker, 2002). Indeed, participants in conversations have been known to exhibit similar types of concurrent behaviors (both vocal and nonvocal), kinesics, proxemics, facial expressions, and word usage, regardless of topic content (Niederhoffer & Pennebaker, 2002).

If, as linguistic style matching suggests, people in conversation adjust their linguistic behavior to that of their partners, then any differences in linguistic behavior by senders across deceptive and truthful communication should also be observed in the receiver's behavior. That is, receiver's behavior should mirror the behavior of the sender in terms of word usage and linguistic variables across deceptive and truthful communication. If receivers engage in linguistic style matching during deceptive interactions, then receivers, like senders, should produce fewer words, fewer first person pronouns, more second and third pronouns, more exclusive words and negations, and more words pertaining to the senses.

Finally, there may also be receiver activities that do not simply match those of the sender. For example, if a receiver becomes suspicious of the sender's truthfulness, the receiver may probe their partner more frequently, perhaps by asking additional questions.

The present study examined the linguistic styles of senders and received engaged in truthful and deceitful conversations. The conversations were conducted in a text-based, computer-mediated setting, in which participants exchanged synchronous messages. A computer-mediated

communication setting was used in the present study for several reasons. The first is that the transcripts were created automatically as the participants interacted. The second is that, because the interaction was entirely text-based, all of the information exchanged by the participants during their interaction was captured in the transcripts (Hancock, in press).

Methods

Participants (n = 66) were upper-level students at a northeastern American university, and they participated for credit in various courses. Participants were randomly paired to form 33 same-sex, unacquainted dyads (15 male and 18 female).

Participants were recruited for a "study of how unacquainted individuals communicate about various conversation topics." Upon reporting to the laboratory, participants were led separately to remote rooms where they completed an initial set of forms, including informed consent.

The general procedure was adapted from Burgoon et al. (2001). All participants were told that they would be having a conversation with an unknown partner. They were instructed that they would discuss 5 topics, which were then provided to the participants on a sheet of paper. The first topic was always "When I am in a large group, I…" This initial topic was designed to allow the participants to become comfortable interacting with their partner, and was not included in any analyses. After this topic, participants began a discussion of the four experimental topics which included: "Discuss the most significant person in your life", "Talk about a mistake you made recently", "Describe the most unpleasant job you have ever had to do" and "Talk about responsibility." There was no time limit and participants were asked to discuss each topic until they had exhausted it and understood each other's responses.

One of the two participants was randomly assigned to the role of sender, and the other to the role of receiver. Senders were asked to sometimes deceive their partners. In particular, they were instructed "to NOT tell 'the truth, the whole truth, and nothing but the truth'" (Burgoon et al., 2001) on two topics, and to be truthful on the other two topics. The two topics in which the whole truth was not to be told were marked with an asterisk on the sheet of paper given to the sender.

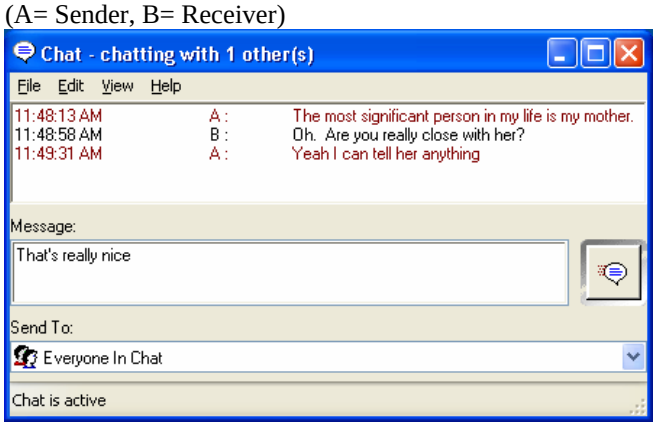
Examples of lies were given to the senders, and it was emphasized that the senders should try to produce lies that were fairly substantial (e.g., saying that they went on a vacation when in fact they did not) rather than small lies (e.g., saying that they went on a vacation from August 4th to the 10th when they actually went from August 5th to the 11th). Senders had approximately five minutes to plan their stories. Receivers were blind to the deception manipulation and were told that they were going to have a conversation with another person and that their role was to keep the conversation going. The same list of topics in the same order was given to the receivers but without any asterisks marking topics.

The sequence in which the topics were discussed, and the order in which the sender lied, was counterbalanced across 16 orders. After the initial ice-breaking topics, senders were instructed to lie on either the next two topics or on the last two topics. Half of the senders followed a truth-first, deception-second order. The remainder followed a reverse order. Because topics followed a diagram-balanced Latin square order within truth and within deception, all topics appeared within a given time period.

Participants discussed the topics in a text-based, computer-mediated setting and performed the task at isolated computer terminals. Participants used one of two desktop computer stations while the experimenter monitored and recorded the interaction from a third station. Once participants were seated at their terminals, the experimenter briefly demonstrated the use of the computer interface, Netmeeting, in which participants typed their message in a private composition window and hit enter to send their message to a shared chat window (see Figure 1). Message transmission was virtually instantaneous.

Once participants finished the discussion task, they were asked to complete a series of questionnaires based on their conversation. The data from these questionnaires are not reported here. After completing the post-interaction questionnaires, each member of the dyad was brought to a common room, and introduced to his or her partner and they were fully debriefed.

Figure 1. Screenshot of the Netmeeting Interface (A= Sender, B= Receiver)



Automated Linguistic Analyses

Both sender and receiver transcripts were converted into separate text files separated by topic. Each dyad produced eight different transcript files: two deception discussions and two truthful discussions for each sender, and two deception discussions and two truthful discussions for each receiver, which produced a total of 264 transcripts.

All transcripts were analyzed using the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker, Francis, & Booth, 2001). This text analysis program was used to create empirically derived statistical profiles of deceptive and truthful communications (Pennebaker et al., 2003), and it

Table 1. Means and (Standard Errors) of the linguistic output variables by role and truth condition.

	Sender		Receiver	
	Lie M (SE)	Truth M (SE)	Lie M (SE)	Truth M (SE)
Word Count	156.11 (17.07)	125.08 (11.20)	157.36 (16.56)	119.61 (10.96)
Words / sentence	10.20 (.97)	9.03 (.53)	8.21 (.42)	9.07 (.59)
1st Person Pronouns	8.01 (.35)	8.52 (.34)	8.08 (.45)	8.92 (.41)
2nd Person Pronouns	2.41 (.31)	2.82 (.32)	2.64 (.32)	2.25 (.22)
3rd Person Pronouns	3.30 (.33)	2.46 (.18)	2.57 (.31)	2.43 (.27)
Negations	2.19 (.21)	1.77 (.16)	2.27 (.19)	2.20 (.21)
Senses	2.47 (.16)	2.09 (.19)	2.49 (.18)	2.18 (.22)
Exclusive Words	4.01 (.27)	4.18 (.32)	3.63 (.22)	3.86 (.31)
Questions	15.88 (2.27)	16.39 (2.32)	15.33 (1.53)	10.84 (1.34)

Note: all statistics represent the percentage of total words in the transcript, with the exception of Word Count, Word per Sentence and Questions variables, which represent absolute totals.

has been used in studies to predict outcome measures like social judgments, personality, psychological adjustment, and health. LIWC analyzes transcripts on a word-by-word basis, including punctuation, and compares words against a file of words divided into 74 linguistic dimensions. For the purposes of this study, only variables relevant to the hypotheses or of potential interest to deception were included, which left 8 variables within the four categories mentioned above: word counts; pronouns; emotion words and words pertaining to the senses; and exclusive words and negations. In addition, question frequency was also analyzed.

LIWC produces the percentage of each variable type by dividing the frequency of the observed variable by the total number of words in the sample. Word counts were not reported as percentages, but as frequency totals.

Results

A 2 (discussion type: truthful vs. deceptive) x 2 (role: sender vs. receiver) repeated measure type General Linear Model (GLM) procedure was conducted on each dependent variable. Table 1 contains the descriptive statistics for each variable.

Overall, more words were produced during deceptive discussions than during truthful discussions F (1,32) = 7.11, $p < .05$. The increase in word count for deception was equivalent for both senders and receivers, F (1,32) < 1, *ns*, and no interaction was observed, F (1,32) < 1, *ns*, suggesting that both senders and receivers used more words when the sender was lying than when the sender was telling the truth.

An analysis of the number of words used per sentence revealed no main effect of discussion type or role. However, a significant interaction between discussion type and role was observed, F (1,32) = 4.07, $p < .05$. Simple effects analyses conducted at each level of discussion type revealed that during truthful discussion senders and receivers produced the same number of words per sentence, F (1,32) < 1, *ns*. In contrast, during deceitful discussion, receivers used marginally fewer words per sentence than senders, F (1,32) = 3.81, $p = .06$. Considered together, these data suggest that receivers used shorter utterances when being lied to than when they were being told the truth, while senders used the same number of words per sentence regardless of discussion type.

The next set of analyses examined pronoun usage. No significant effects were observed for first person pronouns (e.g., "I," "we," "self"). Similarly, no effects were observed for the usage of second person pronouns (e.g., "you"). An analysis of third person pronouns referring to others (e.g., "he," "she," "they"), however, revealed a main effect of role, F (1,32) = 4.68, $p < .05$. Senders used third person pronouns more frequently than receivers. In addition, senders were significantly more likely to discuss others when lying as compared to when they were telling the truth, F (1,32) = 4.57, $p < .05$.

With regard to the production of exclusive words and negations, no reliable effects were observed. Regardless of discussion type, senders and receivers produced the same number of exclusive words and negations.

The next analyses examined the use of words that pertained to the senses (e.g., "see," "touch," "listen"). Participants used significantly more sense words during deceptive conversations than during truthful ones, F(1,32) = 5.34, $p < .05$. No effect of role was observed, F(1,32) < 1, *n.s.*, nor did role interact with discussion type, F(1,32) < 1, *n.s.*

The last analysis was concerned with the number of questions asked during the interactions. A main effect of discussion type was observed, $F(1,32) = 4.02$, $p < .05$. More questions were observed during deceptive discussions than during truthful discussions. This main effect, however, was qualified by a marginally reliable interaction between discussion type and role, $F(1,32) = 3.24$, $p = .08$. Simple effects analyses conducted at each level of role revealed that while senders asked the same number of questions across deceptive and truthful discussion types, $F(1,32) < 1$, *ns.* , receivers asked more questions during deceptive discussions than truthful ones, $F(1,15) = 9.58$, $p < .01$. Considered together, these data suggest that receivers were more likely to ask questions when they were being lied to than when they were being told the truth

Discussion

The primary objective of the present study was to examine the linguistic behaviors of both senders and receivers during dyadic communication that involved both deceptive and truthful discussions. The first question of interest was determining whether the senders' linguistic behavior changed when the sender was being deceptive relative to when the sender was being truthful. The data suggest that, overall, when senders were lying to their partners, they 1) produced more words, 2) used more "other" pronouns (e.g., "he," "she," "they"), and 3) used more terms that described the senses (e.g., "see," "hear," "feel") than when they were telling the truth.

In general, this linguistic profile is consistent with previous research suggesting that senders attempt to construct a more cohesive and detailed story in order to seem believable (Burgoon et al., 1996). For example, the increased number of words observed in the deceptive discussions may reflect the senders' attempts to convey a more complete story when attempting to deceive. Similarly, senders may have increased their use of sense words to enhance the believability of the deception (e.g., "He *saw* her do it."). Finally, the use of other-focused pronouns during deceptive discussions reveals the senders' attempts to shift the focus away from themselves (DePaulo et al., 2003; Ickes, 1986).

The present data, however, differs with previous research in several important respects. For example, previous research suggests that liars tend to use fewer words than truth tellers (Burgoon et al., 2003; DePaulo et al., 2003; Vrij, 2000). Why, then, did senders in the present study produce more words during deceptive discussions than during truthful discussions? One possibility is that the senders in the present study were engaged in conversation with a partner, whereas previous research has focused primarily on deception in monologue formats (e.g., Newman et al., 2003). It is possible, for instance, that senders engaged in conversation used more words in an effort to convince suspicious or skeptical receivers (e.g., Burgoon et al., 2001). Indeed, receivers in the present study asked more questions when they were being lied to than when they were telling the truth (see below), which may have required senders to use more words to address the additional questions.

Similarly, previous linguistic analyses of deception suggest that senders use more negative emotion terms (e.g., Newman et al., 2003; Vrij, 2003). This difference is perhaps not surprising given the differences in discussion topics between the present study and the Newman et al. (2003) study. As noted above, Newman et al. asked participants to lie or tell the truth about highly emotional topics, such as abortion, which may have been more likely to elicit strong emotional verbal content than the more mundane topics employed in the present study (e.g., "Talk about a mistake you made recently.").

The second question of interest was whether the linguistic style of the receivers changed systematically according to whether or not their partners were lying. The data suggest that, in fact, receivers' linguistic profile changed across deceptive and truthful discussion topics. In particular, when being lied to, receivers 1) used more words, in shorter sentences, 2) used more sense words, and 3) asked more questions than when they were being lied to. These observations are particularly striking given the fact that receivers were blind to the deception manipulation.

These data provide relatively robust support for the linguistic style matching model (Niederhoffer & Pennebaker, 2002). First, receivers matched changes in the sender's total word production and use of sense words. Second, like the senders, receivers' use of emotion words and exclusion words did not change across deceptive and truthful conditions. Considered together, the present data suggest that receivers engaged in linguistic style matching.

There were, however, a number of linguistic variables on which receivers and senders diverged. While senders used more other pronouns when lying than when telling the truth, the receivers' use of other pronouns did not differ across discussion types. This observation may reveal the unique motivation of senders to distance themselves from their deception. Perhaps more importantly is the observation that receivers asked more questions and used fewer words per sentence when they were being lied to than when they were being told the truth. These surprising data suggest that the receivers were skeptical of the senders during deceptive conversations. Because senders did not produce more questions when they were being deceptive, the change in the receivers' question-asking behavior does not simply reflect linguistic style-matching. Instead, these data suggest that although receivers were not explicitly aware that their partner was lying to them (i.e., they were blind to the deception manipulation), they were implicitly aware that they were being lied to.

An important limitation of the present study, however, is that participants interacted in a text-based computer-mediated environment. An important question is how these verbal behaviors we observed in text-based conversations will be altered when nonverbal channels of communication are available. While additional research will be required to address this question, communication via the Internet is becoming increasingly ubiquitous. Indeed, millions of people use text-based forms of communication on a daily basis, and previous research suggests that people do tell lies during computer-mediated interactions, such as Email and Instant Messaging, although not as frequently as they do over the phone or in face-to-face contexts (Hancock, Thom-Santelli, & Ritchie, 2004). As such, the present data provide important insights into interpersonal deception in this new communication domain.

Finally, there may not necessarily be a classification of specific words to predict deception, as previous research by Pennebaker et al. (2003) and Newman et al. (2003) may suggest, but deception may be more reliably predicted by looking at the methods of constructing lies. The present research advances our understanding of how linguistic behavior changes according to the truthfulness of the discussion. Lies that take place during conversation tend to include more words, more other-directed pronouns, and more sense words than truths. Equally important, if a receiver is being lied to by someone who fits this linguistic profile, he or she may be more likely to use more overall words and sense terms, and to ask more questions.

Acknowledgments

The authors are grateful to Yufen Chen and Judith Yellin for their assistance in data collection, and to Joseph Walther and Geri Gay for their comments on earlier drafts of the manuscripts.

References

Burgoon, J.K., Bliar, J.P., Qin, T., Nunamaker, J.F. (2003). Detecting deception through linguistic analysis. *Intelligence and Security Informatics, 2665*, 91-101.

Burgoon, J.K., Bonito, J.A., Bjorn, B., Ramirez, A., Dunbar, N.E., Miczo, N. (2000). Testing the interactivity model: Communication processes, partner assessments, and the quality of collaborative work. *Journal of Management Information Systems, 16 (3)*, 33-56.

Burgoon, J. K., Buller, D. B., & Floyd, K. (2001). Does participation affect deception success? A test of the interactivity principle. *Human Communication Research, 27*, 503-534.

Burgoon, J.K., Buller, D.B., Floyd, K., & Grandpre, J. (1996). Deceptive realities: Sender, receiver, and observer perspectives in deceptive conversations. *Communication Research, 23)*, 724-748.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

DePaulo, B.M., Kashy, D.A., Kirkendol, S.E., & Epstein, J.A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*, 979-995.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118.

Friedman, H.S., & Tucker, J.S. (1990). Language and deception. In H. Giles & W.P. Robinson (Eds.), *Handbook of language and social psychology*. New York: Wiley & Sons.

Grice, P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Hancock, J.T. (in press). Verbal irony use in computer-mediated and face-to-face conversations. *Journal of Language and Social Psychology*.

Hancock, J.T., Thom-Santelli, J., & Ritchie, T. (2004). Deception and design: Effects of communication technology on lying behavior. In *Proceedings of CHI 2004*.

Ickes, W., Reidhead, S., Patterson, M. (1986). Machiavellianism and self-monitoring: As different as "me" and "you." *Social Cognition, 4*, 58-74.

Niederhoffer, K.G. & Pennebaker, J.W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*, 337-360.

Newman, M.L., Pennebaker, J.W., Berry, D.S., & Richards, J.M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665-675.

Pennebaker, J.W., Mehl, M.R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547-577.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*, 211-232.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice.* Chichester, England: John Wiley & Sons.

Vrij, A., Edward, K., Roberts, K.P, & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24*, 239-263.

The transfer of logically general scientific reasoning skills

Anthony M. Harrison (anh23@pitt.edu)
Christian D. Schunn (schunn@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara St
Pittsburgh, PA 15260 USA

Abstract

Extending the paradigm introduced by Schraagen (1993), two near-expert groups and novices completed two scientific discovery tasks, one from each of the experts' domains. In this way, both groups designed simulated experiments from within and outside of their domain. The role of domain familiarity on the application of general scientific reasoning skills is explored by contrasting the performance of the experts in their domain to that in the unfamiliar domain. Results indicate that at the graduate level, near-experts are able to apply general scientific reasoning skills across dissimilar domains, while novices still have difficulty with the transfer.

Introduction

How common are scientific reasoning skills across different domains of practice? Schraagen (1993) as well as Schunn and Anderson (1999) address this issue in their experiments. Both studies relied on the same basic paradigm: two groups of expert researchers and one group of novices were asked to design and conduct a series of experiments in a scientific discovery task. One of the expert groups was familiar with the domain that the task was drawn from (e.g. cognitive psychologists working on a memory experiment), whereas the other was less familiar with the domain, but still from the same general field (e.g. social psychology). The scientific reasoning skills (e.g. experimental design, hypothesis generation, and data evaluation) exhibited by both expert groups were similar and mapped cleanly onto those mentioned in the literature (e.g. Klahr & Dunbar, 1988; Dunbar, 1993; Chinn & Malhotra, 1999). The novices showed a similar pattern of failures as those found by other researchers focusing on non-scientists (e.g. Kuhn, Schauble & Garcia-Mila, 1992; Klahr, Fay & Dunbar, 1993; Zajchowski & Martin, 1993).

Almost all studies that have focused on non-scientists have found remarkably poor performance (see Detterman, 1992 for a review) on scientific reasoning skills. Fortunately, those that have studied practicing scientists in their own domain have found just the opposite (e.g. Dunbar, 1997). Why is it that expert scientists do so well outside of their domain of expertise, yet novices do so poorly? Schunn and Anderson (1999), as well as Schraagen (1993), report that the differences found between the practicing scientists and the novices could not be accounted for by general reasoning ability differences. This leaves two alternative explanations: context of the problem influencing the transfer of the skills, and scientific training itself.

While the studies conducted by Schraagen (1993) and Schunn and Anderson (1999) seem to point towards the influence of scientific training, there is a problem with interpreting it in that way. Both of these studies utilized highly similar groups of experts (all were experimental psychologists of some sort). The similarity of the domains of expertise might be confounding the influence of context on transfer. While the experiments were designed such that the task would be unfamiliar to one of the expert groups, it is likely that they were familiar enough to provide sufficient context cues to trigger the use of the appropriate scientific strategies. The novices, however, would not have had the cues to signal which strategies would be appropriate. If the studies had utilized more dissimilar experts this would not have been an issue. As it is, the transfer context remains a confounding factor.

Voss et al. (1986) provide some support for the transfer hypothesis. Their studies looking at the problem solving of novices and dissimilar experts found that the quality of the reasoning was dependent upon the match of the task to the expert's domain. They report chemists' reasoning being roughly equivalent to that of the novices when attempting to solve political science problems, while the political scientists exhibited a higher quality of reasoning on the same tasks. However, while the Schraagen and Schunn & Anderson studies utilized science-based domains that were overly similar, the Voss, et al. study utilized a non-science-based domain. This makes comparisons across the studies problematic.

The difficulty of transferring skills from one context to another has long been a recognized problem (e.g. Thorndike & Woodworth, 1901; Singley & Anderson, 1989; Detterman, 1992). Singley and Anderson propose that transfer can only occur between two situations for identical elements. This transfer then depends on how the elements were encoded. If researchers only use scientific reasoning skills when faced with problems in their domain, it is possible that they will be coded in a manner that is specific to that context. It would therefore be unlikely that they would use the strategies when those context cues were absent. So in this situation, the

superficial elements of the unfamiliar task mask the relevance of using scientific skills. Unless there is a more abstract understanding of these skills (more general encoding), they will fail to be used in other scientific reasoning contexts.

Much like the experiments of Schraagen (1993) and Schunn and Anderson (1999), this study was designed to explore the differences in scientific reasoning in novices, task experts (at experimentation in general) and domain experts (in the problem domain). The core difference is that instead of using a single task, there are two isomorphic discovery tasks from different scientific domains. Each of the expert groups is a domain expert in one of the two tasks, making each group task experts for both and domain experts for only one. This design allows us to look specifically at the influence of domain familiarity on the transfer of scientific reasoning. Should domain familiarity play a key role in the transfer of the scientific reasoning skills, we would expect the skills to only manifest themselves when the experts are within their natural domain. When working in the unfamiliar domain, their performance would be more like that of the novices. If, however, they are able to recognize the deeper scientific structure of the unfamiliar problem, then their performance should be better than that of the novices and qualitatively equivalent to that seen in their natural domain.

Methods

Participants were recruited from two major universities: 33 undergraduates, 16 from one university, and 17 from the other. The graduate samples were each drawn from a different university (due to enrollment). 11 biology graduate students and 17 industrial/organizational psychology students were recruited, all had completed at least two years of study. Participants were paid for their participation.

Participants were asked to complete two isomorphic experiment-based exploration tasks: they were to determine how each of six task relevant independent variables (IV) affects the outcome of the dependent variable (DV). For example, the biology task had participants design experiments to determine how each variable (water temperature, turbidity, dissolved oxygen, pH, fecal and phosphorus contents) influenced the growth of a certain bacteria that was responsible for the development of open sores on fish. The industrial psychology task had participants design experiments to determine the role of manager characteristics (e.g. technical, critical reasoning, writing skills) on the objectivity of employee appraisals.

Experiments were designed and conducted in a computer simulated laboratory where participants were able to manipulate each IV in question; run and view experiments; take notes on hypotheses, experiments, and outcomes; as well as assign conclusions as to the influences of the IVs on the DV. No data analysis or graphing tools were provided. Each task was self-paced, allowing participants to conduct as many experiments as they needed in order to draw their conclusions. The simulated experiment lab was built with Java™ allowing the recording of all user actions for playback[1].

The task domains were selected based on graduate enrollment in the domains across two universities. Experts in the domains (instructors and researchers) were recruited during the design of the tasks to maximize the external validity of the tasks. The tasks, from microbiology and industrial/organization psychology, were based on experiments found in the literature. The IVs' qualitative effects on the DVs were maintained for 4 of the 6 variables. Two IVs from each task were modified so that they would produce anomalous results. One was anomalous from a theoretical perspective (TA), which was merely an inversion of the qualitative trend. The second was data anomalous (DA) in that a 20% subset of the variable's range produced extreme values. The anomalous variables were added to test the sensitivity of the participants to data anomalies (e.g. Chinn & Brewer, 1993).

Results

The critical comparisons in this study are two orthogonal comparisons: the experts vs. novices (graduates and undergraduates), as well as between the two groups of experts. The expert analyses are between the experts within their domain and outside their domain (domain and task experts respectively). The In-Domain group consists of the biology graduates working on the biology task and the psychology graduates solving the psychology task. The Out-Domain group consists of the same participants, merely performing the opposite task. Unless otherwise noted, all tests are repeated measure ANOVAs with the two tasks as the repeated factor. Specific statistical measures are reported in table 1; significance is assumed at $p<0.05$. Task presentation was counter-balanced, and there were no significant effects of task order on any of the reported measures.

Experimental Design Measures

Participants spent on average 43 minutes on each of the two tasks. While the graduate students spent a little longer on each task (approx. 47 minutes) than the novices (approx. 39 minutes), the differences were nonsignificant. Likewise, the number of experiments designed and conducted in each task did not differ significantly either between the novices and the experts (37 and 47

[1] The experiment can be downloaded from the author's website http://simon.lrdc.pitt.edu/~harrison/

respectively) or within the In-Domain and Out-Domain groups (48 and 45 respectively).

The next design measure considered the breadth of the experimental space that the participants searched (Klahr & Dunbar, 1988). Each IV that they could manipulate had a fixed range, which were divided into five equally-sized bins. For each unique experiment that manipulated a given variable, the total number of bins visited was computed. The more complete the variable range covered, the more informative the results will be with respect to that variable on the whole.

As expected, the novices covered a significantly smaller range than the graduates (see figure 1). However, there were no differences between the In-Domain experts (e.g. biology graduate students performing the biology task, and vice versa) and the Out-Domain experts (e.g. psychology students performing the biology task).

Looking at the breadth of search for the two anomalous variables yielded similar findings. The novices searched a much narrower space for the theory anomalous and data anomalous variables. Likewise, there were no significant differences between the domain and task experts.

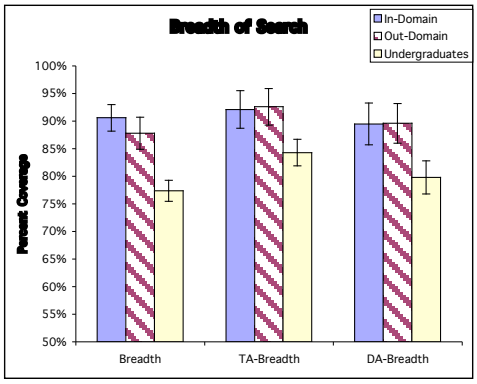


Figure 1. Breadth of experimental space search for all, theory anomalous (TA), and data anomalous (DA) variables.

Another design measure looked at the conservativeness of the experimental designs. With the lack of any data analysis tools, maximizing the interpretability of each simulated experiment was very important. Participants conducted experiments one at a time. The only way to draw conclusions was to compare outcomes of successive experiments. If the comparisons were confounded (multiple IV manipulations), accurate conclusions would be very difficult to draw. Two versions of a VOTAT (vary one thing at a time) score were computed for each task (Tschirgi, 1980). The local VOTAT was computed by averaging the number of variables manipulated in one experiment when compared to the immediately previous experiment. The global VOTAT was computed between the current experiment and the most similar previous experiment (regardless of when it occurred). Since there were no significant differences between the two measures across groups and tasks, the two were combined into an average composite. The novices consistently manipulated multiple variables per experiment, averaging 1.42 changes per experiment, which was significantly more than the graduate sample's 1.23. There were no differences between the experts within or outside their domains.

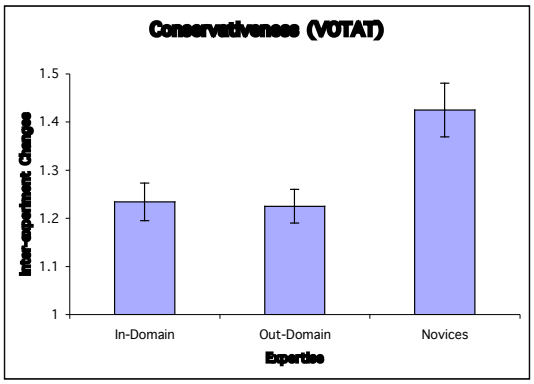


Figure 2. Average number of variables manipulated per experiment. One change at a time yields optimal interpretability in this scenario.

Note-taking

It is hard to argue against the importance of meta-processing skills in scientific reasoning. For this study, we chose to look at the note taking behavior of the participants. This is analogous to practice of scientists keeping a detailed lab notebook (Dunbar, 1997). The first measure is merely one of length, how much note taking is going on. Novices took very few notes (if any), which is in stark contrast to the experts who took significantly longer notes. For the experts, there were no differences in note-taking length when in or out of their natural domain.

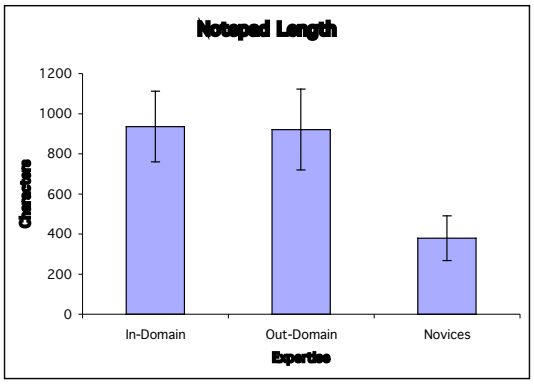


Figure 3. Length of notes (in characters) taken during experiment. Significant difference between novices and experts, none between the experts.

Knowing how much note taking was occurring lead us next to ask how they were utilized. A simple three-category scheme was developed a priori. Notes could be categorized as non-existent (including uninterpretable and

irrelevant notes), effects tracking (documenting variable values and experiment outcomes, effectively duplicating the provided experiment log), or hypothesis tracking (documenting hypotheses, suspected relationships, etc.). As can be seen in figure 4, all groups did an equal amount of effects tracking. The significant differences were in the amount of hypothesis tracking that the experts did.

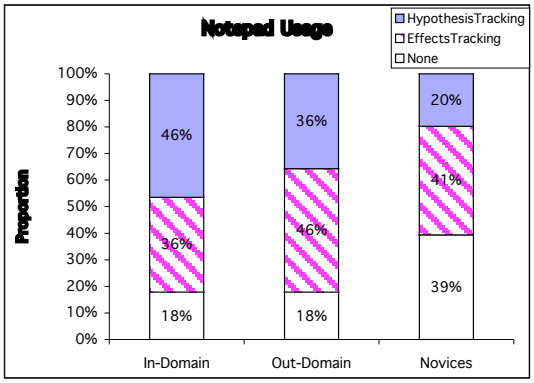


Figure 4. Type of notepad usage. Significant difference between novices and experts, $\chi^2(2,N=61)=5.9$, $p<0.05$. No significant differences between expert groups.

Data Interpretation

The final set of measures where designed to assess the participants' ability to interpret the data and draw conclusions regarding the relationships between the independent variables and the outcome measure. Participants were asked to write out their conclusions for each variable including the qualitative trend, critical values, magnitude, and any "strange" properties. One point was awarded for each property that was correct. Averaging across each of the variables yielded an overall interpretation accuracy score. Making it conditional on the breadth of the experiment space that was searched further refined each variable's score. For instance, any conclusions about critical values would be erroneous if they had not explored the variable range within which the values occurred. The left most graphs in figure 5 show both the raw and the conditional overall interpretation accuracy scores. Once more the novices scored lower than the experts with no significant differences between experts in or out of their domains. The differences between the raw and conditional scores were nonsignificant; participants did not appear to be drawing conclusions beyond what was possible given the data collected.

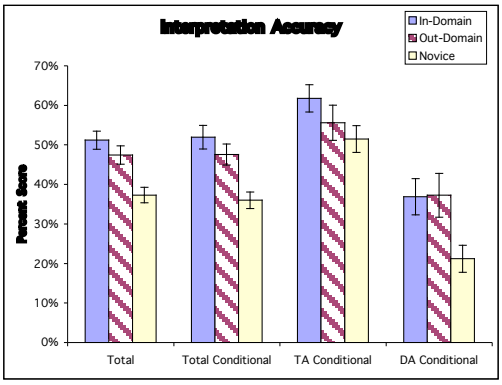


Figure 5. Interpretation accuracy scores, conditional on breadth of search.

The conditional accuracy scores for the anomalous variables were also examined (right half of figure 5). For the theoretically anomalous (TA) variables there were no differences among the three groups. The difference between In- and Out-Domain participants for the TA variable would likely be magnified if established domain experts had been used instead of the graduate students. There was a significant difference between novices and experts for the data anomalous (DA) variables, but this is directly attributable to the differences in the breadth of search for the data anomalous variables (see figure 1).

IQ Surrogates

Since this study had to be conducted across multiple universities, one of the first concerns was general IQ differences between the sampled populations. Using SAT and GRE scores as IQ surrogates, simple ANOVAs were computed. There were no significant differences within the novice undergraduate samples across the two universities, allowing the collapsing of the two groups. There were significant differences between the graduate samples (one from each university) and the novices. The post-hoc LSD showed no differences between the two graduate samples. Figure 6 shows the average combined SAT/GRE scores of the three groups. Further more, regression models were run on all the key dependent measures to test for IQ effects. None of the models approached even marginal significance: higher IQ participants did not perform better than the lower IQ participants. This data suggests that the IQ confound cannot account for the expertise effects found.

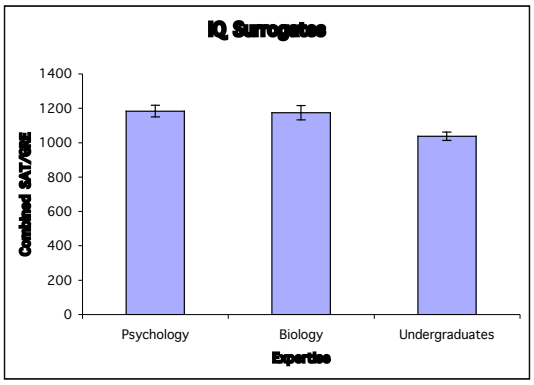


Figure 6. Combined SAT/GRE scores as IQ surrogates. No significant effects due to IQ.

Measure	Undergraduates v. Graduates (p-value)	In-Domain v. Out-Domain (p-value)
Breadth of search		
Overall	0.002	0.429
Theory	0.04	0.894
Data	0.02	0.987
VOTAT	0.02	0.86
Note length	0.001	0.93
Note usage†	0.05	0.875
Data interpretation		
Overall	0.00	0.234
Theory	0.41	0.301
Data	0.03	0.952
IQ surrogates††	0.000	0.34

Table 1. Significance of the planned comparisons. All measures are repeated measure ANOVAs with the tasks as the repeated factor, except for † Note usage (Chi-squared) and the †† IQ surrogate (standard ANOVA).

Discussion

Across the board expert groups performed significantly better than the novices, both inside and out of their domains. Additionally, there were no significant differences between the experts in or out of their natural domains. The findings mirror those of Schunn and Anderson (1999) and Schraagen (1993), even with the use of graduate students as opposed to established and practicing experts. Regardless of the domain, experts used the same general strategies in solving the discovery problems. Had the Out-Domain performance been significantly different from the In-Domain performance, bringing it closer to that of the novices, we could conclude that domain familiarity was influencing the transfer of the scientific reasoning skills.

The only major divergence from the former studies was the lack of a difference between the experts when it came time to draw conclusions from the data collected. While data interpretation per se has nothing to do with a particular domain, both Schraagen (1993) and Schunn and Anderson (1999) found qualitative differences between their task and domain experts. Both expert groups in this study performed equally poorly (but still significantly better than the novices). However, this could have been due to the fact that these tasks produced more data than those in other studies and lacked any analysis tools, such as graphs or tables (a common criticism leveled by the graduate students during the debriefing questionnaire).

Another difference in this study was the inclusion of anomalous variables (both theoretical and data based). Given that anomalies often draw the attention of scientists (Dunbar, 1997; Chinn & Brewer, 1993), it was interesting to see how sensitive the two expert groups were to them. One might expect that upon detection of an anomaly, that participants would search that section of the experimental space more thoroughly. Unfortunately, this is not seen for any of the expert groups (see figure 1). There are no reliably significant differences between the search patterns in terms of breadth or VOTAT (data not shown). This is not to say that the experts did not notice the anomalies; they just didn't exploit them in their experiments, in contrast to the findings of Dunbar (1997) and Tricket, et al (2001). This difference is likely due to the use of graduate students as opposed to established researchers.

Summary

Much like the studies of Schunn and Anderson (1999) and Schraagen (1993), we were interested in exploring the generality of scientific reasoning skills. Both studies concluded that while the quality of the reasoning was domain specific, the processes used were general. As was seen in these studies there are strong differences between the experimental skills exhibited by novices and the experts, or in this case, the developing experts. These differences cannot be attributed to a simple variable such as differences in intelligence. What's more is that the two expert groups behaved qualitatively the same whether they are working in their familiar domain or in a novel one. They were able to transfer the appropriate skills based on the deeper scientific structure and were not negatively influenced by the unfamiliar domain. The bidirectional transfer across a wider context difference is more evident in this study because of the utilization of genuinely dissimilar target domains, as opposed to the psychological domains used by Schunn & Anderson and Schraagen or the chemistry and (non-scientific) political-science domains used in the Voss, et al (1986) studies.

Acknowledgments

I would like to thank Lelyn Saner and Sasha Palmquist for their comments on earlier drafts of this work.

References

Chinn, C., & Brewer, W. (1993). Factors that influence how people respond to anomalous data. Proceedings of the 15th Annual Conference of the Cognitive Science Society, USA, 318-323

Chinn, C. A., & Malhotra, B. A. (1999). Epistemologically authentic scientific reasoning. In Crowley, K., Schunn, C.D., & Okada, T (Eds.), Designing for Science:Implications from Professional, Instructional, and Everyday Science. Mahwah, NJ: Erlbaum.

Detterman, D. K. (1992). The case for the prosecution: Transfer as an epiphenomenon. In Detterman, D. K., & Sternberg, R. J. (Eds), Transfer on Trial. Norwood, NJ.: Ablex Publishing.

Dunbar, K. (1993). Concept of Discovery in a Scientific Domain. Cognitive Science, 17, 397-434.

Dunbar, K. (1997) How scientists think: On-line creativity and conceptual change in science. In Ward, T., Smith, S., & Vaid, J. (Eds), Creative thought: An investigation of conceptual structures and processes. Washington D.C.: American Psychological Association.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-48.

Klahr, D., Dunbar, K., & Fay, A. L. (1991). Designing experiments to test 'bad' hypotheses. In J. Shrager & P. Langley (Eds.), Computational models of discovery and theory formation (pp. 355-401). Sam Mateo, CA: Morgan-Kaufman.

Klahr, D. & Fay, A. L., Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. Cognitive Psychology, 25, 111-146.

Kuhn, D., Schauble, L., & Garcia-Mila, M., (1992). Cross-domain development of scientific reasoning. Cognition and Instruction, 9, 285-327.

Schraagen, J. M. (1993). How experts solve a novel problem in experimental design. Cognitive Science, 17, 285-309.

Schunn, C. D., & Anderson, J. R. (1999) Acquiring expertise in science: explorations of what, when, and how. In Crowley, K., Schunn, C.D., & Okada, T (Eds.), Designing for Science:Implications from Professional, Instructional, and Everyday Science. Mahwah, NJ: Erlbaum.

Singley, M. K., & Anderson, J. R. (1989). The transfer of cognitive skill. Cambridge, MA: Harvard Press.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in mental function upon the efficiency of other functions. The Psychological Review, (8), 3, 247-261.

Trickett, S., Trafton, G., Schunn, C., & Harrison, A. (2001). "That's odd!" How scientists respond to anomalous data. Cognitive Science Society Conference, 2001.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. Child Development, 51, 1-10.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1986). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. Cognition and Instruction, 3, 269-302.

Zajchowski, R., & Martin, J. (1993). Differences in the problem solving of stronger and weaker novices in physics: knowledge, strategies, or knowledge structure? Journal of Research in Science Teaching 30, 459-470.

Learning from collaborative problem solving: An analysis of three hypothesized mechanisms

Robert G.M. Hausmann (bobhaus@pitt.edu), Michelene T.H. Chi (chi@pitt.edu), and Marguerite Roy (mar982@pitt.edu)

Department of Psychology and the Learning Research and Development Center
University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260

Abstract

It has been well established that collaborative learning is more effective in producing learning gains than individuals working alone. The present study investigates three potential mechanisms responsible for learning from collaborative problem solving: other-directed explaining, co-construction, and self-directed explaining. College undergraduates were trained to criterion on the first four chapters of a popular physics textbook. They were then asked to collaboratively solve three physics problems. Preliminary evidence suggests that other-directed explaining was effective in half of the cases, whereas co-construction led to proportionally more generated knowledge. Self-directed explaining was particularly effective for the individual generating the solution; however, there was only a modest gain for the partner who listened to the explanations. The relative impact of these three mechanisms is compared.

Introduction

Collaboration is a ubiquitous part of life and can be found in scientists' laboratories, the business world, the military, and the classroom. Given its usefulness in the real world, peer collaboration has become an important instructional intervention. The literature evaluating the effectiveness of peer collaboration has generally produced positive results (Dillenbourg, Baker, Blaye, & O'Malley, 1995); however, it is far from being an educational panacea (Barron, 2003). The open question remains, "Why is collaboration effective?" Prior research implicates three potential mechanisms responsible for learning during collaboration: other-directed explaining (Ploetzner, Dillenbourg, Praier, & Traum, 1999; Roscoe, 2003), co-construction (Damon, 1984; Rafal, 1996), and self-directed explaining (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, DeLeeuw, Chiu, & LaVancher, 1994). The goal of the present study is to investigate the relative contributions of all three mechanisms to individual learning.

The first mechanism, other-directed explaining, occurs when one peer instructs or explains to another partner how to solve a problem. Other-directed explaining may benefit only the speaker, but not the listener because the speaker is the one actively engaged in conveying the to-be-learned material (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Webb, Troper, & Fall, 1995). However, because both partners have opportunities to explain to their partners, it is conceivable that other-directed explaining is a mechanism that accounts for learning during collaboration.

The second hypothesized mechanism for successful learning from collaborative problem solving is co-construction. Co-construction is defined as the joint construction of knowledge. The process of constructing knowledge may proceed in a variety of ways, but the most natural is for peers to either elaborate or critically evaluate their partners' contributions.

Elaborative co-construction is the generation of knowledge by extending the ideas of one's partner (Tao & Gunstone, 1999) and has been shown to be an effective dialog pattern (Hogan, Nastasi, & Pressley, 1999; van Boxtel, van der Linden, & Kanselaar, 2000). Similarly, knowledge might be constructed through the critical discussion of ideas. Critical co-construction occurs when interacting peers critically evaluate each other's ideas. Support for this particular type of interaction comes from the literature on argumentation. Schwartz, Neuman, and Biezuner (2000) found dyads that successfully solved fraction problems were most likely to engage in argumentation.

Co-construction and other-directed explaining are likely candidates to explain the potential successes of learning from collaborative problem solving. However, there is another potentially overlooked mechanism, which is to learn from listening to someone self-explain (i.e., self-directed explaining). Learning from another person's explaining is analogous to learning from a worked-out example; however, in a collaborative problem-solving context, the source of the worked-out example is not a textbook, but a peer. When a dyad is faced with solving a problem, it is natural for one person to begin solving, while the other listens to the ensuing solution attempt (Shirouzu, Miyake, & Masukawa, 2002). The speaker may be talking out loud while solving a problem while her partner listens. This is a form of self-directed explaining. As has been shown in prior research, self-explaining is an effective learning strategy (Chi et al., 1989). What is unclear, however, is if a partner can benefit from listening to self-directed explaining. One goal of the present study is to provide initial evidence for the utility of self-directed explaining (or self-explaining with an audience).

The structure for the remainder of the paper is as follows. First, we will provide a brief description of the study, which

will then be followed by evidence for the three hypothesized mechanisms discussed above: other-directed explaining, co-construction, and self-directed explaining. The final section will compare the relative impact on learning from the three mechanisms.

Method

Participants

Students were recruited via advertisements placed in a university newspaper. The study used a between-subjects design with a total of ten undergraduate pairs (n=20) participating in the experimental group (i.e., Collaboration condition) and a total of ten (n=10) undergraduates in the control group (i.e., Text-only condition). Upon completion of the experiment, participants were paid for their time. To control for prior knowledge, eligible participants were required to have taken only one high school physics course.

Materials

The domain chosen for the present study was kinematics. Some of the topics covered were vector addition and subtraction, average and instantaneous velocity and acceleration, and an emphasis on Newton's second and third laws. The material was taken from the first five chapters of a popular physics textbook (Halliday & Resnick, 1981).

Measures Four mastery tests were developed to assess participants' understanding of each of the first four chapters. Participants were required to solve 80% of the problems correctly before advancing to the next chapter. After the first four chapters were learned to mastery, participants read the fifth chapter on force. The test administered after the participants had read the fifth chapter served as the pretest to the learning intervention.

The pretest consisted of three problems. The three problems were decomposed into a total of nine solution steps. Each solution step was then labeled with a concept from physics. For example, Problem 1 asks the individual to find the acceleration of two boxes in contact. To correctly solve the problem, the two boxes should be conceptualized as a *compound body*. To demonstrate an understanding of the compound body concept, the solver is required to sum the masses. The labeled solution steps will henceforth be called "concepts" (see Appendix for the mapping between problems, concepts, and the groups that were assigned to each problem). There were a total of nine different physics concepts across all three problems.

The text was available to the participants during the mastery tests, the pretest, and during the instructional phase, but it was not available during the posttest. The reason for making the text available during the pretest, but not the posttest, was to provide the most stringent test for learning, in the sense that we did not want our participants' inability to remember details of formulas to hinder their performance.

During the collaboration session, each pair was asked to solve three physics problems. For the present paper, only 18 of the 30 problems were analyzed.[1] Nine groups solved the first problem, which contained 4 different concepts; 5 groups solved the second problem, which contained 3 different concepts; and 4 groups solved the third problem, which was composed of 2 different concepts. Henceforth, there were 59 concepts assessed across the three problems and ten groups.

Finally, a posttest measured the amount of material learned during the instructional phase (administered M=5.0; SD=3.6 days after the collaboration session). The posttest contained three problems, which were isomorphic to the pretest and collaborative problems (i.e., the same nine concepts were tested on the posttest).

Procedure

All of the participants were asked to study each of the first four chapters individually in the way that they found natural. Participants solved the problems on the mastery tests either while they read the text or after they were finished. When the participants were confident in their solutions, they submitted their answers to the experimenter, who then immediately scored their performance. If the student correctly answered 80% or more of the questions, then he or she was permitted to advance to the next chapter. If the criterion was not met, then the student was shown which problems were incorrect and encouraged to read the text and correct the mistakes. This cycle of reading and testing continued for the first four chapters until criterion performance was met. On average, students spent a total of 6.5 hours to reach mastery of the four chapters.

The pretest was administered after the students read chapter five. Once they were finished with the pretest, an instructional phase was scheduled. For the Collaborative condition, the pretest was not scored immediately, so that the participants were not paired on the basis of their pretest scores. Dyads were formed under the constraints that they finished the background material relatively close in time, and they were the same gender.

During collaboration, the dyads solved three force problems. They were encouraged to use their partners as a resource and to work together to understand and solve the problems. The entire text (chapters 1-5) was available to the dyads during the collaboration session. The Text-only group solved the same problems, but did so individually with the text available. After the instructional phase, the posttest was individually administered.

The sessions were recorded (both audio and video) and later transcribed. The transcription was based on the audiotapes of the dialogues, using information from the video for interpretations when necessary.

[1] A subset was used because the performance data for the present study comes from a larger study in which all of the items were relevant.

Analyses and Results

Coding scheme

The first coding step was to segment the transcribed protocols. The segments were taken at the level of problem-solving episodes (i.e., several turns dedicated to a single concept). The boundaries of a problem-solving episode began with a proposed equation, and ended with either the final solution of the equation or the abandonment of a solution path altogether. Across all groups and problems, there were a total of 87 problem-solving episodes.

Episodes were then coded as other-directed explaining (ODE), self-directed explaining (SDE), or co-construction. Other-directed explaining occurred when a more-knowledgeable partner explained a concept to a less-knowledgeable partner. Pretest performance for each participant determined his or her knowledgeability status for each concept. Because each problem was composed of several concepts, the individual's status could change from one problem-solving episode to the next, depending on his or her pretest performance.

When the less-knowledgeable peer explained a concept during a problem-solving episode, either to a more-knowledgeable or equally knowledgeable partner, then the episode was coded as self-directed explaining. Again, pretest performance was used to determine knowledgeability.

Finally, when both partners were being generative in the conversation by adding significant and relevant contributions, the episode was coded as co-constructed. Co-constructed episodes were further decomposed into elaborative and critical co-construction, which will be defined shortly.

Once the problem-solving episodes were coded in terms of the conversational elements, the content (i.e., the physics concepts) was also coded. The content was then linked to the episode analysis, which allowed us to track the impact of dialog on posttest performance. For example, if a more-knowledgeable peer explains how to solve the compound body concept to her less-knowledgeable partner, then that episode was coded as "other-directed explaining *about* the compound body." To measure the learning effect of other-directed explaining on the listener, we then looked at her posttest performance on the compound body concept. Because some problems involved multiple solution attempts, only the final problem-solving episode was linked to the posttest concepts (N=59).

Collaborative problem solving resulted in learning gains

Did the individuals learn from the collaborative problem-solving session? To answer this question, we calculated gain scores for each individual, which controlled for pretest knowledge: $g = (post - pre)/(100\% - pre)$ (Crouch & Mazur, 2001). Thus, the gain scores reflect the increase (or decrease) in learning per concept, per person. Overall, there was an average net gain of 26% (while controlling for pretest knowledge), which was significantly different from zero (p=0.002).

Evidence that individuals learned from collaborative problem solving can also be found in the analysis of the control (Text-only) group. The gain from pre- to posttest for the Collaboration group was significantly greater than zero, while the gain for the Text-only group was not ($F(1,9)=0.756$, p=0.41) even through the two groups did not differ at pretest.[2] This suggests that the learning gains are due to the activities the dyads engaged in during collaborative problem solving, which is presented in the next three sections.

Other-directed explaining during collaborative problem solving

As stated in the Coding Scheme section, both the content and episode were coded together to give us a sense of the impact of other-directed explaining on learning. Table 1 contains an excerpt of one student explaining her reasoning to another. The example begins with Beth asking Abby[3] to elaborate on a previous line of reasoning. There are two features to note in this example. First, Abby's style is definitely instructional. Her intent is to explain, as clearly as she can, how to solve the problem (see Appendix problem 1.ii.). Second, Beth does not contribute much to the conversation, but merely indicates that she is attending to Abby's explanation with her use of continuers.

Table 1: Example of other-directed explaining

Beth:	So like 14 newtons would be the net force acting on B?
Abby:	No, this-the overall force is ten,
Beth:	Mm-hmm.
Abby:	but if you split it, if-if-both of the blocks, as we know, are accelerating at two meters per second. If they're in contact then they have to be accelerating at the same, rate.
Beth:	Mm-hmm.
Abby	And, because, by Newton's second law F-F [pause] equals mass times acceleration. And we know the acceleration,
Beth:	Mm-hmm.
Abby:	of each block and we know the mass of each block. So you can calculate the force-the force of each block. Or the force acting on each block.

Of the 59 final problem-solving episodes, there were a total of 11 other-directed explaining episodes (11/59=19%). On posttest, the listener (i.e., the less-knowledgeable peer) correctly used 5 concepts that they had previously used incorrectly on pretest. The data are summarized in the left segment of Figure 1. The black bars represent the

[2] The data for the Text-only and Collaboration comparisons is taken from the full pretest and posttest (see footnote 1).

[3] All names are pseudonyms.

percentage of the corpus dedicated to a particular dialog type, while the grey bars represent the gain scores of the listeners (controlling for pretest knowledge). While a gain of 5 concepts is encouraging, especially given that the text was unavailable during the posttest, the probability of learning from listening to other-directed explaining is low (5/11=45%). This is not entirely surprising, given the finding that receiving elaborated help does not always lead to learning gains (Webb, 1989).

It is also informative to measure the performance of the more-knowledgeable speaker. Figure 2 suggests that the speaker (ODE), who knew the concepts on pretest, maintained 82% of her knowledge by correctly demonstrating her knowledge of the concepts on the posttest (see white bars in Fig. 2).

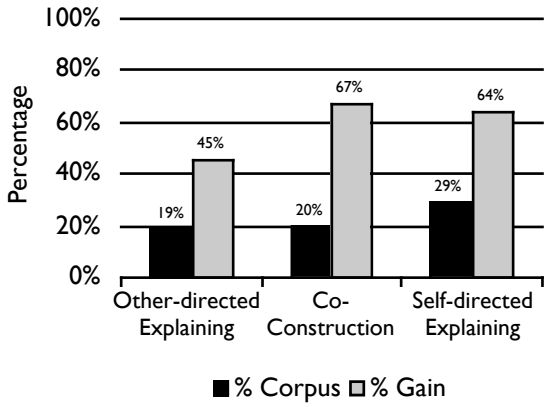


Figure 1: Gain scores associated with the absolute frequency for each dialog pattern.

Partners co-construct answers during collaboration

As stated in the introduction, co-construction is a hypothesized mechanism that has been proposed to account for learning from collaborative problem solving. A problem-solving episode was coded as co-construction when both partners were actively constructing new knowledge. The co-constructed solutions were further categorized as elaborative and critical co-construction. Elaborative co-construction was defined as one partner adding a significant contribution to the discourse that develops another person's idea. Here is an example of elaborative co-construction (Problem 2.ii.):

Table 3: Example of elaborative co-construction

Ron: It's the weight of the crate, which is ten, times gravity right? [R writes 10kg(9.8m/s^2)="] So it's, 98N, plus the five [R writes "98N+5"]-oh no, cause we don't know what it is yet, really. Well I mean, it's-
Ben: Mg-Mg is the force exerted by the block, on the Earth.
Ron: Mg, that's,
Ben: Weight.
Ron: Mass times gravity, right?
Ben: Mm-hmm.

Episodes were coded as critical co-construction when they contained conflicts between the two partners (Druyan, 2001). The difference between partners' solutions led to a discussion where each attempted to convince the other how to solve the problem.

The following protocol excerpt is taken from Jill and Sara solving the compound body problem (see Appendix, Problem 1.i.). The question is difficult because it requires the solvers to represent the blocks as a single body; neither student demonstrated an understanding of this on the pretest. Sara believes the question implies that the acceleration should be found separately for each block, but Jill makes a case for the compound body. The conflict is between treating the blocks separately or jointly. Here is their argument:

Table 2: Example of critical co-construction

Sara: Yeah. It's just-it didn't say? I thought it said each of them. [Reads: **Find the acceleration of the blocks.**] To me that says find the acceleration of each block. You know like, since they're two different kilograms.
Jill: It's going to be, the same though.
Jill: Because like, if we, go like this [pushes a book and pencil], and I do this, they're both moving at the same acceleration.
Sara: [Talks to Experimenter: 4 turns]
Sara: Because if you-well you can get a different acceleration by breaking it up though.
Jill: Oh wait. You know what? The acceleration will be the same for both of them. Acceleration is the same for both of them. Force acting on block B, is different from force acting on block A.
Sara: Ok. Because their mass, is different.
Jill: Yeah. Because-yeah.

The frequency of co-constructive episodes is summarized in Figure 1 (see the middle bars). Two results are of particular interest. First, the amount of co-construction is similar to the frequency of the other-directed explaining episodes. More importantly, however, is the proportion of co-construction episodes that led to learning. Of the 12 episodes where the solution was jointly constructed, 8 of them led to the correct application on posttest (8/12=67%). Although co-construction was a relatively rare conversational pattern (12/59=20%), the reported frequency replicates prior estimates from a different domain (McGregor & Chi, 2002). Furthermore, the knowledge produced during collaboration was useful to both the individuals, which suggests the viability of group-to-individual transfer. That co-construction lead to a high proportion of learned concepts further supports the constructivist perspective that being active, as opposed to merely listening to a didactic explanation, is important for learning (Chi et al., 2001; Webb et al., 1995).

Co-construction was further decomposed into elaborative and critical co-constructive episodes. Of the 12 instances of co-construction, 5 were elaborative (5/12=42%) and 7 were

critical (7/12=58%). Elaborative led to a gain of 3 concepts (3/5=60%), whereas critical co-construction led to the correct application of 5 concepts on posttest (5/7=71%). Because of the small numbers, it is difficult to tell if elaborative or critical co-construction was more effective in subsequent learning. Follow-up research needs to be done to gain a better understanding of what drives learning from co-construction.

Learning occurs from self-directed explaining for speakers and listeners

Prior research has shown that good students spontaneously self-explain while learning from worked-out examples (Chi et al., 1989). Subsequent research has shown that prompting students to self-explain can lead to learning gains, above and beyond those who spontaneously self-explain (Chi et al., 1994).

Figure 1 suggests that self-directed explaining (SDE) also operates in a collaborative problem-solving context. The frequency of self-directed explaining is high relative to the other conversational patterns (i.e., other-directed explaining and co-construction). We observed 17 episodes of self-directed explaining, which accounts for 29% of the corpus. In terms of the average gain, self-directed explaining episodes lead to a 64% increase (see Fig. 1).

The effects of self-directed explaining can be further differentiated into the gain observed by the speaker and listener. In the present context, the listener is also trying to learn the material; therefore, she has a stake in the problem-solving process. Instead of being a passive recipient, the collaborative partner listens to and could potentially monitor the ensuing self-explanation.

As expected, the gain was proportionally high for the speaker (71%; see Fig. 2). While self-explaining is effective for the explainer, the question becomes, does listening to a self-explanation benefit the listener? The answer to this question seems to be mixed. To a certain extent, listening to another person self-explain can produce learning. Specifically, there was a net gain of 5 concepts for the listeners (5/17=29%; see Fig. 2). Therefore, it appears that observing reasoning in action (i.e., being the listener) is about as effective as listening to other-directed explaining (ODE). Further coding is needed to gain a better understanding of what the listener is doing while listening to a partner self-explain. One might hypothesize that the listeners benefit only when engaged in a constructive activity, which has received some empirical support (Webb et al., 1995).

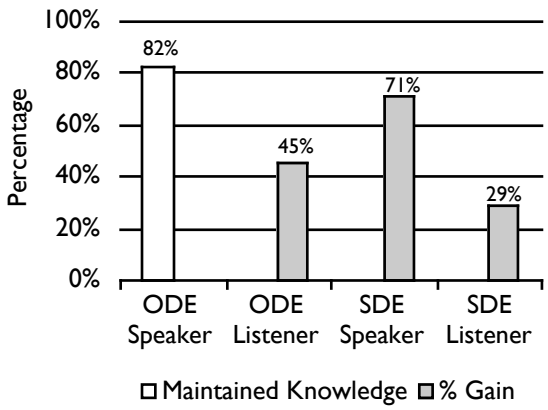


Figure 2: Gain scores as a function of dialog pattern and dominant speaker or listener.

Discussion

The primary goal of the present study was to demonstrate that several different mechanisms contribute to learning from collaborative problem solving. These three mechanisms were other-directed explaining, co-construction, and self-directed explaining. All of these mechanisms were associated with learning, but did so to different degrees. In terms of the overall proportion, self-directed explaining produced the strongest learning gains, with the caveat that the learning gains were greatest for the speakers. Other-directed explaining also lead to learning gains for the listener, but only to a limited extent. Several explanations given by the speaker during other-directed explaining did not translate into increased problem solving behaviors on posttest. Finally, co-construction, although relatively infrequent, led to increased problem-solving performance. Two-thirds of the co-constructed concepts were correctly used on the posttest.

A secondary goal of the present study was to demonstrate that multiple mechanisms operate within dyads. That is, one group may engage in other-directed explaining on a problem that one person understands (whereas the other does not). Then on the next problem, the same dyad may have to co-construct the solution because each individual has a different solution, and they must resolve their differences. Most research on collaborative problem solving measures the influence of one mechanism on learning in isolation of other potential explanations. The results from this study suggest that the pattern of communication is largely shaped by the background knowledge of the participants.

Finally, we attempted to show that self-explaining can take place in a collaborative context. While effective for the speaker, there was marginal utility for the listener. The effect was strongest when the speaker was engaging a partner with low pretest knowledge, but this effect needs to be substantiated by further research. The results from this study agree well with the idea that being constructive while solving problems leads to better learning and understanding.

Acknowledgements

Funding for this research is provided by the National Science Foundation, Grant Number NSF (LIS): 9720359, to the Center for Interdisciplinary Research on Constructive Learning Environments (CIRCLE, http://www.pitt.edu/~circle). The authors are indebted to Mark U. McGregor and Randi A. Engle for their data collection assistance, Stacy Setterberg for transcription, and Rod D. Roscoe, Kwangsu Cho, and 4 anonymous reviewers for their critical comments on an earlier draft.

References

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307-359.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M. T. H., DeLeeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Association of Physics Teachers, 69*(9), 970-977.

Damon, W. (1984). Peer education: The untapped potential. *Journal of Applied Developmental Psychology, 5*, 331-343.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. J. (1995). The evolution of research on collaborative learning. In P. Reinman & H. Spada (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science*. New York: Elsevier Science Inc.

Druyan, S. (2001). A comparison of four types of cognitive conflict and their effect on cognitive development. *International Journal of Behavioral Development, 25*(3), 226-236.

Halliday, D., & Resnick, R. (1981). *Fundamentals of physics*. New York: Wiley.

Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and Instruction, 17*(4), 379-432.

McGregor, M., & Chi, M. T. H. (2002). Collaborative interactions: The process of joint production and individual reuse of novel ideas. In W. D. Gray & C. D. Schunn (Eds.), *24nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawerence Erlbaum.

Ploetzner, R., Dillenbourg, P., Praier, M., & Traum, D. (1999). Learning by explaining to oneself and to others. In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 103-121). Oxford: Elsevier.

Rafal, C. T. (1996). From co-construction to takeovers: Science talk in a group of four girls. *The Journal of the Learning Sciences, 5*, 279-293.

Roscoe, R. D. (2003). *Learning from self-directed versus other-directed explaining.* Paper presented at the 84th Annual Meeting of the American Education Research Association, Chicago, IL.

Schwartz, B. B., Neuman, Y., & Biezuner, S. (2000). Two wrongs may make a right...if they argue! *Cognition and Instruction, 18*(4), 461-494.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science, 26*, 469-501.

Tao, P.-K., & Gunstone, R. F. (1999). Conceptual change in science through collaborative learning at the computer. *International Journal of Science Education, 21*(1), 39-57.

van Boxtel, C., van der Linden, J., & Kanselaar, G. (2000). Collaborative leraning tasks and the elaboration of conceptual knowledge. *Learning and Instruction, 10*, 311-330.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13*, 21-39.

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87*(3), 406-423.

Appendix

Collaboration Problem	*Concepts*	*Groups*
1. Two blocks A and B are in contact with each other on a smooth floor. A force of 10N is applied to the blocks as shown in the figure. Masses of the blocks are 2 Kg and 3 Kg respectively. (i) Find acceleration of the blocks. (ii) Find net force acting on block B. (iii) Find force exerted by block B on block A.	N2Law CB N2Law N3Law	1,2 3,4 5,6 7,9 10
2. A person pushes a crate on a smooth floor. He is applying force at an angle q with the horizontal as in the figure. If the mass of the crate is 10 Kg, magnitude of the force is 5N and q=30 degrees, what will be the acceleration of the crate?	VD W T	5,6 8,9 10
3. Block A is attached to a string which is tied to a wall. The block is resting on a smooth plane inclined at an angle q with the horizontal as shown in the figure. Mass of the block is MA. What is the tension in the sting?	VD N2Law	1,2 4,7

Note. N2Law=Newton' second law; N3Law=Newton's third Law; CB=compound body; VD=vector decomposition; W=weight; T=tension.

The Adaptability of Language Specific Verb Lexicalization Biases

Catherine Havasi (havasi@mit.edu)
Department of Computer Science, 77 Massachusetts Ave
Cambridge MA 02139 USA

Jesse Snedeker (snedeker@wjh.harvard.edu)
Department of Psychology, 33 Kirkland St.
Cambridge, MA 02138 USA

Abstract

Languages vary in how they encode motion events. For example, English motion verbs often encode the manner of the motion while Spanish motion verbs encode the path. Efficient verb learning has been argued to involve the acquisition of language specific lexicalization biases. When given a novel verb paired with a single motion event, English speakers interpret it as a manner verb, Spanish speakers as a path verb. The present study examines the nature and plasticity of this lexicalization bias. Do lexicalization biases result in a permanent alteration of the semantic interface? Or are these biases continually shaped through our experiences with word learning? English-speaking adults were taught 12 motion verbs. The composition of the set of verbs was varied from 100% manner to 100% path with 3 levels in between. Lexicalization biases were monitored by testing verb extension after the first ambiguous exemplar of each verb. We replicate the finding that English speakers have an initial manner bias. However, we find that this bias changes over time in response to the input: Participants who learned path verbs developed a path lexicalization bias. Experiment 2 replicates this result with a different syntactic frame.

Introduction

Children's early lexicons are curiously lopsided. Across a variety of linguistic environments, nouns dominate early vocabularies, while verbs are initially scarce (for a review see Gentner & Boroditsky, 2001). There are a number of explanations for initial noun dominance, which are by no means mutually exclusive. Verbs differ from nouns in the frequency with which they occur in isolation or in salient positions within the utterance and they also differ in the types of concepts that they encode and the types of entities that they pick out in the world. All of these factors have been argued to play a role in early noun dominance (Gleitman, 1990; Tardif, Shatz & Naigles, 1997; Caselli, Casadio & Bates, 1999; Snedeker & Gleitman, 2004).

Gentner and her colleagues have argued that nouns are prominent in the early lexicon because they typically denote physical objects which can be individuated (and presumably conceptualized) on the basis of the child's perceptual experience of the world (1982; Gentner & Boroditsky, 2001). Verbs, they argue are more difficult for novice language learners because perception does not package events into stable individuals. Instead languages decide how to conflate the conceptual components of events into lexical items. This results in greater cross-linguistic differences in the meanings of verbs than in the meanings of nouns. To learn verbs, they argue, children must first discover how their language chooses to package events. To the extent that lexicalization patterns are systematic within a language, children should be able to draw generalizations from known instances, developing lexicalization biases which allow the pace of verb learning to accelerate.

The parade case for systematic cross-linguistic variation in lexicalization is the conflation patterns that occur in verbs of motion (Talmy, 1975). A motion event consists of a thing that is moving (the figure), the location it is moving relative to (the ground), the manner in which it is moving and the path along which it moves. All languages have ways of expressing these elements, but how they do so varies. 'Manner' languages, such as English and Mandarin, typically pack manner of motion into the verb, leaving path for an optional prepositional phrase ("He ran into the store"). In contrast, 'path' languages, such as Spanish and Greek and typically encode path in the verb and fob off manner on an optional gerund ("Él entró en la tienda corriendo"). In English, path verbs are relatively scarce (Gutiérrez, 2001; Talmy 1975). This cross-linguistic difference in verb use shows up in distributional analyses and production studies with both children and adults (Aske, 1989; Jackendoff, 1990; Berman & Slobin, 1994).

This systematic difference in lexicalization patterns also results in differences in how the speakers of manner and path languages learn new motion verbs, consistent with the predictions of Gentner's relational relativity hypothesis (1982). When confronted with a novel verb used to describe a single motion event, English speaking adults and seven year olds will extend the word to other events with the same manner of motion but not to events that have the same path (Naigles & Terrazas, 1998; Hohenstein & Naigles, 2000). In contrast Spanish speaking adults and seven-year olds extend the verb to events that have the same path but not the same manner. Thus each group has developed a lexicalization bias that is consistent with the primary verb lexicalization pattern in their language.

While verb lexicalization biases clearly exist, we know little about how they might develop or how they are mentally represented. One intriguing hypothesis comes from the literature on the development of the shape bias in noun learning. Smith and colleagues have argued that the shape bias is a generalization based on the words that the child has previously acquired (Smith, Jones, Landau, Gershkoff-Stowe & Samuelson, 2002). Children, they

claim, are initially unbiased learners who acquire their first nouns by patiently waiting for the situational concomitants of word use to tease apart the many alternate hypotheses about how a word might be extended. In this way, they manage to acquire a sizeable number of nouns, many of which are well-organized by shape. They argue that the shape bias is simply a second-order generalization of these known words. This account is supported by two lines of evidence. First, in the studies of Smith and her colleagues children fail to show a systematic shape bias until they have acquired a substantial number of nouns (but see Waxman, 1999). Second, toddlers who are trained on shape-based categories develop a shape bias and show accelerated acquisition of nouns, while those who are trained on substance-based categories or given an unsystematic training set do not.

While Gentner makes no specific proposal for how verb lexicalization biases could be acquired, the mechanism laid out by Smith seems consistent with the relational relativity hypothesis. Children learn a number of verbs that follow a language specific lexicalization pattern and then form the expectation that verbs in the same semantic field will be extended in a parallel fashion. But what does this expectation consist of? There are at least three possible explanations for how cross-linguistic differences in word learning biases could be instantiated.

First, children's word learning experiences could permanently alter their conceptual systems resulting in a change in the repertoire of possible concepts or in their relative salience or stability. This is an unlikely explanation for the manner-path lexicalization bias. Speakers of the two languages show similar behavior on nonlinguistic memory and categorization tasks, suggesting that verb conflation patterns do not affect the accessibility of manner or path concepts (Papafragou, Massey & Gleitman, 2002). Furthermore, since both languages have ways of expressing both manner and path, the linguistic evidence itself radically limits the degree to which conceptual alteration can be invoked. Thus we have to look to changes in the semantic interface which maps between linguistic forms and concepts.[1]

Second, lexicalization biases could be permanent alterations in the semantic interface. The mappings between linguistic forms and concepts could be altered so that certain conceptual dimensions are unavailable as candidates for verb meanings, although they might be used in nonlinguistic tasks or even as meanings for other terms. This mechanism would be the semantic parallel of Werker's functional reorganization hypothesis for phonological development (1995). Finally, lexicalization biases may be more plastic mappings between linguistic forms and concepts which can modified as the child gains access to new information sources. Critically, if lexicalization biases are generalizations on the basis of known words, then they may be dynamically updated as children learn a larger and more varied set of verbs. To date there has been little work on the plasticity of the semantic interface between words and concepts. While several studies have examined the effects of age of acquisition on semantic processing of a second language, the results vary with the measures and contrasts that are studied (compare e.g., Weber-Fox and Neville, 1996 with Munnich, 2002). To the best of our knowledge no one has looked at the plasticity of lexicalization biases. Experiment 1 explores the possibility that manner lexicalization bias for motion verbs continues to be malleable into adulthood and is shaped by the set of verbs that a person learns. Experiment 2 replicates this finding when the verb is presented in a different syntactic context.

Experiment 1

Each participant learned twelve new motion verbs. For each novel verb, participants (1) saw a single ambiguous scene with a salient path and manner of motion, (2) were tested to determine their initial interpretation of the verb, (3) saw five additional instances of the new verb which disambiguated its meaning (e.g. five scenes with same manner but a novel path), and finally, (4) were tested again to ensure that they had learned the novel verb.

Critically, the proportion of path and manner verbs was varied across participants. Some participants learned only manner verbs, some learned only path verbs, and others received different proportions of both types. We predicted that our adult participants would have little difficulty learning either the manner or the path verbs. The critical measure was the participants' responses to the initial test trials, which followed the first ambiguous scene. Because a single verb-scene pair is consistent with either a manner or path interpretation, responses to this test sequence reveal the participants' verb lexicalization bias. Since our participants are English speakers, we expect that they will begin with an initial bias to interpret the novel verbs as encoding manner of motion. However, if these estimates of prior probability are updated in response to the verbs that the participant has learned, then responses on the initial test trials should change in response to novel verbs. Thus we predict that over the course of the experiment participants who learn path verbs will develop a path bias, while those who learn manner verbs will retain the manner bias. [2]

Methods

Participants 56 adult native English speakers participated in this study. Since our goal was to determine how previously learned verbs influence the interpretation of future verbs, we eliminated all participants who failed to

[1] We follow Jackendoff's (2002) suggestion that language specific semantics are most parsimoniously described as an interface between linguistic forms and conceptual representations, rather than as a separate level of representation, but nothing in our argument rests upon this distinction.

[2] Similar issues have been explored in artificial category learning studies. Critically, Kersten, Goldstone & Schaffert (1998) found that adults who learned manner event categories were more likely to focus on the manner feature of an ambiguous category. They used simple animated events with bug-like agents and no sentential context. These stimuli did not appear to engage participants' prior lexicalization biases: English speakers showed a strong initial path bias. Unlike the present work, the study did not examine generalization after a single ambiguous exemplar.

learn 5 or more of the verbs after viewing the disambiguating scenes. Sixteen participants were excluded for this reason.

Stimuli Participants saw short video clips of motion events. Each event depicted an actor moving in a salient manner and in a salient path with respect to some reference object (e.g., a woman walking tip-toe behind a large sign). Twelve manner and twelve path concepts were selected as target verb meanings. Some concepts corresponded to English verbs, some to English prepositions, and some had no monomorphemic English equivalent. The path verb meanings were: *around, between, down, up, in front of, along, in, diagonal to, over, across,* and *behind*. The manner verb meanings were: *crab-walk, crawl, twirl, flap-walk, hop on 1 foot, hop on 2 feet, march, run, skip, stoop-walk, tiptoe,* and *walk*.

Participants were presented with a block of questions and videos for each of 12 novel nonce verbs. Each block was identical in layout and was made up of 4 phases: an initial ambiguous scene, an initial bias test, training and a final test phase. An example test block for a manner verb is shown in Table 1 and an example for a path verb is shown in Table 2.

Table 1: Sample block for a novel manner verb.

Target Concept: Crab-Walk	Manner	Path
Ambiguous Scene	Crab-walk	Out
Initial Test: Manner	Crab-walk	Behind
Initial Test: Path	Skip	Out
Training One	Crab-walk	Front
Training Two	Crab-walk	In
Training Three	Crab-walk	Between
Training Four	Crab-walk	Across
Training Five	Crab-walk	Diagonal to
Final Test: Path	March	Out
Final Test: Manner	Crab-Walk	Between

Table 2: Sample block for novel path verb.

Target Concept: Out	Manner	Path
Ambiguous Scene	Crab-walk	Out
Initial Test: Manner	Crab-walk	Behind
Initial Test: Path	Skip	Out
Training One	Hop 2 Feet	Out
Training Two	Walk	Out
Training Three	Run	Out
Training Four	Stoop-walk	Out
Training Five	Dance	Out
Final Test: Path	March	Out
Final Test: Manner	Crab-Walk	Between

In the ambiguous scene, the participant saw a written sentence containing a new nonce verb (e.g. *"She is going to torg out the door."*) and a video which illustrates the sentence (e.g., a woman crab walking out of the door). The initial test consists of two clips which are presented sequentially. The participant is asked if clip is an instance of the new verb ("Is this torging?"). One test clip matches the manner of the ambiguous event but not the path; the other matches the path but not the manner. During the training phase, participants are presented with 5 video clips which disambiguate the meaning of the word. If the verb is being taught as a path verb, then all 5 clips will show the same path as the ambiguous training clip but vary in their manner. If the word is being taught as a manner verb, the reverse will be true. The final test parallels the initial test; one video matches the path of the ambiguous clip, the other matches it in manner. This test allows us to determine if the participant has succeeded in learning the verb.

Each manner verb was arbitrarily paired with a path verb. The paired verbs shared the same initial scene and the same test scenes (see Tables 1 & 2). Pairing the items in this way allowed us to examine how participants with different verb learning experiences responded to identical stimuli. The disambiguating videos were different for each member of a pair. Subjects were assigned to one of five conditions which differed in the proportion of the novel verbs that encoded path (0, .25, .50, .75 or 1). The 12 verb pairs were randomly ordered and half of the participants in each condition were tested with the blocks in reverse order.

Procedure Stimuli were presented on a computer which using custom software. The participants were told that they would be watching videos that would teach them new words and answering the questions about these words.

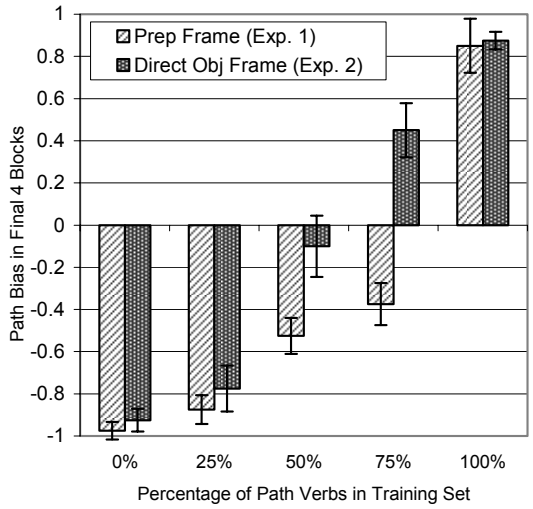


Figure 1: Path Bias on Final Blocks

Results

Responses to the final test questions were used to exclude participants who failed to learn the verbs. Our analyses focused entirely on participants responses in the initial test. To explore how bias changed over time we examined responses to the first four verb blocks and the last four verb blocks. The participants' responses were converted to path bias scores by taking the proportion of blocks where the subject extended the word to the path match and subtracting the proportion of blocks where they extended the word to

the manner match. This number would equal -1 for a perfect manner bias and 1 for a perfect path bias.

The ANOVA of the first four blocks revealed that these English speaking subjects entered the study with a strong manner bias (M = -.58, F (1,40) = 145.69, p < .001). 76% of the participants responded yes to a manner video while only 5% responded yes to a path video. However, there were also differences between the training condition demonstrating that the verbs in the training set were already beginning to shape participants interpretations of the initial ambiguous scenes (F(4,40) = 12.97, p < .001). This effect was driven by participants in the 100% Path Verbs Condition who had no systematic bias in these early blocks (M = .08).

In the final four blocks of the experiment, the initial bias trials are clearly shaped by the set of verbs that the participant has learned (see Figure 1, F(4,40) = 45.14, p < .001). Participants in the 100% Path Condition have developed a strong, consistent path bias in their interpretation of new verbs (M = .85). Those in the 0% and 25% Path Conditions show an equally clear manner bias (M = .98, M = .88). In the 50% and 75% Path Conditions participants flout the input, continuing to show a mild preference to interpret the new word as a manner verb.

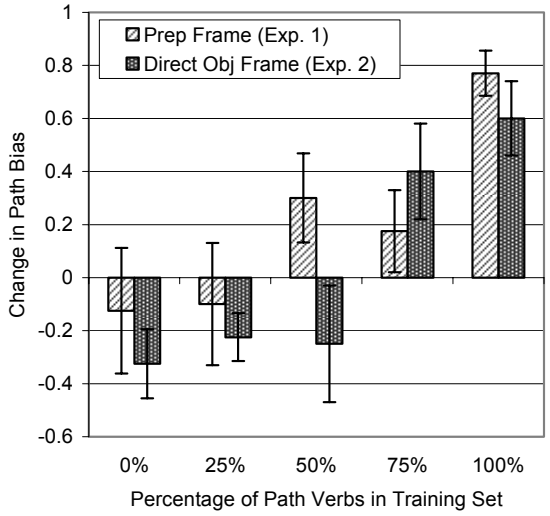


Figure 2: Change in Path Bias between the First 4 and Last 4 blocks.

To further explore how lexicalization biases changed over the course of the experiment, we directly compared the results of the first and final blocks. In Figure 2 this is graphed as the change in path bias. There was a substantial increase in path bias in the final trials (F(1,40) = 7.45, p < .009) and a reliable interaction between the time in the experiment and the Training Condition (F(4,40) = 4.75, p < .003). Participants in the 100%, 75% and 50% Conditions showed an increase in path bias, while participants in the 25% and 0% retained or strengthened their manner bias.

Experiment 2

In Experiment 1, the nonce verbs appeared with prepositional phrase arguments. In English, this syntactic frame is used more frequently with manner verbs, although it can be used colloquially with path verbs as well (e.g. *"She ran around the tree."* or *"She circled around the tree.")*. In English path verbs are often used in simple transitive frames (*"She circled the tree."*) and this usage is typically considered more proper. Naigles and Terrazas (1998) found that both English and Spanish speakers were more likely to interpret novel verbs as encoding manner when syntactic frames with semantically rich prepositions were used. If the participants assume that each component of the motion event is encoded in only one word of the sentence, they may be reluctant to conflate path in the verb when it is already marked in the preposition. In Experiment 2 we used simple transitive sentences to explore whether the syntactic context influenced participants' initial lexicalization biases or the changes in these biases in response to newly learned words.

Methods

Participants 52 English-speaking adults participated in this study. Responses from 2 participants were excluded because they failed to learn 5 or more target verbs.

Stimuli, Procedure and Coding Participants were tested on the same verbs sets as before (100%, 75%, 50%, 25% or 0% Path Verbs). The procedure was identical to Experiment 1 except that verbs were introduced with simple transitive frames. Thus *"He torged down the stairs"* became "*He torged the stairs."*

Results

When the novel words were presented in transitive frames, subjects showed only a weak bias in the first four verb blocks (M = -.14, F(4,40) = 3.56, p = .067) which may reflect the early effect of Training Condition on path bias (F(4,40) = 6.56, p < .001). Participants in the 0% Path Condition show a manner bias (M = -.60) while those in the 100% Condition have a path bias (M = .28). By the final four verb blocks, participants' initial interpretations of the new nonce verbs are essentially categorical and closely match the set of words that they have learned (F(4,40) = 60.17, p < .001). There is a reliable shift in bias between the first and final blocks, which interacts with Training Condition (F(4,40) = 21.56, p < .000). Participants who learned all path verbs showed an increase in path bias (M = .60) while participants who learned all manner verbs showed an increase in manner bias (M = -.33). A direct comparison of Experiments 1 and 2 demonstrates that syntactic frame influenced participants' performance on both the final and first 4 blocks. These differences are limited to the 50% and 75% Path conditions, conditions where subjects are given weak evidence for a lexicalization pattern which differs from the dominant pattern in English.

General Discussion

These experiments 1) replicate Naigles & Terrazas' findings that adult speakers of English have a bias to assume that novel verbs encode the manner of motion rather than the path; 2) demonstrate that this lexicalization bias remains

plastic into adulthood; and 3) demonstrate that this lexicalization bias can be influenced by the words that a person learns. English speaking adults who were taught new motion verbs developed lexicalization biases that matched the verbs in their training set. This pattern was observed regardless of whether the ground nominal was presented as the object of a preposition (consistent with a manner verb) or as the direct object of the verb (consistent with a path verb). In the remainder of the discussion we re-examine Gentner's relational relativity hypothesis in the light of these findings and discuss the development of verb lexicalization biases.

Reexamining Relational Relativity

Gentner's relational relativity hypothesis proposes that the acquisition of verbs is delayed because children must discover how their language packages and categorizes events (1982). Naigles and colleagues have extended this argument by suggesting that efficient verb learning requires the acquisition of language-specific semantic patterns, which we have called lexicalization biases (Naigles & Terrazas, 1998). The current study demonstrates that learners retain a remarkable degree of plasticity in lexicalization biases. They can not only learn verbs which violate the lexicalization pattern of their language; they can actually change their lexicalization biases to reflect the patterns in newly acquired words. These results suggest that verb lexicalization biases are not the result of permanent alterations in conceptual structure or unalterable changes in the semantic interface. Instead these biases appear to be plastic generalizations based on the words the learner has acquired. In essence a lexicalization bias results from a change in the prior probability of a class of hypotheses based on the prior success of hypotheses from that class.

The flexibility of these biases raises questions about the role that they play in potentiating early verb learning. If stable lexicalization patterns are required to repackage relational components into the individuated events, then why do adults have little difficulty in rapidly and spontaneously recombining these components? [3] We would argue that this ability is essential for learning the range of verbs that exist within any one language. While the variability of cross-linguistic encoding of events has received much attention, there is considerable variation in verbs that can be used to describe a single event within a language (Gleitman, 1990). For example, we can refer to an event in which a girl kicks a ball to her mother as *giving, passing, kicking, rolling, receiving, moving, crossing* or *contacting* depending on the components of meaning that we wish to include in the verb or the perspective that we are taking on the scene. In light of such variability, rigid lexicalization biases are likely to be counterproductive. Within language variation in lexicalization also seriously limits the role that these biases can play in constraining word learning, and thus limits the explanatory potential of the relational relativity hypothesis.

If language-specific semantic mappings cannot eliminate the ambiguity inherent in events, then how do children ever become rapid and efficient verb learners? We believe that two factors are at play. First, children may improve in their ability to make use of cross-situational observation. Much of the work in early word learning has focused on what children are able to learn from a single word-scene pair. In the case of nouns it may be possible to make a meaningful conjecture about the meaning of a word on the basis of a single referent. Many of children's early nouns label artifacts and natural kinds. Concepts of these kinds are organized in taxonomic hierarchies, which have multiple levels (animal, mammal, dog, poodle) and categories which are mutually exclusive at a given level (Markman, 1989). This conceptual structure helps bridge the gap between reference and meaning. Once the observer has correctly picked out the referent of an artifact or natural-kind term, then its meaning can be limited to concepts on the path from the individual exemplar up to the top of the hierarchical tree. If there is a conceptually or perceptually privileged basic level, then a single referent might provide enough information to map the word to the correct node of that tree (Rosch et al., 1976; Markman, 1989).

But in the case of verbs, cross-situational observation may be essential. There is little evidence that the concepts encoded in verbs form complex taxonomic hierarchies. Instead most observers have argued that states and events are grouped into semantic fields which are organized as a cross-cutting lattice of concepts rather than as mutually exclusive categories (Talmy, 1985; Behrend, 1995). Identifying a single referent event merely identifies a point in this multi-dimensional conceptual space but it does not tell the observer which dimension(s) of the event are encoded in the verb. Multiple exemplars, however, can be used to rule out the relevance of some dimensions and provide convergent evidence for the importance of others.

Second, children's verb learning also benefits from their increasingly sophisticated representations of the utterances in which new verbs appear. Initially children must learn the meanings of new words by observing the nonlinguistic contexts in which those words are used. This initial information source provides ample support for noun learning but provides inadequate information about meanings of many verbs (see e.g., Snedeker & Gleitman, 2004). More sophisticated learners can use the known words which co-occur with novel verbs to focus their attention on the relevant events. As the child gains knowledge about the syntax of her language, the structural environments in which the verb occurs can also provide increasingly fine-grained information about its meaning

[3] Perhaps that the manner-path distinction is the wrong place to search for stable lexicalization biases, since both manner and path languages have verbs of each type (Ashe, 1989). We challenge the reader to come up with a better example of a systematic lexicalization pattern which applies to a large number of verbs. The explanatory potential of the relational relativity hypothesis depends on the prevalence of this predictable variation.

Examining the Early Development of Verb Lexicalization Biases

Clearly the present experiments cannot rule out the possibility that young language learners have relatively inflexible language-specific lexicalization biases which serve as sharp constraints on children's hypotheses about verb meaning. Furthermore, even plastic and probabilistic biases could provide useful guidance for verb learning. Understanding the role of lexicalization biases in early verb learning clearly requires studying how these biases develop in young children. The limited information that we have about the development of the manner-path bias suggests that this bias may emerge quite late: Hohenstein and Naigles (2000) have found the 3 year old English speakers and Spanish speakers show no differences in their extension of novel motion verbs (both populations prefer to extend the words to events with the same manner of motion). The obvious explanation is that children this age simply lack the ability to derive lexicalization biases from the words they learn. But this seems unlikely in light of Smith and colleagues' finding (2002) that children under two can develop a shape-bias after learning just a handful of exemplars. The alternative explanation is that the verbs that 3-year-olds know simply don't support this generalization. In elicited production tasks English speaking adults show a clear preference for manner, however, speakers of path languages like Greek and Spanish often produce equal numbers of manner and path verbs when describing motion events (Papafragou et al., 2002; Naigles et al., 1998), suggesting that young Spanish speakers may have little evidence for a path lexicalization bias. To determine whether young children can form verb lexicalization biases in response to clear category structure, we are currently testing three- and five-year old children in a modified version of Experiment 1. Our preliminary findings (N = 9) suggest that five-year-olds will rapidly form a bias for the dimension which has been relevant on previous trials. Children who are given six path verbs select path as the relevant dimension on 67% of all trials, while those who are given manner verbs do so only 25% of the time ($p < .05$).

Acknowledgments

This work grew out of conversations with Sourabh Niyogi and Bob Berwick and we are grateful for their ideas, inspiration and assistance. We also thank Melanie Goetz, Liz Sepulveda, Natan Cliffer, Rob Speer, Mahvash Malik and Sylvia Yuan, for their assistance with filming and testing. This project was funded by a grant from the NSF (IIS-0218852) to the second author and Bob Berwick.

References

Aske, J. (1989). Path predicates in English and Spanish: A closer look. *Proceedings of the 15th Annual Meeting of the Berkeley Linguistics Society*, 1-14. Berkeley, CA: BLS.

Behrend, D. A. (1995). Processes involved in the initial mapping of verb meanings. In M. Tomasello & W.E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs. (pp. 251-273).* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc

Berman, R. & Slobin, D., eds. (1994). *Relating events in narrative: A cross-linguistic developmental study*. Hillsdale, NJ: Erlbaum.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S.A. Kuczaj, II (Ed.), *Language development: Vol 2. language, thought, and culture*. Hillsdale, NJ: Erlbaum.

Gentner, D. & Boroditsky, L. (2001). Individuation, relativity and early word learning. In M. Bowerman and Levinson (Eds.), *Language acquisition and conceptual development*. England: Cambridge University Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3-55.

Hohenstein, J., & Naigles, L. (2000). *Preferential looking reveals language specific event similarity by Spanish- and English-speaking children.* Paper presented at the BU Conf. on Lang. Development, Nov 2000, Boston, MA.

Jackendoff, R. S. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford Univ. Press, NY.

Kersten, A.W., Goldstone, R.L., & Schaffert, A. (1998). Two competing attentional mechanisms in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 24, 1437-1458.

Levin, B. (1993). *English Verb Classes and Alternation*. Chicago, IL: U of Chicago Press.

Markman, E. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press, Bradford Books.

Naigles, L. & Terrazas, P. (1998). Motion-verb generalizations in English and Spanish: Influences of language and syntax. *Psychological Science*, *9,* 363-369.

Papafragou, A., Massey, C., Gleitman, L. (2002). Shake, rattle, 'n' roll: The representation of motion in language and cognition. *Cognition*. Vol 84(2): 189-219.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology, 8*, 382-439.

Schwartz, R.G., & Leonard, L.B. (1980). Words, objects, and actions in early lexical acquisition. *Papers and Reports in Child Language Development*, *19,* 29-36.

Smith, L.B., Jones S.S., Landau B., Gershkoff-Stowe L., & Samuelson L. (2002). Object name learning provides on-the-job training for attention. *Psych. Science, 13*, 13-19.

Snedeker, J. & Gleitman, L. (2004). Why it is hard to label our concepts. In Hall & Waxman (eds.), *Weaving a Lexicon.* Cambridge, MA: MIT Press.

Tardif, T., Shatz, M. & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs : A comparison of English, Italian, and Mandarin. *JCL, 24*, 535-565.

Talmy, L. (1975). Semantics and syntax of motion. In J. Kimball (Ed.), *Syntax and semantics* (Vol 4., pp 181-238). New York: Academic Press.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description*, Vol. 3, pp. 57–149. New York: Cambridge University Press.

Scopal ambiguity preferences in German negated clauses

Barbara Hemforth (barbara.hemforth@lpl.univ-aix.fr)
Laboratoire Parole et Langage, UMR 6057, Univ. de Aix en Provence
29, av. Robert Schuman, 13621 Aix en Provence, France

Lars Konieczny (lars@cognition.uni-freiburg.de)
Center for Cognitive Science, IIG, Univ. Freiburg
Friedrichstr. 50, 79098 Freiburg

Abstract

When following a negated matrix clause, adverbial clauses (ACs) like "because it was paid very well" in (1) can be interpreted as residing within the scope of the negation (1b), or outside of it (1).

(1) a. Peter did not quit his job because it was paid very well
b. Peter did not quit his job because it was paid very badly.

Depending on the scope of the negative, the interpretation differs dramatically: Whereas Peter did in fact not quit his job in (1a), he did so in (1b), but for yet unknown reasons. It has been shown for English (see Frazier & Clifton, 1996), that there is a preference to interpret the adverbial clause outside of the scope of the negation so that (1b) appears fairly odd. This observation challenges recency based processing principles, such as late closure, since the high attachment (to IP) appears to be preferred over low attachment (to VP) (see figure 1). In this paper, we will present evidence on German equivalents of (1a,b), varying the order of the negative and the verb (Experiment I), the context in which the ambiguity appears (Experiment II), and the position of the adverbial in relation to the clause boundary where the negation of the main verb is restricted or even retracted (Experiment III). None of these variations reduced the preference substantially. Only an explicit alternative cause reduced it but even this variation did not eliminate the difficulty of the inside scope interpretation. We will argue that incremental interpretation as well as immediate attribution of prosodic structure determine the interpretation of the adverbial clause.

Introduction

Whereas the scope of quantifiers has been the subject of substantial research (e.g., Ioup, 1975, Johnson-Laird, 1969; Kurtzman & MacDonald, 1996; for an extensive discussion see Frazier, 1999) this is much less the case for the scope of negations. In this paper, we will look at sentences like (1) where an adverbial clause can be interpreted as either being within the scope of the negation in the matrix clause or outside of it. Structurally, the adverbial clause has to be attached to the VP if it is interpreted as residing within the scope of the negation whereas it has to be attached to IP if it is interpreted outside the scope of the negation.

Locality based principles of syntactic attachment as they are assumed in most theories of human sentence processing (e.g., Frazier, 1978; Gibson, 1991) predict a preference to attach the adverbial clause to VP, and thus a preference to interpret it inside of the scope of the negation.

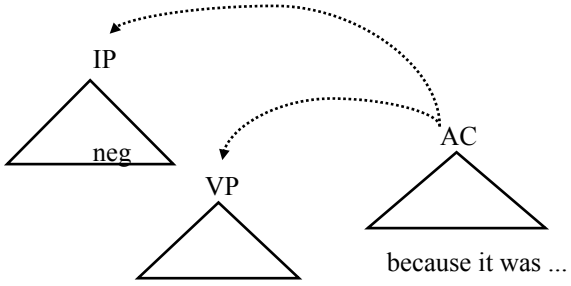


Figure 1

However, Frazier and Clifton (1996) found a clear preference for an interpretation of the AC outside of the negation scope. In their experiments, this preference does not show up in early stages of processing, but only in later off-line measures. In the framework of Construal Theory, the authors argue that only argument like or primary relations are attached to the phrase marker of the sentence immediately. Only primary relations are subject to syntactic attachment principles like Minimal Attachment or Recency. As a non-primary relation, the adverbial clause is only construed as part of the maximal projection of the preceding thematic domain (i.e. the IP). All kinds of factors (syntactic, semantic, pragmatic, or prosodic) jointly determine the final interpretation of the clause.

In this paper, we do not want to dispute the immediacy of the attachment or the interpretation of adverbial clauses. We are much more interested in the question of which factors are driving this final preference, and how general or universal it is. This means testing this preference in different constructions, in different environments, and across languages. Therefore, we have been looking at German versions of Frazier and Clifton's materials. In four experiments, we tried to find a way to override the interpretational preference, by putting the negation in focus position (Exp I), varying contextually given presuppositions (Exp II), and restricting the negation (Exp III). Since our main interest does not lie in the question of when these factors come into play, but only whether they do have an effect at all, in all of our experiments, we applied an off-line acceptability judgment task.

Experiment I

In Experiment I, we wanted to test whether the preference for high attachment of the adverbial clause which has

already been established for English can also be found in German. Additionally, we varied the position of the negation "nicht" (not). Since the clause final position in German is prosodically more dominant than clause internal positions, we assumed that an interpretation of the adverbial clause inside the scope of the negation might be more viable if it is clause final. As a control for lexical effects we included controls without negations.

Methods

Materials. Eighteen sentences were constructed, closely related to the materials from Frazier and Clifton (1996). Each sentence contained a main clause followed by a subclause beginning with "weil" (because). The comma preceding "weil" is obligatory in German for all possible interpretations. It cannot serve as a cue for the interpretation of the adverbial clause. In the main clause, a character was introduced as the subject in the default topic position. The pronoun in the because-clause was meant to refer to the subject of the main clause. This was the most plausible reading according to the intuitions of the experimenters and qualitative interviews after the experiment showed that subjects interpreted the pronouns exactly this way. Six versions of each sentence were constructed. In two, the negation preceded the clause-final participle (condition VP-internal, 2a,b). Two versions had the negation as the last word of the main clause (VP-final, 2c,d). Crossed with the position of the negation, the because-clause could either be plausibly interpreted as being inside of the scope of the negation or not. In order to control for potential plausibility differences in the adverbial clauses, we included two controls (2e,f) which only differed with respect to the adverbial clause, but did not have a negation in the matrix clause.

(2) a. neg. VP-internal, AC outside negation scope
Die Sekretärin hat nicht gekündigt, weil sie ein hohes Gehalt erhielt.
The secretary has not quit her job, because she got a high salary.
b. neg. VP-internal, AC in negation scope
Die Sekretärin hat nicht gekündigt, weil sie ein geringes Gehalt erhielt.
The secretary has not quit her job, because she got a low salary.
c. neg. VP-final, AC outside negation scope
Die Sekretärin kündigte nicht, weil sie ein hohes Gehalt erhielt.
The secretary did not quit her job, because she received a high salary.
d. neg. VP-final, AC in negation scope
Die Sekretärin kündigte nicht, weil sie ein geringes Gehalt erhielt.
The secretary did not quit her job, because she received a low salary.
e. no negation (control-a)
Die Sekretärin war unbeliebt, weil sie ein hohes Gehalt erhielt
The secretary wasn´t liked very much, because she received a high salary.
f. no negation (control-b)
Die Sekretärin war unbeliebt, weil sie ein geringes Gehalt erhielt.
The secretary wasn´t liked very much, because she received a low salary.

Six counterbalanced forms of the questionnaire were constructed. One sixth of the 18 experimental sentences appeared in each version in each form of the questionnaire, and across the six forms, each experimental sentence appeared once in each version. Each sentence was followed by a question concerning its acceptability. These 18 sentences were combined with 36 sentences of various forms varying in complexity (simple main clauses, simple embeddings and doubly nested embeddings) and plausbility (from fully plausible to fairly implausible according to the intuitions of the experimenters). One randomization was made of each form.

Participants. Eighteen participants, mostly undergraduate students from the University of Freiburg, judged the acceptability of sentences presented in a printed questionnaire. They either received course credits or they were paid for their participation. All subjects´ native language was German, none of them was bilingual.

Procedure. The rating technique used was magnitude estimation (ME, see Bard et al., 1996). Participants were instructed to provide a numeric score that indicates how much better (or worse) the current sentence was compared to a given reference sentence (Example: If the reference sentence was given the reference score of 100, judging a target sentence five times better would result in 500, judging it five times worse in 20). Judging the acceptability ratio of a sentence in this way results in a scale which is open-ended on both sides. It has been demonstrated that ME is therefore more sensitive than fixed rating-scales, especially for scores that would approach the ends of such rating scales (Bard, et al., 1996).

Each questionnaire began with a written instruction where the subject was made familiar with the task based on two examples. After that subjects were presented with a reference sentence for which they had to provide a reference score. All following sentences had to be judged in relation to the reference sentence.

Results

Individual judgments were individually standardized and logarithmized. Table 1 contains mean judgments in the six conditions.

Table 1: Acceptability judgments Experiment I

	neg. VP-internal	neg. VP-final	no negation
AC within scope of negation	-.64	-.79	.24
AC outside of scope of negation	.28	.31	.34

Judgments were submitted to a two-factorial MANOVA including the factors "negation" (VP-internal, VP-final, no negation) and "attachment" (IP, VP). There was a main effect of "negation" ($F1(2, 34) = 11.25$, $p < 0.001$; $F2(2,34) = 10.89$, $p < 0.001$), resulting from the fact that on average, sentences without negations were judged more acceptable than those containing negations (VP-internal: - 0,178, VP-final: - 0,240; no negation: 0,287). The main effect of "attachment" as well as the interaction "negation"*"attachment" reached significance as well ("attachment": $F1(1,17) = 51.29$, $p<0.001$; $F2(1,17) = 26.00$, $p < 0.001$; "negation"*"attachment": $F1(2,34) = 7.26$, $p <0.01$; $F2(2,34) = 8.64$, $p < 0.01$). Planned comparisons show that low attachment was judged less acceptable in sentences containing a VP-internal negation ($F1(1,17) = 26,47$, $p < 0.001$, $F2(1,17) = 16.20$, $p < 0.01$) as well as in sentences containing a VP-final negation ($F1(1,17) = 38.31$; $p < 0.001$; $F2(1,17) = 29.29$, $p < 0.001$), whereas there was no difference in acceptability between the control sentences ($F1(1,17) < 1$, ns; $F2(1,17) < 1$, ns).

Discussion

In Experiment I, we clearly replicated the findings Frazier & Clifton (1996) report for English. In German as in English it is harder to interpret the adverbial clause as residing within the scope of the negation. Varying the position of the negation, however, did not exert an influence on the acceptability of this interpretation. The clause-final focus on the negation is obviously not sufficient to render it more viable. In the second experiment we tried to put the negated sentences in contexts that were supposed to bias for either of the two interpretations.

Experiment II

For Experiment II, we constructed four different contexts for each sentence: A neutral context, leaving open whether or not the proposition stated in the following matrix clause holds or not (3), a context biasing an external scope reading (4), and two contexts biasing an internal scope reading, with one context explicitly stating that the proposition stated in the matrix clause should not be negated (5) and another one presupposing that e.g. the secretary actually quit her job (6).

(3) Neutral context
Jeder hat die Neuigkeiten über die Sekretärin gehört.
Die Sekretärin hat nicht gekündigt, weil sie ein (a) hohes / (b) geringes Gehalt erhielt.
Everybody heard the news about the secretary.
The secretary did not quit her job, because she got a (a) high / (b) low salary.
(4) Contextual bias: AC outside neg. scope
Jeder hat sich gefragt, ob die Sekretärin gekündigt hat.
Die Sekretärin hat nicht gekündigt, weil sie ein (a) hohes / (b) geringes Gehalt erhielt.
Everybody wondered, whether the secretary had quit her job.
The secretary did not quit her job, because she got a (a) high / (b) low salary.
(5) Contextual bias: AC within neg. scope (I)
Jeder hat gehört, dass die Sekretärin gekündigt hat.
Die Sekretärin hat nicht gekündigt, weil sie ein (a) hohes / (b) geringes Gehalt erhielt.
Everybody heard that the secretary had quit her job.
The secretary did not quit her job, because she got a (a) high / (b) low salary.
(6) Contextual bias: AC within neg. scope (II)
Jeder hat sich gefragt, warum die Sekretärin gekündigt hat.
Die Sekretärin hat nicht gekündigt, weil sie ein (a) hohes / (b) geringes Gehalt erhielt.
Everybody wondered, why the secretary had quit her job..
The secretary did not quit her job, because she got a (a) high / (b) low salary.

Methods

We applied the same technique as in the previous experiment. 24 subjects, all native German speakers and all students from the University of Freiburg participated in the experiment. Fillers varied along the full range from fully grammatical to ungrammatical, as well as from fully plausible to highly implausible.

Results

There was a reliable main effect of Scope: sentences for which the adverbial clause had to be interpreted inside the scope of the negation were judged less acceptable than their counterparts ($F1(1,23) = 26.25$, $p < 0.001$; $F2(1,15) = 53.4$, $p < 0.001$).

Table II: Acceptability judgments Experiment II

	Neutral Context	Scope external	Scope internal I	Scope internal II
AC inside scope of negation	-.54	-.38	-.16	-.44
AC outside scope of negation	.36	.26	.39	.52

No main effect of context was found ($F1(3, 69) < 1$, ns; $F2(3,45) = 1.23$, ns). Although explicitly stating the fact that

the proposition described in the matrix clause should not be negated rendered the sentences slightly more acceptable numerically, this did not result in an interaction between Context and Scope (F1(3,69) = 1,12, ns; F2(3,45) = 1.61, ns).

Discussion

In all contexts, even in those biasing the inside scope reading, interpreting the adverbial clause outside of the scope of the negation was judged more acceptable. Before we will discuss why this may be the case, we will extend the phenomenon by including a temporal adverb as a possible domain for the negation.

Experiment III

In Experiment III we presented the same ambiguous constructions as in the earlier experiments. However, we added two conditions with a temporal adverb between the negation and the adverbial clause. This has the effect that now, the adverbial clause in (7d) can easily be interpreted outside of the scope of the negation which is restricted by the temporal adverb (The secretary quit the job before yesterday because of the low salary.). The only viable interpretation for (7c), however is an interpretation of the adverbial clause inside the scope of the negation (The secretary actually quit the job yesterday, but not because she got a high salary, but e.g., because she didn't have another job offer before.). The prediction is that we should find the same preference for the outside scope reading in these constructions.

(7) a. AC outside of neg. scope, no restriction before clause boundary
Die Sekretärin kündigte nicht, weil sie ein hohes Gehalt erhielt.
The secretary did not quit her job, because she got a high salary.
b. AC within neg. scope, no restriction before clause boundary
Die Sekretärin kündigte nicht, weil sie ein geringes Gehalt erhielt.
The secretary did not quit her job, because she got a low salary.
c. AC inside of neg. scope, restriction before clause boundary
Die Sekretärin kündigte nicht erst gestern, weil sie ein hohes Gehalt erhielt.
The secretary did not quit her job only yesterday, because she got a high salary.
d. AC outside of neg. scope, restriction before clause boundary
Die Sekretärin kündigte nicht erst gestern, weil sie ein geringes Gehalt erhielt.
The secretary did not quit her job only yesterday, because she got a low salary.

Methods

As in the other experiments, subjects judged the acceptability of the sentences in relation to a reference sentence (Magnitude Estimations). The sixteen experimental sentences were randomly mixed with 48 filler sentences of varying acceptability (some ungrammatical, some highly implausible). Sixteen native German subjects, all students of the University of Freiburg, participated in the experiment.

Results

The sentences where we included a temporal modifier as a possible domain for the negation were generally more complex and judged less acceptable than the shorter versions 5F1(1,15) = 17.69, $p < 0.001$; F2(1,15) = 10.07; $p < 0.01$). More importantly though, the interpretation of the adverbial clause outside of the scope of the negation was clearly preferred in both, the shorter and the longer version (F1(1,15) = 45.10, $p < 0.001$; F2(1,15) = 156.00, $p < 0.001$). The difference between the inside and the outside scope reading was, however, somewhat stronger for sentences without a temporal adverb as indicated by a reliable interaction between the two experimental factors (F1(1,15) = 5.71, $p < 0.05$; F2(1,15) = 7.51; $p < 0.05$).

Table 3: Acceptability judgments Experiment II

	Restriction before clause boundary	No restriction before clause b.
AC within scope of negation	-.54	-.24
AC outside of scope of negation	-.05	.85

Discussion

Although the sentences including a temporal modifier were somewhat less acceptable than the shorter versions, they showed the same preference pattern. Interpretation of the adverbial clause inside the scope of the negation is always far less acceptable than its outside scope interpretation. The interaction between the domain of the scope of the negation and the presence of a temporal modifier can be explained by the fact that the sentences including a temporal adverb were generally semantically more complex. Assuming a preference for incremental interpretation (Konieczny, Hemforth, Scheepers, & Strube, 1997; Crocker, 1995), the negation in these conditions has to be revised (First the secretary did not quit her job, then she did, but not only yesterday.). This local revision is obviously less costly than a revision between clause boundaries as in (7b), but it may have reduced the difference between the final interpretation of (7c) and (7d).

Experiment IV

In our fourth experiment, we presented a continuation of the sentence, explicitly providing the alternative cause (8a).

(8) a. Die Sekretaerin hat nicht gekuendigt, weil sie ein geringes Gehalt erhielt, sondern weil sie ihre Arbeit langweilig fand.
The secretary did not quit her job because she got a low salary bur because she found her work boring.
b. Die Sekretaerin hat nicht gekuendigt, weil sie ein hohes Gehalt erhielt, obwohl sie ihre Arbeit langweilig fand.
The secretary did not quit her job because she got a high salary although she found her job boring.

In this Experiment we found, that although the difference is actually strongly reduced, interpreting the "because"-clause inside the scope of the negation still causes some difficulties (mean ME score: 0.09 for 8a, +0.45 for 8b, $F1(1,15) = 5.83$, $p < 0.05$; $F2(1,15) = 6.51$; $p < 0.05$).

General Discussion

A strong preference to interpret the adverbial outside of the negative was established, which turned out to be very stable across experiments. It is independent of the ordering of the verb and the negative (Experiment I).

It shows up even in very strong contexts biasing the inside scope reading where the secretary actually quit her job (Experiment II), and the same pattern can be established in constructions, where the negation can be interpreted as restricted by a temporal adverb (Experiment III). Finally, Experiment IV shows that the preference for interpreting the adverbial clause outside the scope of the negation is reduced but not eliminated by a continuation that provides an explicit alternative to the negated because clause.

Obviously, the interpretation of the adverbial clause inside of the scope of the negation is very hard even in a context biasing for this reading. The question is why this is the case. One possibility is that the short texts in (5) and (6) are still semantically incomplete, since we now know what is not the reason for the secretary quitting her job but not what the reason for doing so actually is. Since it may be assumed that negative information is not as well presented as positive information (Legrenz, Girotto, & Johnson-Laird, 2003)), mental models for texts like these may be insufficiently specified und thus less acceptable. The data presented for Experiment IV, on the other hand, suggest that this aspect actually plays a major role. However, it does not seem to tell the whole story, since even an explicitly given alternative does not render the inside scope reading as acceptable as the outside scope reading.

An further possibility lies in the interaction of semantic and prosodic information. Ronat (1984) presents the Prosodic Binding hypothesis for French, roughly stating that a prosodic boundary delimits the scope of a quantifier or a wh-phrase (see also the Scope Correspondence Principle (SPC) suggested by Hirotani, 2003, for Japanese). Note, that this is actually a principle of grammar, meant to constrain the interpretational domain of quantifiers.

Assuming that even during silent reading, a prosodic structure of a sentence is constructed (Implicit Prosody Hypothesis; Fodor, 1998), this may play a role in reading as well. Since there is a high probability for a prosodic boundary between the matrix clause and the adverbial clause, the AC cannot be interpreted inside the scope of the negation. The inside scope reading of the adverbial clause is actually only viable with a very marked prosodic contour. At least intuitively (as stated by several native German and English informants), the break before the adverbial clause in this marked prosodic contour is strongly reduced. Interestingly, this seems to be true even though there still has to be a comma in the German clauses. So the comma by itself cannot be the major factor. The remaining difficulty in interpreting "complete" models may thus result from the interaction of semantic and prosodic constraints. This will, however, be a question to be answered in future research (Bradley, Fodor, Fernandez, Hemforth, & Pynte, in prep).

Acknowledgments

This research has partly been sponsored by the German Research Foundation (He 2310/3). We want to thank Regine Becher, Christian Mönke, and Daniel Umber for their assistance in constructing the materials and running the experiments. We also want to thank the two anonymous reviewers for their helpful comments.

References

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72(1),* 32-68.

Bradley, D., Fodor, J., Fernandez, E., Hemforth, B., & Pynte, J.(in prep). The role of prosody in reading: A cross-linguistic study of French and English.

Crocker, M.(1995). Computational Psycholinguistics : An Inter-Disciplinary Approach to the Study of Language, Dordrecht, NL: Kluwer Academic Press.

Fodor, J. D. 1998. Learning to parse? Journal of Psycholinguistic Research, 27, 2, 285-319.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* Doctoral dissertation, University of Connecticut.

Frazier, L., &Clifton, C. (1996) *Construal*. Cambridge, MA: The MIT Press.

Frazier, L. (1999). *On sentence interpretation*. Kluwer Academic Press.

Gibson, E. (1991). *A computational theory of human linguistic processing:* Memory limitations and processing breakdown. Doctoral dissertation, Carnegie Mellon University.

Hirotani, M. (2003) "Prosodic Boundary and Prosodic Contour: Interpreting Wh-scope in Japanese". Paper Presented at the Workshop on Scrambling and

Wh-questions at the 4th GLOW in Asia, August 23rd, 2003, Seoul National University, Seoul, Korea.

Ioup, G. (1975). Some universals for quantifier scope. In J. Kimball (Ed.), *Syntax and Semantics, vol. 4,* New York: Academic Press.

Johnson-Laird, P.N. (1969). On understanding logically complex sentences. *Quarterly Journal of Experimental Psychology, 21*, 1-13.

Konieczny, L., Hemforth, B., Scheepers, C. & Strube, G. (1997) The role of lexical heads in parsing: evidence from German. *Language and Cognitive Processes, 12,* 307-348.

Kurtzman, H.S., & MacDonald, M.C. Resolution of quantifier scope ambiguities. *Cognition, 48*:243--279, 1993.

Legrenz, P, Girotto, V., & Johnson-Laird, P.N. (2003). Models of Consistency. *Psychological Science* 14 (2), 131-137.

Ronat, M. (1984) "Logical Form and Prosodic Islands ". In D. Gibbon and H. Richter (Eds.),*Intonation , Accent, and Rhythm: Studies in Discourse Phonology*. New York: Walter de Gruyter, 311-32

A mechanism of ontological boundary shifting

Shohei Hidaka (hidaka@cog.ist.i.kyoto-u.ac.jp)
Jun Saiki (saiki@cog.ist.i.kyoto-u.ac.jp)
Graduate School of Informatics, Kyoto University;
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, JAPAN

Abstract

Past research on children's categorizations has suggested that children use perceptual and conceptual knowledge to generalize object names. Especially, some researches suggested that the relation between ontological categories and linguistic categories is a critical cue to categorize objects. However, this mechanism has not been specified. This paper reports new insights to reveal children's categorizations based on the survey of adults' knowledge. We estimated the English and Japanese ontological spaces from data and used these results to simulate behavioral experiment of previous research. The results show a possibility that linguistic cues help children to attend specific perceptual properties.

Introduction

Categorization is a form of information compression, one solution to handle an almost infinite number of entities efficiently. Where do these categories come from and how do children know which words to map to which categories?

Quine (1960) pointed to the difficulty of word learning without prior category knowledge. If we hear a novel word in an unknown language, how do we infer its meaning ? For example, suppose we heard 'gavagai' while looking at a rabbit in a field. 'Gavagai' might mean *rabbit*, but it could also mean *rabbit's color* or an infinite variety of other possibilities. By Quine's analysis, word learning should be highly problematic for first language learners.

Constraints to acquire word meanings

Markman & Hutchinson (1984) among others proposed that children learn nouns easily because they do have prior knowledge about kinds of categories. Research over the past 20 years has indicated that this knowledge is considerable. Children know, for example, that animal categories are organized by multiple similarities, that artifact categories are organized by shape, and that substance categories are organized by material. Given a single thing of each of these kinds and told its name, children systematically generalize that name to new instances in ways specific to the kind of thing it is (Landau, Smith & Jones, 1988; Soja, Carey & Spelke, 1991, etc.) One hypothesis is that this knowledge is learned, that as children learn common nouns, they learn the correlations between properties specific to different kinds and the similarities relevant to categorizing those kinds - that things with eyes are classified by multiple similarities, things that are solid and rigid are classified by shape, and things that are nonsolid and nonrigid are classified by material (Jones & Smith, 2002; Yoshida & Smith, 2003). The learning hypothesis is plausible because children's differential categorizations of animates, objects and substances emerge only after they have learned some number of names for these different kinds (Samuelson & Smith 1999), The present paper provides a simulation model of how this might be learned.

Linguistic categories and ontological categories The fact that children distinguish animal, object, and substance categories in noun learning is also interesting because these ontological categories are often related to linguistic individuation, to how different languages quantify nouns. Most of the world's languages treat animates as countable discrete things. Others, like English, also treat inanimate objects as discrete and countable. Few (if any languages) individuate substances in these ways. (Lucy, 1992)

Some have suggested a deep relation between ontological categories and their learning how their language quantifies entities. For example, Quine (1960) hypothesized that children learning English learn to distinguish objects and substances by learning the count-mass distinction. In English, nouns such as "dog" and "cup" that label individuated things are count nouns and mandatorily take the plural if there are more than one instance; in contrast, nouns such as "sand" that label a substance are mass nouns and are not pluralized. Thus this linguistic distinction could teach children that there are two different kinds of categories.

Soja, Carey and Spelke (1991) criticized this idea, because their experiments indicated that 2-year-old children who do not use count-mass syntax nonetheless classify objects by shape and substances by material. Imai & Gentner (1997) reported supporting results in a study comparing English and Japanese speakers. Japanese makes no distinction comparable to the count-mass distinction in its quantification system. Yet Imai and Gentner found that both Japanese- and English-speaking children categorized objects by shape and substances by material.

A boundary shift Imai and Genter also found differences in the range of things treated as objects versus substances by speakers of the two languages. Speakers

of English treated complex and simple solids as objects categorized by shape and nonsolid forms as substances categorized by material. In contrast, Japanese speakers treated complex solids as objects classifiable by shape and treated simple solids and nonsolids as substances classifiable by material. Yoshida and Smith interpreted these results in terms a boundary shift, suggesting that count-mass syntax shifted the object boundary in English relative to that in Japanese so that it also included simple but solid shapes. If this interpretation is correct then linguistic contrasts such as count-mass syntax may play a role in the development of ontological categories.

Yoshida and Smith also predicted and found a corresponding boundary shift at the animate-inanimate boundary. They predicted this from an analysis of Japanese which distinguishes animates and inanimates in ways that English does not, through its quantification system and also via the verbs "iru" and "aru" which mean "exists". In locative constructions such as "There is," animates require the use of iru and inmates require the use of aru. Yoshida and Smith hypothesized that this distinction, like the count-mass distinction in English, would perturb the boundary between animates and objects. In support of this idea, they showed that the range of things treated as animates (and classified by multiple similarities) was broader for Japanese-speaking than English-speaking children.

The purpose of the simulations reported here is to explain the mechanism underlying the boundary shift. Following Yoshida and Smith, we propose that ontological categories are the product of learned correlations among the properties such as shape, material and color and also linguistic contrasts such as the count-mass distinction in English and the iru-aru distinction in Japanese.

Experiment 1

We measured statistical structure of common noun category via adult judgments. We studied the statistical structure of 48 nouns that name common categories typically known by 2 year olds (Fenson, Dale, Reznick, Bates, Thal & Pethick, 1994) [1]We did this in two steps. First, we asked adults to judge how a list of 16 adjectives taken from some studies using the Semantic Differential (the SD; Osgood, 1957) described category relevant properties such as shape, material, movement. Second, in vocabulary survey, we asked adults to rate how the 16 adjectives described the 48 noun categories.

Method

Participants In the adjective survey, we recruited 12 volunteers (from 23 to 25 years old) from Kyoto university. In the vocabulary survey, we recruited 104 students (from 18 to 22 years old) from Kyoto Koka women's university who received a class credit for participation.

Stimuli The stimuli consisted of (1) a list of category relevant perceptual properties: shape, material, color, texture, sound, temperature, flavor, movement, smell, and function, (2) 16 adjective pairs: dynamic-static, wet-dry, light-heavy, large-small, complex-simple, slow-quick, quiet-noisy, stable-unstable, cool-warm, natural-artificial, round-square, weak-strong, rough hewn-finely crafted, straight-curved, smooth-bumpy, hard-soft; and (3) 48 nouns commonly known by young children [2] (see also Table 1).

Adjective survey Participants were asked - 'How do you use these words (adjective pairs) to express familiar objects' perceptual features'. Participants made these judgments using an electronic file of the 16 (adjectives) by 10 (properties) cells. The ratings were on a 5 point scale (1: very inappropriate, 2: inappropriate, 3: neither, 4: appropriate, 5 :very appropriate).

Vocabulary survey Participants were presented with one noun at a time and asked to judge the applicability of the 16 adjective pairs on a 5 point scale. For example, if the adjective pair was big-small, participants would be asked the thing labeled by the noun very small, small, ambiguous, big and very big. Five different orders were used across subjects.

Analysis We used Principal Component Analysis (PCA) to analyze the vocabulary with mean linguistic-scale scores of the all participants. PCA is a popular method to compress information by the least loss of data variance.

We used the results to estimate the English and Japanese ontological spaces. We added 1-dimension syntactic cues which was close to ontological categories (Table 1) to raw data (16 dimensions), and analyzed the combined data (17 dimensions). In the English condition, we added count-mass syntax which was encoded as 1-0. In the Japanese condition, we added iru-aru syntax just as in the English condition. In the neutral condition, we added the value 0.5 for all objects. We decided these parameters of syntactic categories based on the dictionaries. We assumed that (1) our ontology space consists of perceptual and linguistic properties, and that (2) the most important factor of these space is the variance of the object's distribution. These assumptions are reasonable, because (1) our goal is to estimate children's ontology space in the context of generalizing novel names and (2) we name entities different labels based on not similar features but different properties.

Our another goal is to estimate perceptual weights in two language conditions. However, principal components consist of weights of linguistic scales, so we can not directly know which perceptual weights the ontology spaces have. Therefore we defined perceptual weights of principal components as the equation(1) to analyze perceptual weights in English and Japanese conditions.

$$W_{dp} = \mid \sum_{l} C_{dl} M_{lp} \mid \qquad (1)$$

[1]This form of the MCDI is a parental checklist of words designed to measure the productive vocabulary of children between 16 and 30 months of age.

[2]The 9 categories are 'animals', 'body parts', 'clothing', 'food and drink', 'furniture and rooms', 'outside things', 'small household items', 'toys', and 'vehicles'.

Table 1: Linguistic categories of 48 nouns in English and Japanese. E=English, J=Japanese, c=count noun, m=mass noun, i=with-'iru' noun, a=with-'aru' noun

	E	J		E	J		E	J
butterfly	c	i	banana	c	a	water	m	a
cat	c	i	egg	c	a	camera	c	a
fish	c	i	ice cream	c	a	cup	c	a
frog	c	i	milk	m	a	key	c	a
horse	c	i	pizza	c	a	money	m	a
monkey	c	i	salt	m	a	paper	m	a
tiger	c	i	toast	c	a	scissors	c	a
arm	c	a	bed	c	a	plant	c	a
eye	c	a	chair	c	a	balloon	c	a
hand	c	a	door	c	a	book	c	a
knee	c	a	refrigerator	c	a	doll	c	a
tongue	c	a	table	c	a	glue	m	a
boots	c	a	rain	m	a	airplane	c	a
gloves	c	a	snow	m	a	train	c	a
jeans	c	a	stone	c	a	car	c	a
shirt	c	a	tree	c	a	bicycle	c	a

d is a dimension of principal components. l is a index of 16 linguistic scales of the SD (see also Method). p is the index of the 10 perceptual properties (see also Method). W_{dp} is the pth perceptual weight of dth principal component. C_{dl} is the loading of lth linguistic scales of dth principal component. M_{lp} is the estimated expressiveness of the pth perception of the lth linguistic scales. C_{d*} is a unit row vector and M_{*p} is a unit column vector, so W_{dp} is the absolute inner product of two vectors, or $\mid cos\theta \mid$ (θ is the angle of two vectors).

Results and Discussion

First three and six principal components respectively accounted for more than 70% and 90 % of the variability in the data.

Estimated ontological spaces The first two principal components of the vocabulary survey data were displayed as a 2-dimensional plot (Figure 1 is the result of neutral condition). In the neutral condition, we found animates and body parts in upper-right area, vehicles in upper-left area, furniture in lower-left area and substance in lower-right area. This distribution of entities leads us the following interpretation of the first two components. The first principal component axis can be interpreted as 'solidity', because solid and non-solid entities are located in the left and right sides, respectively. The second principal component axis can be interpreted as 'animacy', because dynamic and static entities are located in the upper and lower sides, respectively.

There were no clear boundaries in neutral 2-dimensional space, but we found global boundaries in the English and Japanese space. Furthermore, the English and Japanese spaces had a great difference. The English space also had 'solidity' axis as the first principal component, but the Japanese space had 'animacy' axis as the first principal component. Therefore, we analyzed these distributions of entities by clustering.

First three principal components (total 70% over) were enough to analyze global structure of results, so we analyzed this 3-dimensional data by hierarchical clustering(Figures 2 and 3).

The clustering of the neutral condition showed the clusters like MCDI classes, but did not show any global boundaries. On the other hand, the analyses of the English and Japanese conditions showed the global boundary (Figures 2 and 3). There were two global clusters categorized near by the root of the tree. One cluster mainly consisted of 'objects' category members, and another cluster mainly consisted of 'substance' category members. The second branch occurred in the object cluster. There was the 'animates' cluster near substance cluster in the part of objects cluster. That is why, English ontology space seemed defined by 'individuation' or 'solidity'.

In the Japanese condition, there were two global clusters that mainly consisted of 'animates' members and 'inanimates' members. Despite being inanimates, vehicles(e.g. 'airplane', 'car') and body parts(e.g. 'eye', 'hand') were near the animates members. There seemed an 'animacy' boundary in the Japanese ontological space because animates and dynamic objects make cluster and inanimates make another cluster.

Table 2: The estimated perceptual weights. In the Experiment 2, we used the normalized W_{dp} ($\sum_p^{10} W_{dp} = 1$).

	English	Japanese
shape	0.091	0.047
color	0.067	0.194
texture	0.086	0.09

Perceptual weights of the English and Japanese spaces We estimated perceptual weight in the English and Japanese ontological spaces. Tables 2 shows the results of the estimation.

Compared with the Japanese condition, the English condition showed higher weight on shape. Contrary to the English condition, the Japanese condition showed higher weights on color and texture.

Estimating ontological category One potential problem with the present experiment is that the perceptual ontological space was derived from only Japanese speaker's data.

We are currently collecting the English data. Preliminary results indicate that they are extremely similar to those of Japanese speakers. The Pearson's correlation of mean across participants are .79 in vocabulary survey and .80 in adjective survey. In this work, we have assumed that the adjective - noun ratings and the adjectives - properties ratings reflect the perceptual structure of the categories. It could be argued that these rating reflect instead how predicates and nouns co-occur in a language.

Sommers (1963) claimed that knowledge of ontological categories is intimately related to predicability, that is, to the knowledge of which predicates in a language can be combined with which nouns. For example, the predicate 'is asleep' distinguishes animals and non-animals. Furthermore, Keil (1979, 1981) showed that children's judgments of predicability, like those of adults, yield an

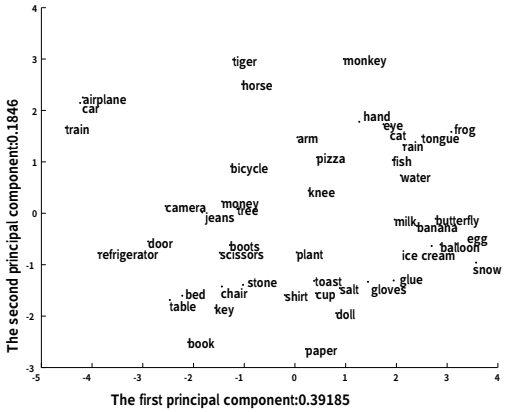


Figure 1: The first two principal components for the neutral condition. The first principal component (x axis) was interpreted as 'solidity' or 'size' of objects. The second principal component(y axis) was interpreted as 'animacy' or 'movement' of objects.

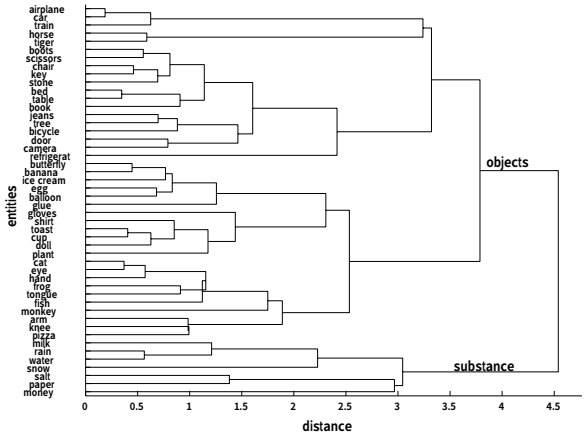


Figure 2: The result of cluster analysis for the English condition. We estimated 'objects' cluster and 'substance' cluster in superior hierarchy.

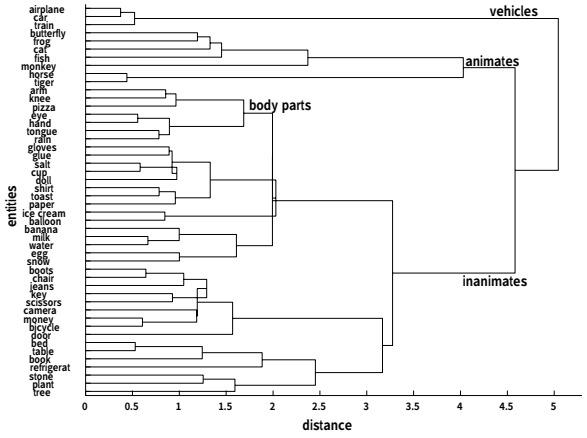


Figure 3: The result of cluster analysis for the Japanese condition. We estimated 'animates' cluster and 'inanimates' cluster in superior hierarchy.

ontological tree, though a less elaborate one than adults. The question of whether our judgments reflect the structure of categories in the world or relations among words is a difficult one. Given our preliminary results from English, if they do reflect relations among words-predicates and nouns,those relations are nearly identical in the two language, a fact one might want to explain by the regularities in the world.

Experiment 2

In Experiment 2 uses the results of Experiment 1 to simulate the boundary shift reported by Yoshida and Smith.

Experiment to be simulated The specific goal is to simulate Yoshida and Smith's second experiment which showed a boundary shift in the animate - inanimate boundary Japanese speakers relative to English speakers. The participants in Yoshida and Smith's experiment were 3-year-old English and Japanese monolingual children. The experimenters presented children with an ambiguous entity that could be seen as depictions of ether animates or inanimates and named it with a novel label (e.g. in Japanese 'Kore-wa teema dayo', in English 'This is a teema.'). Experimenters did not provide any cue such as "iru" or "aru" which might cue these as depictions of animates or inanimates Experimenters then presented children with test objects and asked them whether the test object had the same name Exemplars and test objects matched or did not match in three perceptual features (Table 3).

The results suggested that Japanese speakers treated these ambiguous forms as depictions of animates, extending the name to new instances by multiple similarities. In contrast, English speakers treated them as inanimates objects, extending the name to new instances by shape. Thus, Yoshida and Smith proposed that the Japanese speaking children included a wider range of kinds in the animate category relative to English speakers just as English speakers include a wider range of instances in the objects category than do Japanese speakers. The question for Experiment 2 is whether we can use the adult judgments in Experiment 1 to simulate these results.

Method

Following Yoshida and Smith's method, we assumed that objects categories are defined in terms of shape, color and texture, and that other nonstudied features will have no effect on the similarity of a test object to the exemplar.

We also assume that children's name extensions are based on the psychological distance between a test object and the exemplar. We defined the psychological distance between stimuli by the equation (3). Probability of 'yes' response which means two objects belong to the same category is defined by the equation (2).

$$P_{yes} = exp(-b\delta) \tag{2}$$

$$\delta = \Big(\sum_{i \subset perception} D_i w_{il} \mid (e_i - s_i)^m \mid \Big)^{\frac{1}{m}} \tag{3}$$

Table 3: Experimental conditions of Yoshida & Smith (2003). 'm' means feature match between exemplar and test object, and 'N' means non-match

condition	1	2	3	4	5	6
shape	m	m	m	m	N	N
texture	m	m	N	N	m	N
color	m	N	m	N	N	m
	S+T+C	S+T	S+C	S	T	C

$b > 0$ is the scaling parameter of the transfer between a distance and a yes-response ratio , and $m > 0$ is the metric parameter. $i \subset perception = \{shape(S), color(C), texture(T)\}$ means the population of the perceptual features. e_i represents the ith perceptual dimension of the exemplar, and it is a random value from 0 to 1. s_i represents the ith perceptual dimension of the test stimulus. s_i is a random value from 0 to 1 in case of feature non-match or the same value as the exemplar in case of feature match (see also Table 3). w_{il} is the value of ith perceptual weight of l ($l \subset \{English, Japanese\}$) participant(see also Table 2). D_is are the supplementary terms which represent ith perceptual bias common in English and Japanese. We added these terms to the model because the feature differences of stimuli were not controlled in the behavioral experiment. D_is represent the relative mean difference of perceptual features. The model have four free parameters (b, m, two D_is) ,because D_is are the ratios among three perceptual features.

Results and Discussion

We simulated the second experiment of Yoshida and Smith (Figure 4) by the computational model (Figure 5). We used Monte Carlo simulation to estimate optimal parameters. In the result, we estimated $b = 12$, $m = 0.8$, $(D_{shape}, D_{texture}, D_{color}) = (7, 1, 0.6)$ ($D_{texture} = 1$ is constant) and $R^2 = 0.916$ between the response patterns (12=2 (language of participants)×6 (feature controlled condition)) of simulation and those of behavior. When we did not add two parameters D_is, the fitness of the model was $R^2 = 0.683$. This suggested the methodological problem of estimation by the equation (2).

In the behavioral experiment, the English speakers categorized the stimuli based on their shape and the Japanese speakers categorized them based on their multiple features. These results suggested that the English speakers categorized ambiguous objects as inanimates whereas Japanese speakers categorize them as animates. Thus our model fitted the behavioral results well, and provides a simple account of the crosslinguistic difference.

General Discussion

Recent studies on early word acquisition have shown that some biases, such as shape bias, are not so universal, but dependent on context and language. For example, Children speaking English show stronger shape bias for inanimate objects than those speaking Japanese. These findings are explained by postulating children's linguistic, cultural and category knowledge influences boundaries between ontological categories.

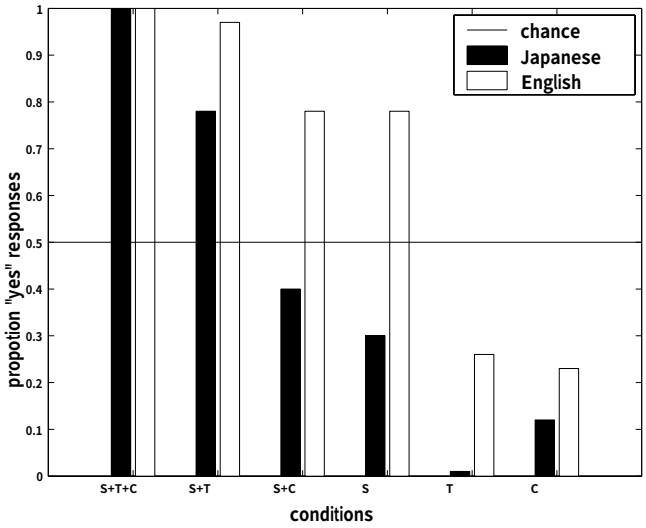


Figure 4: The behavioral data of Yoshida & Smith (2003). The English speakers categorize stimuli based on shape, while the Japanese speakers categorize them based on multiple features.

The present simulation offers a mechanism. The results of the simulation suggest that (1) ontological categories may not be a special nor given but an emergent property derived from multidimensional perceptual and linguistic features, and (2) crosslinguistic differences along this ontological continuum can be explained by a difference in the emergent variable due to different statistical structure of linguistic features. Specifically, we assumed that the emergent property can be extracted by information compression of the multidimensional feature space, such as PCA. To evaluate whether we can account for the behavioral findings, we conducted a survey to obtain the multidimensional feature space of objects, and a series of quantitative analyses to obtain the language specific ontological spaces. Without linguistic features, the compressed perceptual space spanned by two principal components was organized by objects' solidity or size. Thus, a solidity-dominant space can be derived from the perceptual feature space, but there was no principal component representing an "individuation continuum" from animates to objects to substances. More interestingly, addition of linguistic features made the ontological space more well-defined, and the estimated language-specific ontological spaces are quite consistent with previous findings. The estimated English ontological space was solidity-dominant and shape-weighted. This is consistent with Colunga & Smith (2000) and Samuelson (2002) showing that American children attended solidity of objects in object categorization. On the other hand, the estimated Japanese ontological space is animacy-dominant and color-and-texture-weighted, which is consistent with Yoshida & Smith (2001, 2003) showing that

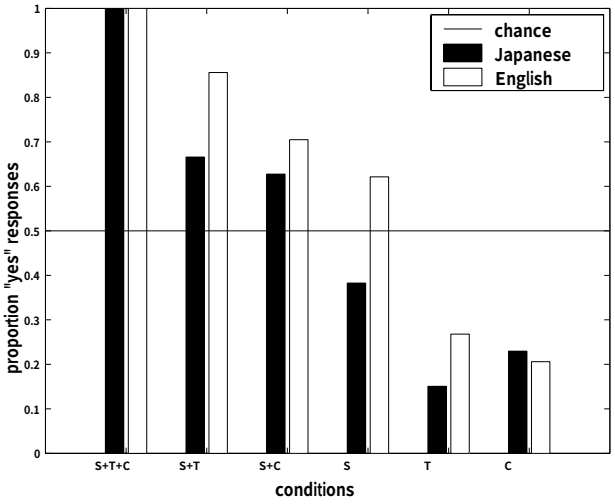


Figure 5: The result of simulation.
The coefficient of determination of 12 responses pattern (R^2) is 0.916

Japanese children attended multiple features of objects.

Furthermore, objects/substance boundary was clearer in the English space than the Japanese space. This results is consistent with Imai & Gentner (1997). In addition to qualitative matches with previous data, our theory make a good quantitative fit to the behavioral data of Yoshida & Smith (2003). With a simple computational model that categorization response is based on similarity derived from a distance on the ontological space, the behavioral data showing difference in shape bias for objects between English and Japanese speaking children with various different stimulus conditions could be simulated quite well.

Beyond the "boundary shift" Our account expands the boundary shift hypothesis in the following senses. First, our theory proposes an underlying mechanism of boundary shift in a quantitative fashion. Second, the individuation continuum is not a separate dimension, but a statistical property embedded in the multidimensional feature space. Ontological features such as animacy and solidity may be extracted from perceptual and linguistic features through statistical learning. This suggests a possibility that more abstract conceptual features are also formed by statistical learning of basic perceptual and linguistic features.

Acknowledgments

This work was supported by Grants-in-Aid for Scientific Research from JMEXT (No. 15650046), and the 21st Century COE Program from JMEXT (D-2 to Kyoto University).

References

Colunga, E. & Smith, L. (2000) Committing to an Ontology: A Connectionist Account, *The Twenty Second Annual Meeting of the Cognitive Science Society*.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59* (5, Serial No. 242) Chicago: University of Chicago Press.

Imai, M. & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence., *Cognition*, *62*, 169-200

Jones, S. S. & Smith, L. B. (2002). How relevant properties for generalizing object names., *Developmental Science*, *5*, 219-232

Keil, F. (1979). Semantic and conceptual development: An ontological perspective, MA: Harvard University Press.

Keil, F. (1981). Constraints on knowledge and cognitive development, *Psychological Review*, *88*, 197-227.

Landau, B., Smith, L.B. & Jones, S.S. (1988). The importance of shape in early lexical learning, *Cognitive Development*, *3*, 299-321.

Lucy, J.A. (1992). Language diversity and thought: A reformulation of the linguistic relativity hypothesis., Cambridge: Cambridge University Press.

Markman, E.M. & Hutchinson, J.E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations., *Cognitive Psychology*, *16*, 1–27 .

Osgood, C.E., Suci, G.J. & Tannenbaum, P.H. (1957). *The measurement of meaning.*, Univ. of Illinois Press.

Quine, W.V.O. (1960). *Word and Object.*, Cambridge, MA:MIT Press,.

Samuelson, L.K. (2002) Statistical Regularities in Vocabulary Guide Language Acquisition in Connectionist Models and 15-20 Month Olds., *Developmental Psychology*, *38*, 1016–1037.

Samuelson, L. and Smith, L.: Early noun vocabularies: do ontology, category structure and syntax correspond?, *Cognition*, Vol. 73, pp. 1–33 (1999).

Soja, N. N. , Carey, S. & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meanings: object terms and substance terms., *Cognition*, *38*, 179–211.

Sommers, F. (1963). Types and ontology., *The Philosophycal Review*, *72*, 327–363.

Spelke, E. S. (1990). Principles of object perception., *Cognitive Science*, *14*, 29–56 .

Yoshida, H. & Smith, L. B. (2001) Early noun lexicons in English and Japanese., *Cognition*, *82*, 63–74.

Yoshida, H. & Smith, L. B. (2003) Shifting ontological boundaries: how Japanese- and English- speaking children generalize names for animals and artifacts., *Developmental Science*, *6*, 1–34.

Perception as Prediction

Stephen A. Hockema (shockema@indiana.edu)
Psychology and Cognitive Science, Indiana University
1101 E. Tenth St., Bloomington, IN 47405 USA

Abstract

Learning is often about prediction. This paper asks whether perception is also. The main idea is that perception is a stream and that perceivers learn the trajectory through which one moment in that stream turns into the next. A behavioral experiment with children is described that tests two hypotheses developed from this idea.. In the experiment, children briefly watch a transformation (e.g., a triangle increasing in size and/or saturation). If children learn the trajectory of change and if prediction is at the core of perception, then a subsequent statically presented object should trigger the perceptual system to anticipate the "next state". To test this, children were asked to make same/different judgments that should, by hypothesis, be interfered with by the learned trajectory. Children became less able to detect pairs that were the "same" when asked to make judgments about the dimension that they had seen varied. Furthermore, there was evidence that two dimensions could be made more integral by covarying them simultaneously. Both of these results were simulated with a simple connectionist model constructed to embody a predictive mechanism. Taken together, these results lend support to the idea that the perceptual system is designed to make predictions in time and that this architecture gives a "dynamic" aspect to perception.

Introduction

Perception is an interaction *in time* between a dynamic mind and a dynamic world. It is crucial to the understanding of this process that *both* sides of the relation are changing in time. This paper examines two key implications of this fact: 1) an element of *prediction* must be inherent in the perceptual process, and 2) the system should be *adaptable* to a changing environment.

There is ample evidence for the adaptability of the perceptual system in the short and long term. In the short term, the evidence shows up in the form of aftereffects and priming. For example, motion aftereffects (MAEs), illusions of motion (without displacement) that occur after viewing real motion in a certain direction for a short time, are considered to be perceptual adaptations that serve to keep the system in balance (Anstis, Verstraten, & Mather, 1998). In the long term, adaptation shows up as the perceptual learning of new psychological dimensions and features and the readjustment of the relative attention paid to existing dimensions (Goldstone, 1998). In brief, our perceptual systems constantly tune themselves to the environment.

Furthermore, our perceptual systems must also be tuned to *anticipate* the future. In the short term, the system must be able to predict, at a very low level, how our environment and the things in it change and appreciate which changes are normal and which are unexpected. Normal changes include regular transformations along dimensions, for example changes in position, size, orientation, luminance, pitch, and so on.

There is evidence for the predictive capabilities of the system in many everyday activities, like tracking moving objects behind occluders, navigation through crowds, simple eye-hand coordination, and in several related experimental phenomena where the system exhibits momentum when tracking predictable changes. For example, in the phenomenon known as representational momentum , subjects learn to anticipate the continuation of a transformation, for example an object moving across a computer screen. When the object suddenly disappears and subjects are probed about its nal position, there is evidence that the remembered position is shifted forward along the path of the trajectory (Freyd, 1992). This has been taken by some (e.g. Freyd, 1992; Hubbard, 1999) as evidence of a dynamic representation that continues to move forward after the real object has disappeared. However, the evidence is also consistent with the possibility that the trajectory of the object is being perceptually *predicted* and thus the perceived position of the object at any given instant in time is actually ahead of its actual position.[1] Such a view would be consistent with an explanation for several other perceptual illusions given recently by Changizi and Widders (2002) and Changizi, Nijhawan, Kanai, and Shimojo (2003).

This paper will present further empirical evidence of a perceptual adaptation that is triggered by predictable variation. A simulation of the experiments using a simple model of a predictive mechanism will then be presented as additional support for these ideas.

Experiment

If an adaptable, predictive mechanism is built into our low-level perceptual systems then it should be possible to prime the system to anticipate and perceive change as in the momentum effects, even when tested with static objects. If the primed change is along a dimension, this might disrupt the perception of the value for this dimension in a subsequently presented static object. If caused by an adaptable, predictive mechanism, this disruption would have two characteristics:

1. a perceptual *shift* in the direction of the primed change; and
2. less certainty about the precise value of the dimension.[2]

[1] Of course it is possible that both of these theories are operating.

[2] This is analogous to uncertainty principles in physics, although it also stems from the imperfection that will be inherent in any physical predictive mechanism, especially in circumstances where it has only had a brief exposure to the trajectory that it is trying to learn to predict.

This experiment deals with the second characteristic. It was designed to test the hypothesis that priming the perceptual system with bidirectional change along a dimension could lead to perceptual spreading along it. Such an effect should make it more difficult for perceivers to make accurate *comparisons* along this dimension subsequent to experiencing this priming. In particular, the spreading should increase the likelihood that perceivers will see two things that are actually the same as being different on the dimension.

In this experiment, this hypothesis was evaluated using a task where the dimensional change was expansion and contraction in size. Thus, it was expected that subjects would have more difficulty judging two shapes as being the same size after exposure to change along the size dimension than after being exposed to no change, or to change along an irrelevant dimension. In this experiment, color saturation was used as the control dimension. The participants were preschool-aged children under the assumption that the developing perceptual system is more sensitive to these priming effects.

The use of preschool children was also motivated by the desire to investigate the usefulness of a predictive mechanism to a developing perceptual system. Based on the simple idea that "things that change together go together", the hypothesis was formulated that if the perceptual system adapts to a coherent transformation along more than one dimension, these dimensions might become perceptually "fused", or more *integral* in the sense formalized by Garner and Felfoldy (1970).

Method

The experiment was a between-subjects design. Subjects were randomly assigned to one of four conditions. There were also four possible orderings of the test trials (see below). Subjects were randomly assigned to one of these four orders, which were counter-balanced across the four conditions.

Participants Thirty-two children from the Bloomington, IN area have participated. Children were all between 4 years and 4 years, 8 months of age. (Mean age was 4.3 years.)

Procedure The experiment was divided into three phases, all of which used a computer to present the stimuli.

In the first, "warm-up" phase, subjects were familiarized with the computer and trained to press the space bar whenever they were presented with a pair of shapes on the screen that were the same *size*. This phase was the same for all subjects, regardless of what condition they were in. The items presented included pictures of various common and colorful objects (e.g. balls, flags, hearts, etc). Across trials matching and mismatching objects and matching and mismatching sizes varied orthogonally so as to instruct the child as to the importance of attention to size only.

Subjects were presented with 32 warm-up comparisons, starting at a slow presentation rate and gradually increasing to 1.25 seconds per pair. Whenever a new pair appeared, the computer would emit a short beep. A different pitched beep would sound whenever the space bar was pressed. If an error was made (either the pair were the same size and the child failed to press the space bar, or the pair were different sizes and the child incorrectly pressed the space bar), the computer would emit a lower beep and the warm-up sequence would stop. The experimenter would then explain the mistake to the child, pointing out what they should have done, and then continue the sequence. If, after 32 trials, the child had not yet grasped the task, the training phase was repeated. If they still had not grasped the task after three passes, the subject's data was replaced.

In the second phase, the "priming" phase, all children were shown a simple video animation that lasted 80 seconds and consisted of 20 repetitions of a "transformation" of two shapes. The transformation that children saw depended on their condition assignment and will be described in the next section, however the instructions were identical for all four conditions. Children were told that they were playing a game in which they were supposed to press the space bar whenever the shapes "stop changing" and "begin to change back". The game was simply a ruse to help keep the children paying attention to the crucial animations.

The final phase of the experiment, the "test" phase, was similar to the training phase. Here subjects were again told to press the space bar as quickly as possible whenever they saw a pair of shapes that were the same size. In the test phase, the shapes being compared were either a pair of circles or a pair of squares. Half of the comparisons were the same size and half were different. The shapes were either both blue or both red, with their saturations varying as described in the next section. As in the warm-up phase, the computer emitted a beep whenever a new pair was displayed and a different pitched beep whenever the space bar was pressed. No feedback was provided during the test phase.

During the test phase, each pair was displayed for exactly 1.25 seconds. After being given the instructions, and prior to starting, subjects were warned that the speed would be fast and that they should get ready. There were 32 test stimuli in this phase, broken down into four sets of eight. After each set of eight, the computer would pause and the subject would be reminded of the instructions, and told to get ready again. Test trials were presented to subjects in one of four random orders. To ensure that children were still on task, any child that pressed the space bar on at least 20% or more than 80% of the test trials was replaced.

Materials The priming phase was designed to "teach" the perceptual system a predictable trajectory of change. There were four possible animations used corresponding with the four conditions. All animations showed the gradual transformation of a pair of side-by-side triangles. The left and right triangle always transformed identically and in synch with one another. The left triangle was always red and the right triangle was always blue. The four conditions were as follows:

1. Control: increasing and decreasing saturation;
2. Size-Only: increasing and decreasing size;
3. Correlated: both size and saturation increasing and decreasing: the bigger the triangles got, the more saturated, and the smaller they got, the less saturated;
4. Anti-correlated: both size and saturation increasing and decreasing: the bigger the triangles got, the less saturated, and the smaller they got, the more saturated.

Since *size* is at least one of the transformed dimensions in Conditions 2-4, the three conditions will collectively be referred to as the "Size-change" conditions. Conversely, since *saturation* is at least one of the transformed dimensions in Conditions 1, 3 and 4, these three conditions will be referred

to together as the Saturation-change conditions.

The minimum size (area) of the triangles in the three Size-change animations, as displayed on the monitor, was $1.83cm^2$ (base $1.63cm$, height $2.25cm$); the maximum size was $29.25cm^2$ (base $6.5cm$, height $9.0cm$). The minimum saturation in the Saturation-change animations was 0.1 on a scale of 0 to 1; the maximum was 1.0. For Condition 1, the size of the triangles remained constant at $29.25cm^2$. For Condition 2, the saturation of both triangles remained constant at 0.8. For the Size-change animations, the triangles always started and ended at their smallest point.

In the Testing phase, there were four different types of comparison possible. These were among shapes that were either big, medium or small in size (see Table 1) and high, medium or low in saturation (1.00, 0.45 and 0.20 respectively in the range of 0 1).

Table 1: Actual On-Screen Areas (in cm^2)

Term	Squares	Circles
Big	27.56	9.33
Medium	12.25	5.25
Small	4.86	2.33

The four types of comparison were as follows:

1. Identical: Pair being compared were identical in both size and saturation. There were four variations of this: big/high, big/low, small/high, small/low.

2. Saturation Different: Pair being compared were the same size, but differed in saturation. There were two variations of this: big/high compared to big/low and small/high compared to small/low.

3. Size Different: Pair being compared were the same saturation, but differed in size. There were four variations of this. In the rst two, the pair were both high saturation. In the second two, the pair were both low saturation. One of the shapes was always medium size and the other was always either small or big.

4. Both Different: Pair being compared were different both in size and saturation. In these pairs, the bigger shape always had the higher saturation. (This was done for reasons related to other experiments not reported here.) As in the Size Different comparisons, one of the shapes was always medium size and the other was always either small or big.

Results

Children's errors were classi ed by the type of comparison trial in which they occurred and the condition to which the subjects were assigned. There were two broad classes of errors: *Misses*, where the shapes were the same size yet the subject failed to press the space bar, and *False Alarms* where the shapes were different size and yet the subject incorrectly pressed the space bar.

Figure 1 shows the average number of Misses broken down by the four priming conditions and two types of relevant test trial (Identical trials on the left and Saturation Different trials on the right in each group). As can be seen, there were signi cantly more Misses in the Size-Change conditions as compared to the Control Condition ($p < .02$ for the Identical trials, $p < .002$ for the Saturation Different trials).

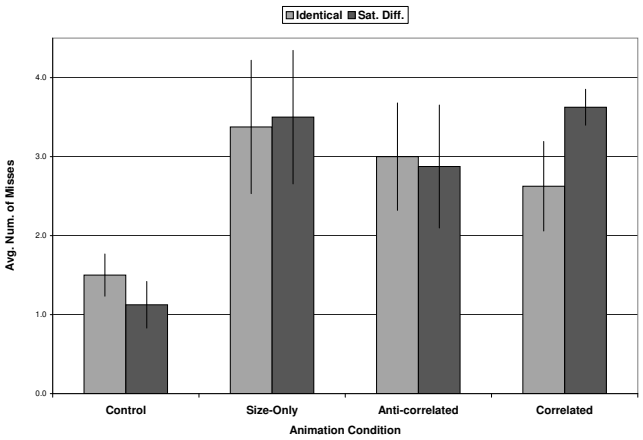


Figure 1: Average Misses for Same-Size trials

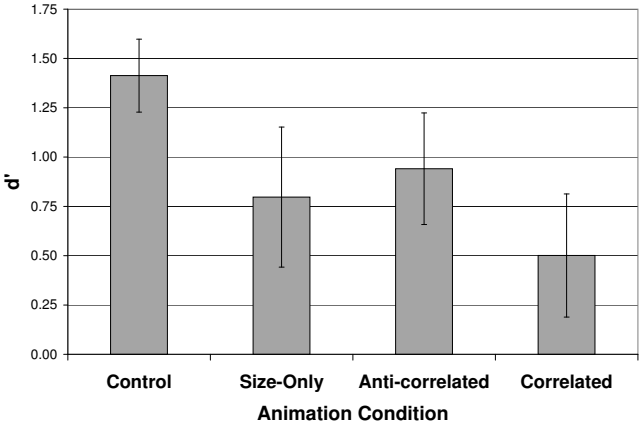


Figure 2: Average disciminibility

An analysis from the perspective of Signal Detection Theory (SDT) was also performed. Figure 2 shows the average d' (discriminability or sensitivity) for each of the four conditions.[3] As can be seen, the three Size-change conditions had a signi cantly lower average d' value ($p < .03$) than the Control Condition. (In this context, lower d' values mean that it is harder to distinguish Same Size shapes from Different Size shapes.) The three Size-change conditions also had a marginally signi cant ($p < .06$) higher average β value (criterion for pressing the space bar): they were less likely to respond Same Size in general. Also, the Correlated condition by itself also had a signi cantly lower d' ($p < .01$) and higher β ($p < .03$) than the Control condition.

Figure 3 shows the False Alarms, again broken down by the four conditions and two types of relevant test trial (Size Different trials on the left and Both Different trials on the right). Considering Figure 3, notice that the gap between the number of False Alarms in the Different Size, *Same* Saturation trials and the number of False Alarms in the Different Size, *Different* Saturation trials increased in the conditions where size and saturation were covaried (Correlated and Anti-correlated) over what it was in the conditions where they were varied in-

[3] For one subject in Condition 2, d' was infinite because the subject had a Hit Rate of 1.0. For this subject, d' was estimated using an adjusted Hit Rate of $15.5/16 = .9688$, yielding a d' of 3.01.

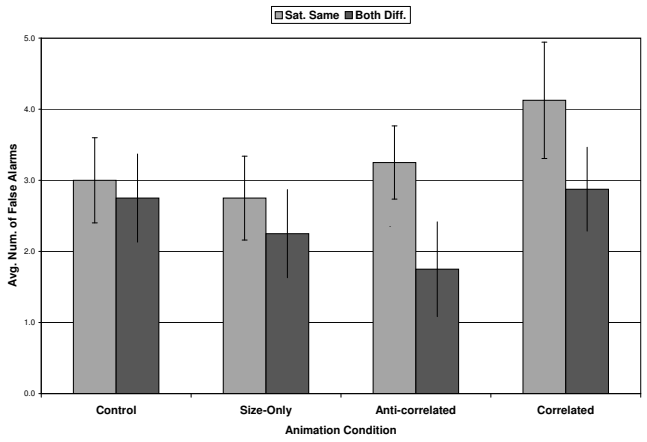


Figure 3: Average False Alarms for Different-Size trials

dependently. The graph shows the group averages and the error bars are the standard error of these means. For each subject, the difference between their number of False Alarms on the Different Size, Same Saturation trials and the Different Size, Different Saturation trials was computed. This difference was found to be signi cantly above zero in both the Correlated ($p < .006$) and Anti-correlated ($p < .028$) conditions. Comparing to the Control and Size-only Conditions, it can be seen that no such difference existed where the dimensions were varied independently during the priming event. When the within-subject difference scores of the two covaried conditions are compared to the difference scores of the two independent conditions, the covaried differences are signi cantly bigger ($p < .04$). Thus, differences in the irrelevant dimension of saturation had a larger effect on subjects' comparisons of size during test when size and saturation were covaried together during the priming event.

To see if the effect of the priming event looses its potency over time, the probability of error on any given test trial was correlated with how long after the priming phase it occurred. Table 2 shows the correlation coef cients between error frequency and trial. In no case was the correlation signi cant. Thus, performance did not change throughout the test phase.

Table 2: Correlation between trial number and error prob.

Condition	r
Control	$-.12$
Size-only	.12
Correlated	$-.17$
Anti-correlated	$-.29$

Discussion

The results appear to support the rst hypothesis of perceptual spreading. Priming with size transformations in the three Size-change conditions led to a decrease in the ability to detect shapes that were the same size in the Test phase. This is consistent with spreading (bidirectional propagation) along the size dimension due to the anticipation of change by the perceptual system. Such spreading apparently leads to dif - culties in comparison.

The speeded comparison task focusing on just size was chosen speci cally to induce errors it is known that children have dif culty ignoring variation on an irrelevant dimension to focus on just one dimension (Smith & Evans, 1989). The point of interest here was how the type of errors would be affected by the prior experience with a systematic trajectory of change. It is noteworthy that in the Control and Size-only conditions, performance was *not* signi cantly different for test comparisons where the saturation was different than it was for comparisons where the saturation was the same. Thus, size and saturation were fairly separable in these two conditions. The fact that the gap in the number of False Alarms widened in the two covaried conditions over the two independent conditions is taken as support for the second hypothesis: It appears that coherent, predictable change along both the size and saturation dimensions caused them to become more integral, such that saturation differences had signi cantly more of an effect on size comparisons. This nding has developmental implications, providing a possible account for how low-level *features* and *properties* that start out as perceptually distinct can congeal into perceptual *dimensions* when they are experienced as covarying in a predictable and coherent way. One thing that is very interesting about the present results is that it took only a relatively short exposure (80 seconds) to such regularity to produce this effect! Further, the effect appeared to decay slowly (not measurably) over the course of the experiment. Taken together, the quick adaptation and slow decay of the effect suggest that an interesting avenue of future research will be to explore the relationship between the amount of experience with certain types of transformation and the duration of the adaptation. This type of predictive learning of correlations may well be a potent part of the developmental process.

Signal Detection Theory: An Alternative Explanation
Signal Detection Theory (SDT) is a type of analysis that assumes that the system responsible for making decisions is inherently noisy, such that, for example in the case of this experiment, even when the shapes are obviously very different in size there is still a chance that the subject will respond Same Size . What makes SDT really useful here is that it provides a way of separating subjects' propensity to respond Same Size (their *bias* or *criterion*) from their ability to tell the difference between the Same Size test trials and the Different Size trials. The predicted perceptual spreading effect should result in a decrease in this second ability, not merely a change in bias. And, importantly, there was a signi cant difference in discriminability (d') triggered by experience with the Size-change transformations. Subjects had a harder time distinguishing between Same Size and Different Size test trials in these conditions, as the spreading hypothesis would predict.

The assumption of inherent noise that underlies SDT also provides another way to explain the data. It is possible that priming the system with changing size (dynamic priming), increased the internal noise in the perceptual system related to size judgements. (In SDT terms, this amounts to increasing the variance of both the signal absent and signal present distributions.) This would show up in the analysis as a decrease in d', as seen. This provides a potentially useful description at a different level of abstraction. Indeed, the uncertainty stemming from the proposed predictive mechanism

might provide a mechanistic explanation (at a lower level) for the increased noisiness.

Yet another possibility, and perhaps a simpler account of the present data than the adaptive prediction hypothesis, is that the dynamic priming of size increased the children's sensitivity to size differences, enabling them to make ner-grained distinctions of size. In this case, the inherent noisiness of the system would have more of an effect. Subjects would be more sensitive to very slight discrepancies in size caused by noise and would be less likely to respond Same Size in general (i.e., they would both Miss more and False Alarm less). This would show up in the analysis as a difference in their response bias, β.

The fact that there was a marginally signi cant increase in β in the three Size-change conditions means that this hypothesis cannot be ruled out here as a possibility.[4] It should be noted that this alternative only offers an explanation for the overall decrease in performance in the three Size-change conditions relative to the Control condition, but it does not address the effects related to increased Integrality. Nor does this bias-shift account explain the (more signi cant) shift in d', which indicates that subjects in the size-change conditions really did have more dif culty distinguishing the Same Size test trials from the Different Size trials. For this, an adaptive, prediction mechanism still seems to be reasonable.

A motion aftereffect? The analysis of the probability of error as a function of trial is interesting because it shows that the induced effect does not decay rapidly. If it were a typical motion aftereffect, 80 seconds of exposure to the animation motion might be expected to trigger on the order of 10 seconds of aftereffect, as motion aftereffects typically decay with the square root of the time exposed to the inducing motion (Anstis et al., 1998). Yet the error rate showed only a very slow decay, lasting over a minute.

Furthermore, the present effects are also set apart from typical motion aftereffects in that they occur after seeing *bidirectional* motion. Motion aftereffects typically occur in the opposite direction from the direction of inducing motion (Anstis et al., 1998). The theory behind this is that the visual system adapts to correct what it (mistakenly) takes to be drift in the neurons detecting motion in the inducing direction and lowers their weight relative to neurons sensitive to motion in the opposite direction. If this accurately re ects what really happens (in a MAE), then bidirectional motion should not produce the aftereffect because there will be no incentive for the visual system to suspect drift in the rst place, the opposing motions will cancel each other out.

Thus, it would appear that the effect observed in this experiment is a new and different type of perceptual adaptation that is related to traditional motion aftereffects, but underwritten by a potentially different mechanism.

Model Simulation

The basic principle of perceptual prediction was embodied in a connectionist model consisting of a simple recurrent network (Elman, 1990). Its task was to actively sample its sensory input and try to predict how it changes in time. By its nature, such a network allows for supervised learning in the sense that it can validate its predictions by what eventually happens. Thus, whenever the model encounters consistent, gradual, continuous variation (for which it cannot already account), it might actively train itself to predict this variation. The model was constructed such that its predictions could be fed back into its input units, enabling extrapolation in time and giving perception temporal extent .

There is not room in this paper to go into the details of the input representation and training procedure. Basically, a shape was represented by its values on the size and saturation saturation dimensions. There were three different training sequences corresponding to the four animation conditions in the experiment. (The Correlated and Anti-correlated conditions were equivalent in the model representations, so they were combined into a single simulation condition called Covaried .) For each pattern in a sequence, the network was trained to predict the next pattern. A small amount of noise was injected into this process.

Sixteen networks were trained and tested in each condition on the same sequences of test pairs that the children saw in the experiment. The model made *comparisons* as follows: Given two patterns to compare, one was chosen to go rst, passed through the network and the outputs were buffered. Then the other pattern was passed through the network and its outputs on the nine size units were compared to the buffered outputs of the rst pattern using a cosine distance metric. If they fell within 30 degrees of one another, the process was stopped. If not, the buffered outputs from the two patterns were then fed back in as inputs (keeping the same context layer activations).[5] This process repeated until either the output vectors eventually came within $\lambda = 30$ degrees of one another or 10 iterations had gone by. Whenever this process terminated, the following value was calculated:

$$\text{How Different} = \text{Iterations Required} + \frac{1 - cos(\theta)}{1 - cos(\lambda)}. \quad (1)$$

This equation gives an estimate on how different the two (*dynamic*) representations were from one another. The more iterations that were required, the more different the patterns were. θ is the nal angle between the two output vectors.

Finally, the model decided whether or not to say Same Size for the patterns based upon how different they were from one another. The likelihood of saying same was inversely proportional to the score computed by Equation 1. This was operationalized with:

$$\text{Say Same} = (rand < \exp(\frac{-\text{How Different}}{\beta})) \quad (2)$$

where *rand* was a random number uniformly distributed between 0 and 1, and β was a bias parameter that was set to 2 for this simulation.

Results and Discussion Figures 4 and 5 show the average Misses and False Alarms (respectively) over the simulation runs. These are comparable to Figures 1 and 3.

As can be seen, the basic trends that were present in the child data were also present in the model simulations. In particular, there were signi cantly more misses in the Size-change conditions as compared to the Control condition. Furthemore, the degree of integrality signi cantly increased in

[4] This has been taken up in other experiments not reported here.

[5] The prediction of dynamic spreading stems from this.

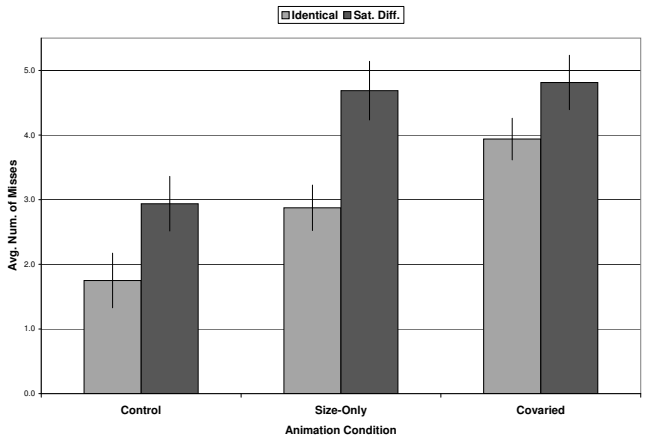


Figure 4: Misses from model simulation

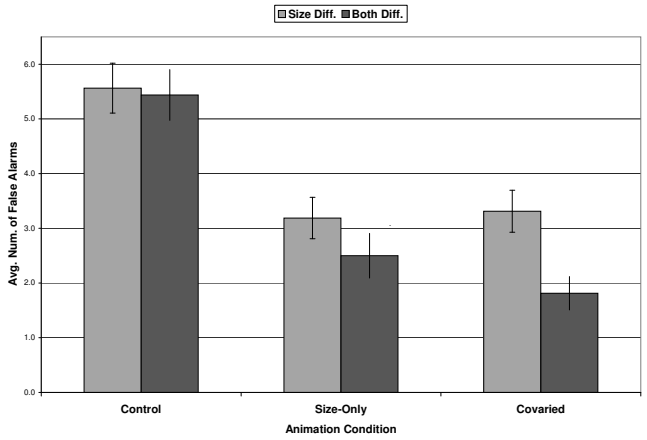


Figure 5: False Alarms from model simulation

the Covaried condition as seen in the increased False Alarm gap of Figure 5. This was due to the fact both the size and saturation dimensions shared the hidden layer and the network learned that each of these dimensions was a good cue to predict the next state of the other.

General Discussion

At the core of this paper is the idea that our perceptual systems are oriented around *transformations*. Transformations contain rich information about the structure of the world. Indeed, it is through this temporal structure that we perceive atemporal structure. (For example, movement is necessary for the detection of occluding edges.) Furthermore, perception itself is a process: it has temporal extent. There is never a single instant when we achieve a percept.[6] Moreover, given that the objects of our perception can be changing at the same time we are perceiving them, it behooves us to learn to anticipate their transformations.

The experiment presented herein was designed with this in mind. It provided evidence for a perceptual adaptation in response to brief experience with a predictable trajectory. It showed that there is a dynamic component involved in perception, even with *static* shapes that are *perceptually present*. It also showed that experience with coherent transformations might have developmental consequences, given that the adaptation caused two dimensions that were initially fairly separable to become more integral. The model simulation then tried to esh out one way a simple predictive mechanism could simultaneously explain both of these effects.

Taken together, these data support the idea that perceptual adaptations go beyond being temporary adjustments to unusual environments and can have important developmental consequences. Indeed, this is the real stuff of development. Long term effects are achieved as the accumulated result of many small tweaks to the system occurring on a situation-by-situation basis. Thus, understanding how the perceptual system can and does adjust in short time windows to speci c situations should be useful towards increasing our understanding of the type of lasting changes that the architecture can achieve. These results are admittedly only a rst step in that understanding, but they do indicate several directions for future work in this regard.

Acknowledgements

Thank you to Linda Smith, Michael Gasser, Robert Goldstone and Robert Port for their feedback on this research.

References

Anstis, S., Verstraten, F. A. J., & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences*, *2*(3), 111–117.

Changizi, M. A., Nijhawan, R., Kanai, R., & Shimojo, S. (2003). Perceiving-the-present and a general theory of illusions of projected size, projected speed, luminance contrast and distance. (Under Review. Found online on 12/1/2003 at URL http://www.geocities.com/changizi/pp3.pdf.)

Changizi, M. A., & Widders, D. (2002). Latency correction explains the classical geometrical illusions. *Perception*, *31*, 1241–1262.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Freyd, J. J. (1992). Dynamic representation guiding adaptive behavior. In *Time, action, and cognition: Toward bridging the gap.* Dordrecht, The Netherlands: Kluwer Academic Press.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*, 225–241.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Helmholtz, H. L. F. von. (1866). Concerning the perceptions in general. In *Treatise on physiological optics* (Vol. III, pp. 1–37). (Translated by J. P. C. Southall, 1925, *Optical Society of America*. Reprinted in 1962, New York: Dover Publications, Inc.)

Hubbard, T. L. (1999). How consequences of physical principles influence mental representation: The environmental invariants hypothesis. In *Fechner Day '99: The end of 20th century psychophysics — Proceedings of the 15th annual meeting of the International Society for Psychophysics.* Tempe, AZ, USA: International Society for Psychophysics.

James, W. (1890). The Perception of Time. In *The Principles of Psychology* (Vol. 1, pp. 605–642). Cambridge, MA, USA: Harvard University Press.

Smith, L. B., & Evans, P. (1989). Similarity, identity, and dimensions: Perceptual classification in children and adults. In *Object perception: Structure and process.* Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

[6] These ideas have a long history in psychology, going back through Gibson (1979) at least as far as James (1890) and Helmholtz (1866).

Time is of the Essence: Processing Temporal Connectives During Reading

John C. J. Hoeks (j.c.j.hoeks@let.rug.nl)
BCN NeuroImaging Centre, PO Box 716,
9700 AS Groningen, The Netherlands.

Laurie A. Stowe (l.a.stowe@let.rug.nl)
BCN NeuroImaging Centre, PO Box 716,
9700 AS Groningen, The Netherlands.

Charlotte Wunderink (c.wunderink@let.rug.nl)
BCN NeuroImaging Centre, PO Box 716,
9700 AS Groningen, The Netherlands.

Abstract

An important study by Münte, Schiltz, and Kutas [Nature *395* (1998) 71-73] using ERPs (=Event-Related brain Potentials) suggested that sentences starting with the temporal connective *before* are more taxing for working memory than sentences starting with *after*, as evidenced by a slow negative shift for *before* sentences. According to Münte et al., *before* sentences present events out of the correct chronological order, as in *Before the author submitted the paper* [=second event], *the journal changed its policy* [=first event]. In order to come up with the correct discourse representation of the sentence, the correct chronological order has to be restored, leading to extra memory load. In the present experiments using a self-paced reading paradigm it will be shown that *before* sentences are not more difficult to process than *after* sentences, but that they are even read faster than *after* sentences. In addition, it is shown that *before* sentences in which events are presented in the *correct* chronological order, as in *The journal changed its policy* [=first event], *before the author submitted the paper* [=second event] are read more slowly than corresponding sentences with *after*. Implications for Münte et al.'s theory are discussed and objectives for future research are formulated.

Introduction

Readers do not wait with the interpretation of a sentence until they have received the final word. On the contrary, the process of understanding sentences occurs in a highly incremental fashion, approximately as each word is encountered (e.g., Altmann & Steedman, 1988). A striking illustration of this phenomenon is provided by Münte, Schiltz, and Kutas (1998). In a study using ERPs (=Event-Related brain Potentials), they showed that sentences starting with the temporal connective *before* were processed differently from sentences starting with *after*. Almost immediately after presentation of the temporal connective, the ERP waveforms for the *before* and the *after* sentences started to diverge, with the more negative values for the *before* sentences. Münte et al. argued that this negative shift reflected the additional discourse-level processing that is necessary to deal with sentences that present events out of their correct chronological order. Consider, for instance, sentence 1a, which is an example sentence of the materials used by Münte et al. (1998).

1a. Before the author submitted the paper, the journal changed its policy.

Here, the event of submitting a paper precedes the event of policy change in this specific *sentence*, but in *reality*, the policy change happened first, which can be described as `Before [Event2], [Event1].` In contrast, sentences starting with *after*, such as 1b, present the events in their correct chronological order, exactly as they purportedly happened in reality: first a submission, then change of policy, so: `After [Event1], [Event2].`

1b. After the author submitted the paper, the journal changed its policy.

As the size of the negative shift in the ERP waveforms turned out to be highly correlated with the individual working memory spans of the participants (the higher the memory span, the larger the effect), Münte et al. concluded that the problem with *before* sentences is really a working memory problem. In other words, it is claimed that when reading a sentence starting with *before*, readers immediately realize that the events that they are going to read about will have to be re-ordered at some stage to arrive at a coherent and valid semantic representation of the sentence. Thus, the temporal connective *before* may act as a kind of cognitive operator instructing the language processor to hold in memory the event reported on in the first clause, in order to enable the reconstruction of the events in their correct chronological order, presumably after the sentence has been read.

There are, however, a number of problems with this interpretation. First of all, there is no *a priori* reason to interpret a *negative* shift as evidence for processing difficulty or any other form of effortful (memory-related) processing. For instance, a well-known ERP component such as the 'P600' is a *positive* component (occurring about 600 ms post-onset of a critical stimulus), which can be evoked by a number of syntactic problems, such as ungrammatical sentences (Osterhout & Holcomb, 1992), correct sentences with an unpreferred syntactic structure

(Hagoort, Brown, & Groothusen, 1993), syntactically complex sentences (Kaan, Harris, Gibson, & Holcomb, 2000), and even in sentences with a correct syntactic structure that are semantically anomalous (Hoeks, Stowe, & Doedens, 2004). In other words, it cannot be excluded that for some reason or other the *after* sentences are the most difficult, and that this processing difficulty is reflected as a *positive* shift in the ERP signal.

But even if the negativity does indicate processing problems, and the *before* sentences are actually the most difficult, there is another reason why the interpretation of Münte et al might be wrong. For example, if we take a look at the data that Münte et al. provide on the participants with the high and the low working memory score, visual inspection of the waveforms suggests that the ERPs for the *before* sentences actually do not differ between the two working memory groups. Shouldn't these *before* sentences be extra taxing for the group with the smallest working memory capacity, as compared to the high working memory group? What we see instead is a difference between the groups for the *after* sentences, which are more *positive* for the group with the high working memory score. This is quite unexpected, given that the *after* sentences are relatively 'easy' and do not tax memory at all, at least much less than *before* sentences, as Münte et al. claim. In addition, there is only a very slight difference in the low working memory group between the the 'difficult' *before* sentences and 'easy' *after* sentences, which is also rather unexpected. It is not immediately clear how this pattern of results should be interpreted, but it is clear that it does not support Münte et al.'s hypotheses.

In the light of these problematic aspects it seems necessary that two specific issues regarding the processing of sentences with temporal connectives be resolved. First, it is very important to find out whether *before* sentences are more difficult than *after* sentences, or whether it is the other way around. Once this is known, we also know how to interpret the negative shift for *before* sentences reported by Münte et al. Indeed, we might be looking at a *positive* shift, if the *after* sentences turn out to be the most difficult. Secondly, if *before* sentences are more difficult than *after* sentences, we should be able to establish whether this is caused by the chronological order of the events described in the sentence or perhaps by other factors. In the present experiment we will focus on exactly those issues using a self-paced reading paradigm.

The first issue can be tackled rather straightforwardly: by measuring the time people take to read the sentence in either the *before* or the *after* version, we can establish which condition is the most difficult, as it will be read more slowly. The second issue is more complicated, but can be investigated in the following way. Consider sentence 2a, which is an example sentence from the present experiment (with English translation in brackets).

2a. Voordat Piet de sinas dronk, at Stefan de koekjes op.
(Before Piet drank the soft drink, Stefan ate the biscuits)

This sentence presents the events out of chronological order, as did sentence 1a. The 'drinking' event which is mentioned first, actually happened later than the 'eating' event. However, in a sentence such as 2b, the events are presented in their chronological order again.

2b. Stefan at de koekjes op, voordat Piet de sinas dronk.
(Stefan ate the biscuits, before Piet drank the soft drink)

Thus, sentence 2b should not be problematic at all, and be processed faster than a similar sentence with *after* in the second clause (e.g., Stefan ate the biscuits [event2], after Piet drank the soft drink [event1]).

Experiment 1

This experiment is a reading time experiment in which participants read sentences for comprehension and made semantic plausibility judgments after reading each sentence.

Method

Participants Forty native speakers of Dutch were paid for participating in this experiment (28 female; mean age 21 years, age range 18-30). All were currently receiving a university education.

Materials & Design For this experiment, 80 sets of sentences were constructed, each set consisting of eight versions of a given item. Experimental lists were constructed with 10 experimental items per condition, and no list containing more than one version of a given item. All 80 experimental sentences were plausible as determined by two expert raters. An equal number of implausible filler sentences (see sentence 4 below for an example) were added such that each list contain an equal number of plausible and implausible items. The purpose of the semantic plausibility test and the implausible fillers was to encourage deep semantic processing of the experimental sentences.

The order in which experimental and filler items appeared was determined semi-randomly and was the same for each list. Each list was presented to an equal number of participants (i.e., five) and each participant saw one list. Only the first four of the eight conditions belong to the present Experiment 1; the other four conditions were part of a related experiment that will be discussed below as Experiment 2. The experimental sentences for the first experiment appeared in the following forms:

3a. **Before (first clause), Incorrect order (E2 - E1)**
Before Piet drank the soft drink [E2], Stefan ate the biscuits [E1].

3b. **After (first clause), Correct order (E1 - E2)**
After Piet drank the soft drink [E1], Stefan ate the biscuits [E2].

3c. **Before (second clause), Correct order (E1 - E2)**
Stefan ate the biscuits [E1], before Piet drank the soft drink [E2].

3d. **After (second clause), Incorrect order (E2 - E1)**
Stefan ate the biscuits [E2], after Piet drank the soft drink [E1].

The filler sentences had exactly the same form as the experimental sentences (in exactly the same quantities) but were semantically implausible, as sentence 4.

4. Before the murder was committed, the police found the dead body.

A practice session consisting of 30 items preceded the actual experiment.

Procedure Participants were seated behind a computer screen in a sound-proof cabin. Each sentence was preceded by an asterisk indicating the start of a new sentence. Participants were instructed to use the 'b'-key on a keyboard before them to read the sentence clause-by-clause. That is, after the first key-press the asterisk disappeared and the first clause appeared (e.g., "Before Piet drank the soft drink,"); after the second press the first clause disappeared and the second clause appeared (e.g., "Stefan ate the biscuits."); at the next press the second clause disappeared and the question "Goed?" (="Correct?") appeared. Participants had to press the right SHIFT button to indicate that the sentence was semantically plausible, and the left SHIFT button if they felt it was not. Each response was followed by feedback on the correctness of the answer (i.e., "Correct!" / "Wrong!"). Participants were asked to read the sentences carefully and to respond as quickly as possible without compromising accuracy. After the feedback the asterisk reappeared. In all, the experiment took approximately 20 min.

Results

Analysis First, reading time data were screened for outliers. Reading times less than 200 or greater than 4000 ms were excluded. After that, all observations were excluded which deviated more than 2.5 *SD*s from either the participant or the item mean of each clause in each condition. Two analyses were performed: an *F1*-ANOVA on the condition means for each participant and an *F2*-ANOVA on the condition means for each item. The factors Temporal Connective (*before* vs. *after*), Connective Position (in first clause vs. in second clause), and Clause (first clause vs. second clause) were treated as within-participants and within-items factors. In the participant-based analyses, the factor List (i.e., grouping together participants that were presented with the same list) was also included in the analyses as a between-participant factor, and in the item-analyses the factor Itemgroup (i.e., grouping together items that appeared in the same condition in each list) was entered as a between-items factor. Both factors had 8 levels as there were 8 lists and 8 itemgroups (see design section above). In addition, accuracy percentages were calculated per condition. Mean reading times and accuracy are presented graphically in Figure 1.

Accuracy As can be seen in Figure 1, accuracy was high for each condition (overall accuracy 90 %). No significant interactions or main effects were found (all F-values < 1).

Reading Times The 3-way interaction between Temporal Connective x Connective Position x Clause was significant in the analysis on items (F2(1,72)=4.12; p<.05) and marginally significant in the participant analysis (F1(1,32)=3.23; p=.08). Post Hoc analyses showed that there was no significant difference between *before* and *after* sentences as far as the first clause is concerned (though reading times of first clauses containing *before* were numerically smaller than those of first clauses containing *after*). Much larger differences were found in the second clause. The second clause of sentences with *before* in the first clause was read significantly faster than of sentences with *after* in the first clause (p<.05). The opposite pattern, however, was found for the sentences with the temporal connective in the second clause: here, the *before* sentences were read more slowly than the *after* sentences, though this difference was only marginally significant (p=.09).

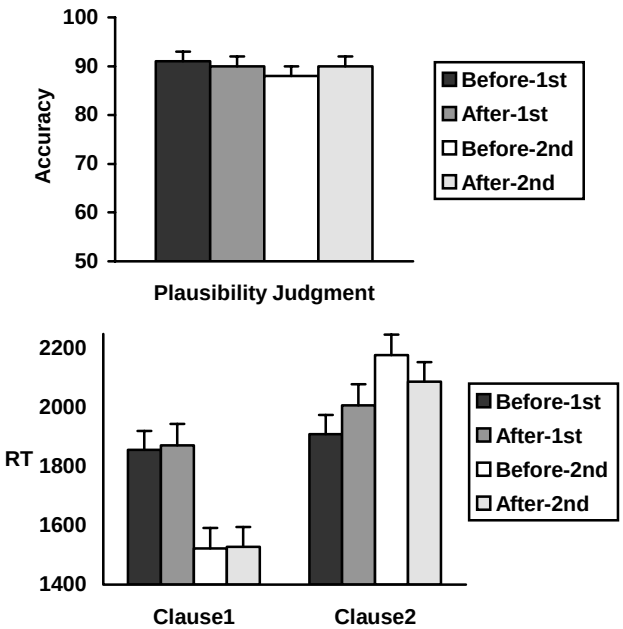


Figure 1. Accuracy (in percentages, upper panel) and Reading times (in ms, lower panel) for Experiment 1. "Before/After-1st"=temporal connective in first clause; "Before/After-2nd"= connective in second clause.

A number of other effects were significant, which should of course be interpreted with caution in the presence of the significant 3-way interaction. For instance, the 2-way interaction of Connective Position and Clause was significant (F1(1,32)=107.16; p<.0001; F2(1,72)=87.80; p<.0001), reflecting longer reading times for the clause in which the temporal connective was present (connective in first clause: first clause: 1865 ms, second clause: 1525 ms; connective in second clause: first clause: 1959 ms, second clause: 2138 ms). More interestingly, there was also an interaction between Connective Position and Temporal Connective (F1(1,32)=5.06; p<.05; F2(1,72)=4.80; p<.05), indicating that *before* sentences as a whole were read faster when the temporal connective appeared in the first clause (*before*: 1884 ms; *after*: 1940 ms) than when it appeared in the second clause (*before*: 1854 ms; *after*: 1808 ms). No other effects concerning Temporal Connective were significant (all F-values < 1). The factors Clause (Clause 1: 1695 ms; Clause 2: 2048 ms) and Connective Position (connective in first clause: 1912 ms; in second clause: 1831

ms) had significant main effects ($F1(1,32)=34.80$; $p<.0001$; $F2(1,72)=84.98$; $p<.0001$, and $F1(1,32)=10.08$; $p<.005$; $F2(1,72)=13.69$; $p<.0001$), respectively).

Discussion

This experiment yielded two important results. First, in the conditions where the temporal connective appeared in the first clause, there was no evidence at all for *before* sentences being more difficult than *after* sentences as expected on the basis of Münte et al.'s arguments (1998). Quite on the contrary, the first clause of *before* sentences was read numerically faster than the first clause of *after* sentences. More importantly, the second clause of *before* sentences was read *significantly faster* than that of *after* sentences, with an average advantage of 97 ms for the *before* sentences. This finding clearly indicates that *before* sentences are in fact easier to process than *after* sentences, contra Münte et al.'s predictions. So perhaps the slow negative shift for the *before* sentences is actually a slow *positive* shift for the more difficult *after* sentences.

The second important result comes from the conditions where the temporal connective was placed in the second clause. Here, we see no difference in reading times in the first clause, which is as one would expect given that there is no difference between the conditions yet, as the temporal connective only appears in the second clause. We do see substantial differences in the second clause, but in a direction opposite to Münte et al.'s predictions. Recall that the *before* sentences with the temporal connective in the second clause present the events in the correct chronological order (see sentence 3c), in contrast to sentences with *after* in the second clause (see sentence 3d). This should have solved the problems of increased memory load and thus have led to a processing advantage for the *before* sentences as compared to the *after* sentences. However, what we find is a 110 ms *dis*advantage for *before* sentences with the events in the 'correct' temporal order. This strongly suggests that presenting events out of chronological order does not lead to processing difficulty. It even seems that presenting events in the correct chronological order leads to an increase in processing difficulty.

Summarizing, this experiment showed, contrary to expectation, 1) that sentences starting with *before* are easier to process than sentences starting with *after*, and 2) that presenting events out of chronological order does not cause processing difficulty.

Experiment 2

Experiment 1 was intended to answer two straightforward questions about the processing of temporal connectives: 1) are *before* sentences more difficult than *after* sentences, and 2) is that the case because before sentences present events out of chronological order? We have seen that neither one was true. Experiment 2 was more exploratory, focussing on the possible interaction between *temporality*, or the chronological ordering of events, with the processing of *referential expressions* which is another important aspect in the construction of a coherent discourse representation (Garnham, 1999). The main issue here is whether temporal and referential processing draw on the same resources, or whether they are processed in parallel by independent mechanisms.

It is assumed that the use of referential elements such as pronouns (e.g., 'he' or 'she') in a sentence may increase the processing load during comprehension. When a pronoun is encountered, a search process needs to be initiated in order to find the intended *referent* for the pronoun (i.e., the person or thing that is referred to by the pronoun). It has been shown that this search process can be more costly than for instance having a proper name (e.g., *Stefan*) where no search process is necessary (Streb, Rösler, & Hennighausen, 1999).

In Experiment 2, the materials from Experiment 1 were used, except that in each sentence a pronoun was inserted, as in sentence 5a.

5a. Stefan at de koekjes op, voordat hij de sinas dronk.
(Stefan ate the biscuits, before he drank the soft drink)

In this sentence, the pronoun *he* appears in the second clause and is used *anaphorically*, that is, it refers back to an entity that is mentioned earlier (in this case *Stefan*). Although there is a pronoun present that might induce a search process, it does not seem likely that in this specific case this search process is in any way difficult. In fact, in a sentence such as 5a there is only one possible referent (i.e., *Stefan*) and the use of a proper name, permitting immediate identification of the referent, would even be sub-optimal (see e.g., Gordon, Grosz, & Gilliom, 1993, regarding a phenomenon called the 'repeated name penalty'). However, if the pronoun were to *precede* its referent, this might be taxing for working memory, or lead to other processing difficulty, because it will not be possible to fully process the clause that contains this pronoun before the referent is known. Consider 5b, for an example of such a sentence.

5b. Voordat hij de sinas dronk, at Stefan de koekjes op.
(Before he drank the soft drink, Stefan ate the biscuits)

In this sentence, the pronoun is used as a *cataphor*, or *backwards anaphor* (Garnham, 1999), and refers to an entity that will be mentioned later. Because the first clause cannot be fully interpreted as it lacks crucial information on whom the pronoun refers to, and because the reader does not know when this referent will be presented, it seems reasonable to assume that these sentences are difficult to process, as compared to sentences such as 5a. If this kind of effortful processing of cataphors is handled by the same mechanism that is responsible for temporal processing one would expect an interaction between these two factors. If, on the other hand, these two kinds of processes proceed in parallel and are carried out by independent systems, then there will be no interaction.

In summary then, Experiment 2 aims to clarify two things:

1) whether cataphoric constructions are more difficult to process than anaphoric ones, and

2) whether this difference in processing load has a (possibly differential) effect on how *before* and *after* sentences are handled. In other words, do temporal and referential processes interact?

Method

Participants, Design, & Procedure Participants, design and procedure are described in the method section of Experiment 1.

Materials Most aspects of the materials are described above in the materials section of Experiment 1, except for the conditions of the current experiment, which were the following:

6a. **Before (first clause), Incorrect order, cataphor**
Before he drank the soft drink [E2], Stefan ate the biscuits [E1].

6b. **After (first clause), Correct order, cataphor**
After he drank the soft drink [E1], Stefan ate the biscuits [E2].

6c. **Before (second clause), Correct order, anaphor**
Stefan ate the biscuits [E1], before he drank the soft drink [E2].

6d. **After (second clause), Incorrect order, anaphor**
Stefan ate the biscuits [E2], after he drank the soft drink [E1].

Note that a full factorial design is not possible, as sentences with a pronoun in the first clause and a connective in the second clause are ungrammatical when the pronoun is intended to refer to the entity mentioned in the second clause (as in: "$He_{(i)}$ drank the soft drink, before $Stefan_{(i)}$ ate the biscuits"). Instead, a reduced design was chosen that would enable us to answer some important questions regarding the interaction of temporal and referential processing.

Results

Analysis After screening for outliers (see Experiment 1), mean RTs and mean accuracy percentages in each condition were calculated for both participants and items. Figure 2 presents the mean reading times and accuracy for Experiment 2.

Accuracy As can be seen in Figure 2, accuracy was high for each condition (overall accuracy 87 %). The main effect of Temporal Connective was marginally significant in the participant-based analysis ($F1(1,32)=2.80$; p=.10), but not in the item-based analysis ($F2(1,72)=1.74$; p=.19), indicating a trend for slightly greater accuracy in the *before* sentences (89%) as compared to the *after* sentences (86%). There was no significant effect of Connective Position (F-values < 1).

Reading Times The 3-way interaction between Temporal Connective x Connective Position x Clause was not significant in the present experiment (both F-values < 1). The interaction of Temporal Connective and Clause was significant in both participant-based and item-based analyses (F1(1,32)=5.89; p<.05; F2(1,72)=8.95; p<.005). This interaction is caused by *before* sentences taking longer than *after* sentences in the first clause (i.e., 1558 vs. 1519 ms, respectively) with a reverse pattern for the second clause (i.e., *before*: 1917 ms vs. *after*: 1999 ms). None of these two separate contrasts were significant, however (p-values>.20). Perhaps more importantly, there was also a trend towards an interaction between Temporal Connective and Connective Position (F1(1,32)=3.07; p=.09; F2(1,72)=1.67; p=.20), suggestive of shorter times for *before* sentences as a whole when the temporal connective appears in the first clause (i.e., *before*: 1704 ms vs. *after*: 1762 ms), contrasting with the pattern of results when the temporal connective appears in the second clause (i.e., *before*: 1771 ms vs. *after*: 1756 ms).

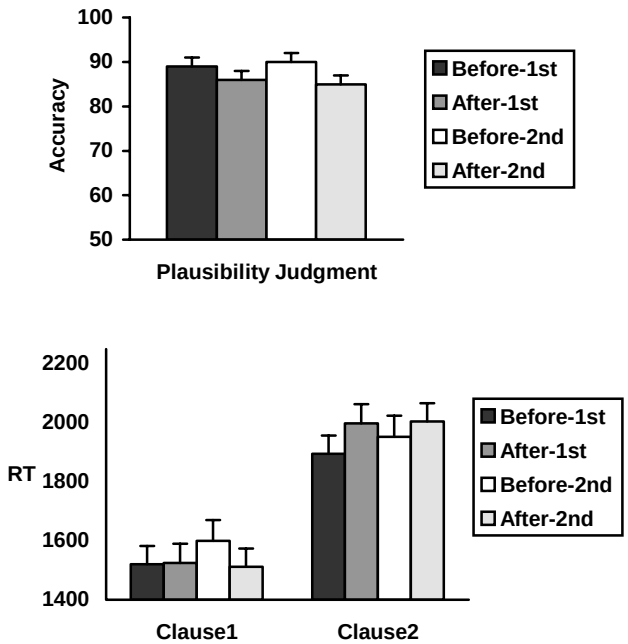


Figure 2. Accuracy (in percentages, upper panel) and Reading times (in ms, lower panel) for Experiment 2. "Before/After-1st"=temporal connective in first clause, pronoun is cataphor; "Before/After-2nd"=temporal connective in second clause, pronoun is anaphor.

As to the main effects, there was a marginally significant main effect of Connective Position (F1(1,32)=1.55; p=.22; F2(1,72)=3.06; p=.09), suggesting that sentences take longer to read when the temporal connective appears in the second clause than when it is present in the first clause (i.e., 1764 ms vs. 1733 ms). Finally, there was a significant main effect of Clause (F1(1,32)=56.37; p<.0001; F2(1,72)=195.57; p<.0001), reflecting the large difference in reading times between first clause (1539 ms) and second clause (1958 ms). There was no main effect of Temporal Connective (*before*: 1737 ms; *after*: 1759 ms; p-values >.30).

Discussion

The aim of this experiment was to establish whether cataphoric constructions were more difficult to process than anaphoric ones, and also whether any difference between them would affect the processing of temporal structure. Some tentative evidence for such an interaction seems to come from the finding that *before* sentences are faster than

after sentences when containing cataphors, but not when containing anaphors. However, this might very well be the result of an unfortunate 'blip' in the data, that is, the fact that reading times for the first clause differ between conditions that had identical first clauses (i.e., of sentences with the connective in the second clause). It is not unlikely either that the marginally significant main effect of Connective Position, another indication of a possible difference between cataphors and anaphors, is caused by just that spurious effect. So what can we say about cataphors and anaphors then?

What we can say about cataphors is that they do not seem to be hard to process. Perhaps the most striking difference between the present two experiments is the large reduction in first clause reading times when proper names / NPs (i.e., in Exp. 1) are replaced by cataphoric pronouns (i.e., in Exp. 2), indicating that inserting cataphoric pronouns makes sentences easier. However, because these clauses differ between experiments in the lexical material they contain, no strong conclusions can be drawn from this outcome. What is equally apparent, however, is that cataphors do not *create* a difference between *before* and *after* sentences: there is no difference at all in the first clause and only a slight difference in the second clause, which is numerically almost identical to the pattern of results in Experiment 1 (for sentences with the temporal connective in the first clause, see also Figure 1). As for anaphors, it seems clear that they do not cause processing difficulty either. On the contrary, they seem to make the processing of *before* sentences easier, if we compare the results of both experiments: In Experiment 1 *before* (with the correct order of events) was read more slowly than *after* (with the incorrect order of events) in the sentences with the connective in the second clause; in Experiment 2 this is (numerically at least) the other way around. In summary then, the outcome of this experiment strongly suggests that cataphors are not difficult to process, that anaphors are even easier, and that chronological order of events is not a factor of importance.

Conclusion & Future Directions

The present experiments have convincingly shown that *before* sentences are actually *easier* to process than *after* sentences. In addition, it has become clear that the chronological order of events does not strongly influence ease of sentence processing. Finally, as there was no strong evidence for an interaction between temporal and referential processing, it is still a bit unclear whether these two kinds of information are processed by the same or by different cognitive mechanisms.

When we try to understand the Münte et al. results in the light of the present findings, we must conclude that the slow potential difference building up while the sentence is read should not necessarily be interpreted as a negativity for the *before* sentences, but rather as a positivity for the *after* sentences. In addition, this slow wave difference does not seem to be related, or at least not directly, to presenting events in or out of their correct chronological order, nor with memory processes *per se* (recall that the low working memory group from Münte et al. did not show a difference between *before* and *after* sentences). This leaves us with a lot of new questions: why are *after* sentences more difficult to process than *before* sentences? and how should we then conceive of the relationship between working memory capacity and temporal processing, if it does not work as Münte et al. hypothesized? It is possible that connectives (and also pronouns) evoke certain processing strategies that do not tax memory, or only minimally. Hoeks and Stowe (2002), for instance, have speculated that *before* might activate a relatively cost-free 'temporal ordering frame' (maybe only available for individuals with high working memory capacity?) that allows for fast sentence integration, whereas *after* does not. More research focussing on these processing aspects of temporal connectives is definitely needed. But also research using language corpora in order to establish both form and function of different kinds of temporal expressions in text and communication.

Acknowledgements

We gratefully acknowledge Ingeborg Prinsen for her help with the construction of the materials and the programming of the experiment.

References

Altmann, G. T. M., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30, 191-238.

Garnham, A. (1999). Reference and anaphora. In: S. Garrod & M. Pickering (Eds.), *Language processing* (pp. 335-362). Hove, UK: Psychology Press.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science,* 17, 311-347.

Hagoort, P., Brown, C.M., & Groothusen, J. (1993). The Syntactic Positive Shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439-483.

Hoeks, J.C.J., & Stowe, L.A. (2002). *Temporal processing and sentential complexity.* Poster presented at AMLaP 2002.

Hoeks, J.C.J., Stowe, L.A., & Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research,19,* 59-73 .

Kaan, E., Harris, A., Gibson, E., & Holcomb, P.J. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes, 15,* 159-201.

Münte, Th.F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature, 395,* 71-73.

Osterhout, L. & Holcomb, P. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31, 785-806.

Streb, J., Rösler, F., & Hennighausen, E. (1999). Event-related responses to pronoun and proper name anaphors in parallel and nonparallel discourse structures. *Brain and Language, 70,* 273-286.

Learning Predictive Models of Memory Landmarks

Eric Horvitz (horvitz@microsoft.com)
Microsoft Research, One Microsoft Way
Redmond, WA 98052 USA

Susan Dumais (sdumais@microsoft.com)
Microsoft Research, One Microsoft Way
Redmond, WA 98052 USA

Paul Koch (paulkoch@microsoft.com)
Microsoft Research, One Microsoft Way
Redmond, WA 98052 USA

Abstract

We describe the construction of statistical models that provide inferences about the probability that subjects will consider events to be memory landmarks. We review methods and report results of experiments probing the classification accuracy and receiver-operator characteristics of the models. Then, we discuss opportunities for integrating models of memory landmarks into computing applications, and present a prototype time-line oriented content browsing tool.

Introduction

Studies of memory support the assertion that people make use of special landmarks or anchor events for guiding recall (Shum, 1994; Smith, 1979; Smith, Glenberg & Bjork, 1978)) and for remembering relationships among events (Davies & Thomson, 1988; Huttenlocher & Prohaska, 1997). Such landmarks include both public and autobiographical events. More generally, there has been significant study and modeling of episodic memory, where memories are considered to be organized by *episodes* of significant events, including such information as the location of an event, attendees, and information about events that occurred before, during, and after each memorable event (Tulving, 1983; Tulving & Thomson, 1980). Memory has been shown to also depend on the reinstatement of not only item-specific contexts, but also on more general context capturing the situation surrounding events.

We believe that automated inferences about important memory landmarks could provide the basis for new kinds of personalized computer applications and services. Rather than focusing on specific machinery proposed as models for recall (*e.g.,* Malmberg, Steyvers, Stephens, *et al*.,. 2002;. Raaijmakers & Shiffrin, 2002; Shiffrin & Steyvers, 1997), we set out to investigate the feasibility of directly learning models of memory landmarks via supervised learning. We focus here on the construction, testing, and application of predictive models of memory landmarks, based on events drawn from users' online calendars.

We first review experiments with the construction of personalized models of memory landmarks. We describe how we construct models that can be used to infer the likelihood that events will serve as memory landmarks, reviewing the extraction of data from subjects' online calendars, the collection of assessments about landmarks with tools that enable subjects to label their calendar events, and the learning of models via Bayesian learning procedures. After reviewing the performance of the models, we describe, as a sample direction for the use of predictive models of memory landmarks in computing applications, a prototype, named MemoryLens Browser. MemoryLens Browser employs the inferences about landmarks in visualizations for browsing files and appointments. Finally, we review research directions aimed at enhancing coverage and discriminatory power of models of memory landmarks.

Accessing Events and Event Properties

We will focus on the construction of models of memory landmarks derived from users' online calendar information. Electronic encodings of calendars provide rich sources of data about events in users' lives. People who rely on electronic calendars, often encode multiple types of events in an online format. Such items include appointments, holidays, and periods of time marked to indicate such activities as travel and vacation. In large enterprises that rely on computer-based calendaring systems, appointments and events are typically formulated, accepted, displayed and managed via schemas capturing multiple properties of the events.

We developed a calendar event crawler that works with the Microsoft Outlook messaging and appointment management system. The crawler analyzes a user's online calendar to create a case library of events and properties associated with each event. The calendar crawler extracts approximately 30 properties for each event. Most of these properties are obtained directly from the online data and metadata stored for events. These properties include the *time of day* and *day of week* of events, *event duration*, *subject*, *location*, *organizer*, *number of invitees, relationships between the user and invitees*, the *role of the user* (*i.e.,* user was the organizer, a required invitee, or an optional invitee), *response status* of the user to appointment invitations (*i.e.,*

user responded *yes*, responded *tentative*, *no response*, or *no response request made*), whether the meeting is *recurrent* or *not recurrent*, whether the time is marked *as busy or free* on the user's calendar, and the nature of the *inviting email alias*—the alias used to send the meeting invitation.

In addition to properties in the database schema employed by Outlook, a subsystem of the crawler accesses the Microsoft Active Directory Service to identify organizational relationships among the user, the organizer, and the invitees, noting for example, whether the organizer and attendees are *organizational peers*, *direct reports*, *managers*, or *managers of the user's manager*.

Beyond the use of data from Outlook and Active Directory Services, we created several derived properties representing statistics about atypical situations, based on the intuition that rare contexts might be more memorable than common ones. In particular, we developed procedures for computing *atypical organizers, atypical attendees,* and *atypical locations.* We compute a measure of the rarity for these properties of events by considering the portion of all meetings over all events under consideration or for a fixed period of time (*e.g.,* events over a year) in which the property under consideration has the same value it has in the event at hand. For the studies reported here, we computed atypicality based on all events under consideration.

To compute the value of *location atypia* for events, we first compute the number of times each location has appeared in a user's calendar over a fixed period. The system then discretizes the *location atypia* variable into a set of states, capturing a range of percentiles, and the location atypia variable for each event acquires a particular value based on the rarity of the location associated with that event.

An analogous derivation is used for computing *organizer atypia* and *attendee atypia*. For these variables, all people attending all of the appointments for the fixed period under consideration are analyzed, and the portion of a subject's appointments attended or organized respectively by each attendee is noted. A meeting acquires the *organizer atypia* or *meeting atypia* value associated with the least frequent attendee or organizer of the meeting.

Building Models of Memory Landmarks

We recruited 5 participants from our organization for data collection and tagging. We asked the subjects to review a list of all of the appointments, holidays, and other annotations stored in their calendars that were extracted automatically by a calendar crawler, and to identify the subset of events that they viewed as serving as salient, memory landmarks. More specifically, we directed the subjects to do the following:

> *Please review the events on your calendar and identify those events that would serve as key memory landmarks on a timeline of events for such purposes as searching for files and appointments.*

Each subject downloaded software components and executed the event-collection program to crawl their calendars and to create a case library of labeled data. The cases typically spanned several years of presentations, trips, meetings, tasks, and holidays, and included several thousand items. We provided subjects with a memory-landmark assessment tool that lists events drawn from their online calendar within a scrollable window, ranked from most recent to most distant events. The tool provides fields, adjacent to each event, that subjects use to label items as landmark or non-landmark events.

We pursued the construction of predictive models of memory landmarks from the supervised training data. We elected to employ Bayesian-network learning methods so as to have the ability to visually inspect key probabilistic dependencies among variables and, in particular, to understand key variables and states of variables influencing the likelihood of events being called memory landmarks.

We partitioned the data into training and testing cases, with an 80/20 split; that is, we built the models for each individual using 80% of their labeled data and evaluated the learned model on the remaining 20% of the labeled data. We employed a Bayesian structure-search procedure, developed by Chickering, Heckerman & Meek (1997), to build Bayesian-network models for event landmarks for each subject The procedure employs a greedy search through a large space of dependency structures and computes, for each plausible dependency structure, an approximation for the likelihood of the data given the structure. A model score is computed as a function of this likelihood and a model-prior parameter that penalizes for complexity. The model with the highest score is selected.

We optimized the model-prior parameter by splitting the training set 80/20 into subtraining and subtesting data sets, respectively, and identifying a soft peak in the Bayesian score. This value of the parameter at the soft peak was used to build the model from the full training set.

We inspected the predictive models constructed for each subject, noting dependencies among key variables, the discriminatory power of variables, and classification accuracy of the models at predicting the data held out from the training procedure.

Figure 1 displays a Bayesian network built from the data from one of the participants in the study (subject S1), showing all of the variables and the dependencies among them. A sensitivity analysis demonstrated that key influencing variables in this model for discriminating whether an event is a memory landmark are the *Subject, Location string, Meeting sender, Meeting organizer, Attendees,* and whether the meeting is *Recurrent*.

We explored the strength of dependencies for variables in the model for each subject and found similar influences of key variables across subjects. For subjects in our study, *atypically long durations*, *non-recurrence of events*, a user *flagging a meeting as busy* or *out of office*, and *atypical locations* or *special locations* had significant influence on the inferred probability that events would be considered a landmark event. We found that meetings marked as *recurrent* meetings rarely served as memory landmarks.

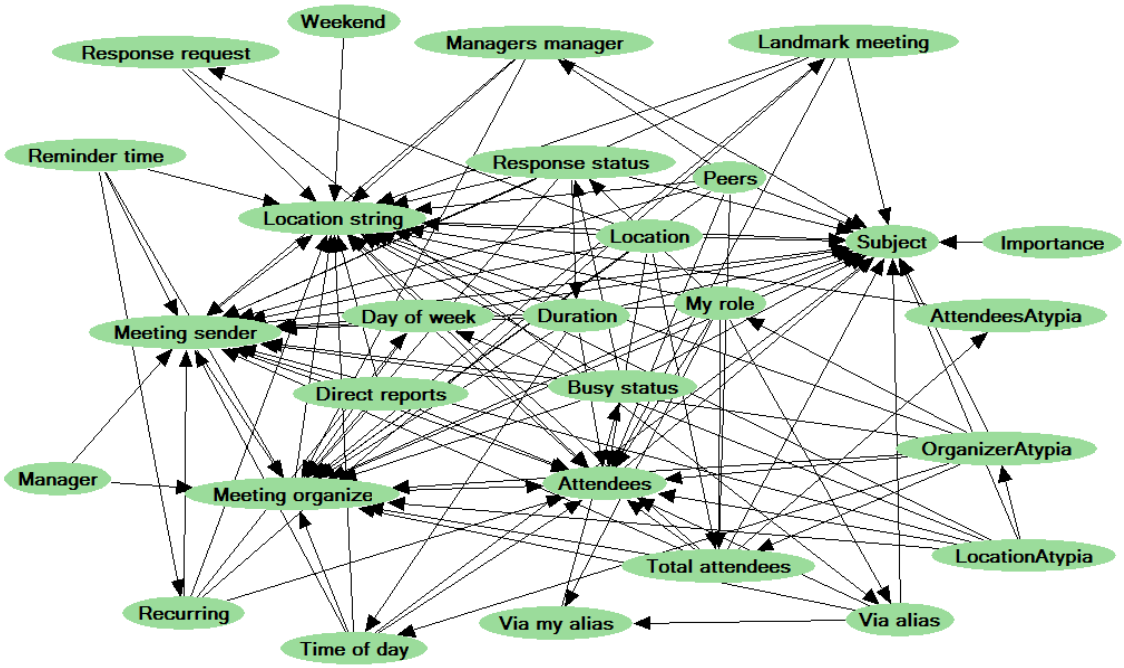


Figure 1: Bayesian network learned from online calendar data (subject S1) showing dependencies among event properties and likelihood that an event will be considered a memory landmark by a subject.

Table 1 shows the classification accuracies of the learned models of landmarks. For each test case, the values of the properties of the event are identified (or computed for derived properties) and then input to the model which provides a probability that the event is a memory landmark. That is, we compute *p*(*Event selected as a memory landmark*|*E*), given evidence *E*—the multiple properties of associated with each event on the subject's calendar. The models range in classification accuracies for the five subjects from 0.78 to 0.95.

In addition to looking at overall classification accuracies, we swept out receiver-operator (ROC) curves to visualize the relationship between false negatives and false positives at different thresholds for admitting events as memory landmarks. The false-positive rate is varied by changing the threshold of the probability score that is required for classifying an event as a memorable landmark, and the corresponding false negative rate is noted. The curves for the subjects in the study are displayed in Figure 2. We note that the ROC curves show a trend toward lower false positives and false negatives with increases in the size of the training sets.

The ROC curves are particularly important for understanding the value of employing such predictive models of memory landmarks in computing applications. As we shall explore in the next section, one class of computing applications centers on the use of a user-controlled threshold on the probability of events used to identify landmark events. In such uses of predictive models of landmarks, users may be given the ability to define, *e.g.,* via a slider control, the subset of all events that should be admitted, say, for displaying within a rendering of a timeline of events. Such timelines could provide useful "memory backbones" when searching for content in a large personal store. Models for inferring the likelihood that events will serve as memory landmarks promise to endow such computing applications with the ability to minimize clutter by limiting the revelation of events to those which are likely to be useful landmarks. Moving beyond basic timelines for searching for desktop content, applications include the use of the inferential models for constructing hierarchical views of events for browsing large quantities of time-based content, such as autobiographical corpora. We shall now explore a sample application we have constructed to investigate prospects for harnessing statistical models of memory landmarks.

Applications of Models of Memory Landmarks

To motivate ongoing work on the use of supervised and unsupervised machine learning to construct models of landmark events, we developed a prototype that demonstrates how such predictive models might be used. We have integrated components for learning and reasoning about memory landmarks into a prototype named *MemoryLens Browser.* The prototype is focused on providing users with a timeline of landmark events to assist them to find content across their computer store. We recently distributed the prototype to a limited group of users within our organization and are pursuing feedback about the system.

MemoryLens comes in the spirit of recent work on developing tools for assisting computer users to better locate information from their personal stores (Adar, Karger & Stein, 1999; Dumais, Cutrell, Sarin, Cadiz & Jancke, 2003). Ringel, Cutrell, Dumais & Horvitz, (2003) recently reported on results of a set of user studies that showed that memory

Table 1: Training data and classification accuracies for predictive models tested on hold-out data for five subjects.

	S1	S2	S3	S4	S5
Total events	3864	3740	2770	1743	1996
-Train	3091	2992	2216	1394	1596
-Test	773	748	554	349	400
Accuracy	0.87	0.94	0.95	0.88	0.78

landmarks can be used to help computer users find relevant results in searches across personal corpora. Significant decreases in the time required to identify search results was found when memory landmarks were used in comparison with the no-landmarks condition. That system employed informal, heuristic rules for selecting memory landmarks.

MemoryLens Browser allows users to train models of memory landmarks on a portion of events from their calendar; the prototype offers menu options which provide access to the training and modeling capabilities that we described earlier. Users invoke a personalization component that executes a crawl of their calendar. The prototype provides assessment and machine-learning tools, allowing users to identify a subset of events in their calendars as landmark events, and to build predictive models by invoking machine learning from the labeled data.

In use, the models constructed by users serve to infer the likelihood that each evemt drawm from the user's calendar will be considered a landmark, *p*(*Event will be viewed as a memory landmark*|*E*), given multiple evidential properties, *E*, extracted from unlabeled calendar items. These likelihoods are considered in the generation of a timeline of inferred landmarks adjacent to files gathered from across a user's file system. The files are positioned at places along the timeline in accordance with the times that they were created or last modified. An event-detail slider control provides users with a means of changing the threshold on the inferred likelihood of memory landmark that is required for displaying events. The slider control allows users to specify thresholds for admitting items for display with successively smaller inferred likelihoods. Only calendar items representing events that have a probability of being a landmark that is greater than a user-set threshold are displayed; as the slider is moved from "most memorable" to "least memorable," the required probability for display of events is lowered, thus bringing in greater numbers of events.

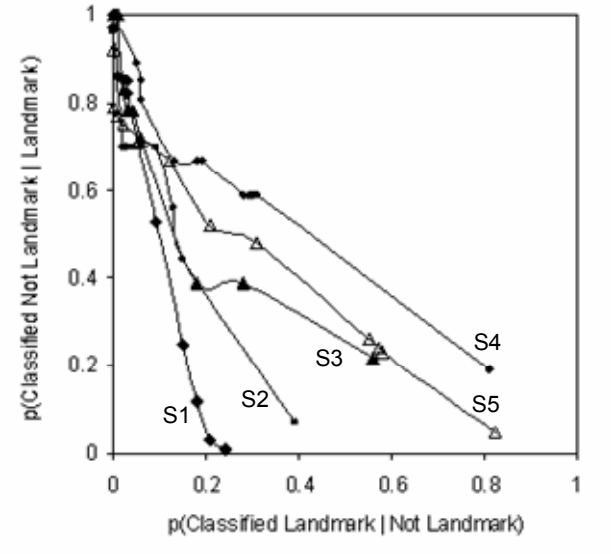


Figure 2: Receiver-operator curves showing the relationships of false negatives and false positives for five subjects at a range of thresholds on probabilities for admitting an event as a memory landmark.

A screenshot of the user interface of MemoryLens Browser is displayed in Figure 3. Thumbnails of file types are sorted in the right-hand column of the browser, in a traditional time-sorted view manner that computer users are familiar with. Within the left-hand column, a list of relevant dates associated with the files are displayed, including the year, month, and relevant days that files were created or modified. The middle column contains memory landmarks that have been assigned through inference a landmark probability exceeding a user-set threshold. The titles of memory landmarks are displayed in the appropriate temporal location, adjacent to the files.

Figure 3 shows three different screenshots of the graphical interface of MemoryLens, each representating a different setting of the probability threshold for the same span of time. Of the three snapshots, the view at the right is set to the highest probability threshold, thus revealing the fewest events. In this case, only the events representing two major conferences, for which the subject had to travel afar to attend, are displayed. As the threshold is lowered, a wedding, an editorial board meeting, a conference call, and a one-on-one meeting are included in the display. Further diminishing of the threshold for admitting events brings larger numbers of events into view. Beyond the use of thresholds for admitting versus excluding events from the landmarks column, the saturation of color of the text used to title events is faded as the probability of memory landmark diminishes—providing an additional cue about the likely value of using the event as a memory landmark.

We have been interested in probing the ability of models with the discriminatory performance represented by the family of ROC curves displayed in Figure 2, to construct useful time-line views. Such views should contain recognizable landmarks, while bypassing the clutter associated with showing a great number of events, and should allow users to work with such models in an exploratory, interactive manner (Horvitz, 1999) with tools embodied in MemoryLens' controls and display.

To relay a qualitative feel for the quality of timelines constructed with the use of the predictive models that we have generated, consider the ROC curve for a model of subject S1. The curve tells us that, at a probability threshold for accepting events as landmarks where ninety percent of the events on the timeline are correctly identified as

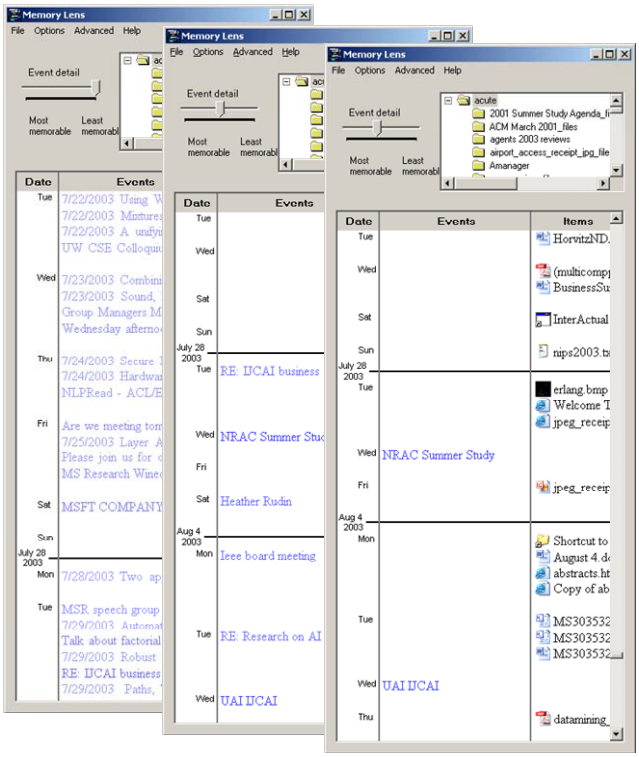


Figure 3: MemoryLens Browser with memory-landmark timeline displayed at three different settings of the threshold on the likelihood required for an event to be considered a memory landmark.

important landmarks, fifty percent of the important landmarks will not be displayed. Such precision and recall may be quite tolerable for navigating to target periods of time, given the overall density of landmarks for users; we found that subjects in our study typically showed 2-4 landmark events per week over the span of their assessments. A recall of half of these events would still tend to identify landmark events for every week.

Understanding the comprehensive value of providing users with selective views of landmark events on timelines will require detailed user studies of the use of specific prototypes and artifacts. We are interested in such studies of the value of specific designs built on predictive models of memory landmarks. Such studies would serve to enhance our understanding of the sensitivity of particular features and services to the performance of the predictive models.

Research Directions

We have focused in this paper on the construction and performance of predictive models that can be used to infer the probability that events drawn from online calendars will be considered memory landmarks by users. We provided, as a motivating example, a prototype application to highlight potential applications. Although we did not dwelled on comprehensive evaluations of the value of the use of memory landmarks in such prototypes, we are nonetheless interested in pursuing a deeper understanding of the value to users of rendering memory landmarks of different types and in different settings. We also seek to better understand the value of employing accurate predictive models of memory landmarks, based on a well-defined probabilistic semantics, versus using simple sets of heuristics to choose events for display.

In addition to pursuing a better understanding of the value of memory landmarks for users performing search and retrieval in computing applications, we are exploring several avenues of opportunity with refining and extending models of memory landmarks.

Generalization of Models. In one area of work, we seek an understanding of the accuracy of *inter-subject* predictions. Inter-subject classification accuracy probes the potential for using models constructed from one subject's training data, or a composite model built from multiple subjects, to predict hold-out data from other participants. Validating generalization across users would suggest that it is possible to field software applications that would require minimal personalization effort, via the use of pre-trained "seed" models. Such models would have a poor ability, without additional training, to consider highly personalized information such as variables containing specific text strings representing labels on meeting locations and subjects. This information tends to vary highly among the subjects.

Beyond Calendar Events. Events captured on users' calendars are convenient, but only a small subset of "events" users may wish to have captured, reasoned about, and harnessed in computing applications appear on a calendar. We are interested in building and refining predictive models for other items that could serve as additional memory landmarks or bolster event landmarks by providing richer context. As an example, we are pursuing, in a parallel project, the construction of predictive models that can identify the likelihood that images drawn from a large online personal photo library represent landmark events. To date, image analysis tools have been used in conjunction with several heuristics to select pictures when a user wishes to only review a subset of photos from a large library. Such methods include the use of a measure of the representativeness of each image to other images in the same session or event, based on such evidence as features derived from an analysis of color histograms (Platt, 2000).

In another realm, we are interested in learning from data predictive models that can automatically select the most important national and world developments, as captured by news events over time.

Beyond calendar-centric events, images, and news, online interactions, communications, and patterns of interactions with computer-based content may serve as memory landmarks. For example, particular email exchanges, or documents associated with clusters of items that have been reviewed or created in patterns of activity over time may provide an important source of events.

Taken together, multiple models of memory landmarks may be used in conjunction to build rich, multi-source timelines, providing views at different scales of time and for

different quantities of events, triaged by the likelihood that events will serve as memory landmarks.

New Classes of Evocative Features. We are also exploring the value of adding new observations features to the modeling of memory landmarks. For example, we are interested in the value of introducing a consideration of observations that assist with inferences about the likelihood that a meeting has been attended, given desktop activity over time and the sensed location of systems. Prior work has demonstrated the feasibility of performing relatively accurate inferences about the likelihood that a meeting has been or will be attended, based on an analysis of meeting properties, including activity monitored during meetings (Horvitz, Jacobs, & Hovel, 1999; Horvitz, Koch, Kadie & Jacobs, 2002; Mynatt & Tullio, 2001). Information about the likelihood of meeting attendance promises to have influence on the probability that the meeting will be viewed as a memory landmark. Other factors include capture and analysis of acoustical energy during meetings, and preparatory or follow-up activity associated with appointments.

Learning Models of Forgetting. Finally, we believe that there are opportunities for developing analogous statistical models of events and tasks that will be forgotten via supervised training. Recent longitudinal studies of office workers have identified classes of important events that are forgotten and have demonstrated the value of heuristics for ways to provide reminders about such events (Czerwinski & Horvitz, 2002). Beyond applications for people in good health, we see the feasibility of developing models for supporting people suffering with pathologies of memory associated with various forms of dementia.

Summary

We reviewed research highlighting prospects for developing and harnessing predictive models of events that will be viewed as landmarks. We focused in particular on the construction and evaluation of models that infer subsets of events drawn from subjects' calendars. After reviewing the classification and ROC curves associated with training sets obtained from five subjects, we discussed the potential to employ predictive models of memory landmarks in computing applications. We described as an example, the MemoryLens Browser prototype. Before concluding, we touched on several current research directions, including opportunities to perform additional studies to evaluate the value of displaying memory landmarks in search tasks, on seeking to define and understand the discriminatory power of additional evidential distinctions in building predictive models of landmarks, and developing models of landmark events for online images, news stories, and other items encountered or created by users in their daily lives that might be encoded as important landmarks in episodic memory.

Acknowledgments

We thank Johnson Apacible for assisting with data collection and analysis.

References

Adar, E., Karger, D. R. & Stein, L. (1999). Haystack: Per-user information environments. In *Proceedings of CIKM'99*, 413-422.

Chickering, D.M., Heckerman, D. & Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In: *Proceedings of UAI '97*, 80-89.

Czerwinski, M. & Horvitz, E. (2002). An investigation of memory for daily computing events. *Proceedings of HCI 2002*, 230-245.

Davies, G. & Thomson, D., eds. (1998). *Memory in Context: Context in Memory*. Wiley: Chichester, England.

Dumais, S., Cutrell, E., Sarin, R., Cadiz J., Jancke, G. Stuff I've Seen: A System for personal information retrieval. *Proceedings of 26th International SIGIR 2003*. 72-79.

Horvitz, E. (1999) Principles of mixed-initiative user interfaces. In: *Proceedings of the SIGCHI '99*, 159-166.

Horvitz, E., Jacobs, A., & Hovel, D. (1999). Attention-sensitive alerting. In: *Proceedings of UAI '99*, 305-313.

Horvitz, E., Koch, P, Kadie, K., Jacobs, A. (2002) Coordinate: Probabilistic forecasting of presence and availability. In: *Proceedings of UAI '02*, 224-233.

Huttenlocher, J., & Prohaska, V. (1997) *Reconstructing the Times of Past Events. Memory for Everyday and Emotional Events.* Mahwah, NJ: Lawrence Erlbaum Associates, 165-179.

Malmberg, K.J., Steyvers, M., Stephens, J., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, 30, 607-613.

Mynatt, B. & Tullio, J. (2001). Inferring calendar event attendance. In: *Intelligent User Interfaces 2001*,. 121-128.

Platt, J. (2000). AutoAlbum: Clustering digital photographs using probabilistic model merging. *IEEE Workshop on Content-Based Access of Image and Video Libraries 2000*, 96-100.

Raaijmakers, J. G. W., & Shiffrin, R. M. (2002). Models of memory. In Pashler, H., & Medin, D. (eds.) Stevens *Handbook of Psychology 3rd Edition*, Vol. 2: Memory and Cognitive Processes, 43-76.

Ringel, M., Cutrell, E., Dumais, S., Horvitz, E. (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. *Interact 2003, Ninth IFIP Interact 2003*, 184-191.

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2):145 166.

Shum, M. (1994). The role of temporal landmarks in autobiographical memory processes. *Psychological Bulletin, 124*, 423-442.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*: 460-471.

Smith, S. M., Glenberg, A. & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6(4),* 342-353.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford University Press.

Tulving, E. & Thomson, D. (1980). Encoding specificity and retrieval processes in episodic memory. *Psychological Review 80*, 352-373.

On the Tip of the Mind: Gesture as a Key to Conceptualization

Autumn B. Hostetter (abhostetter@wisc.edu)
Department of Psychology, University of Wisconsin–Madison
1202 W. Johnson Street, Madison, WI 53706 USA

Martha W. Alibali (mwalibali@wisc.edu)
Department of Psychology, University of Wisconsin–Madison
1202 W. Johnson Street, Madison, WI 53706 USA

Abstract

Why do people gesture when they speak? The reasons are not entirely clear. This paper tests two hypotheses about the role of gesture in speech production: the Lexical Access Hypothesis, which holds that gesturing aids in lexical access, and the Information Packaging Hypothesis, which holds that gesturing aids in conceptualization. Participants were asked to describe dot patterns that were either easy or difficult to conceptualize in terms of geometric shapes. Patterns that were more difficult to conceptualize elicited more gesture than the patterns that were easier to conceptualize. This result supports the Information Packaging Hypothesis.

Introduction

It is often said that a picture is worth a thousand words. In the case of speech production, it sometimes seems that creating pictures with our hands can help our audience understand what we are saying. However, despite the intuitive feeling that we gesture primarily to help our audience, some research suggests that gestures contribute little to an audience's understanding of a gesturer's speech (Krauss, Morrel-Samuels, & Colasante, 1991; Krauss, Dushay, Chen, & Rauscher, 1995; but see Kendon, 1994 for an alternative perspective). Speakers often produce representational gestures even when they know that their audience cannot see them, making it unlikely that their intended purpose is solely to help the audience (Alibali, Heath, & Myers, 2001).

This evidence that gesture does not help comprehension has led some investigators to propose that gesture has a more direct role in the speech production process, by facilitating the planning of speech. Specifically, gesture may play a role in speaking about ideas that are highly spatial or motoric in nature (Kita, 2002; Krauss & Hadar, 2001). It has been shown, for example, that gestures are more likely to coincide with words that are spatial and concrete (e.g., *spin, under,* or *cube*) than with words that are non-spatial and abstract (such as *evil*) (Krauss, 1998; Morsella & Krauss, in press; Rauscher, Krauss, & Chen, 1996). By actively engaging spatial-motoric ideas through gesture, it may become easier to speak about them.

Although gesture may be an overt manifestation of spatio-motoric thought, exactly how gesture may facilitate speech production is still the subject of some debate. The majority of research in this area has followed the speech production model proposed by Levelt (1989), which divides the speech production process into three broad stages: conceptualization, formulation, and articulation. During conceptualization, the prelinguistic thoughts of a speaker are generated and combined into propositional form. During formulation, these thoughts are translated into the appropriate linguistic units by searching through the mental lexicon and identifying the proper lemmas and lexical entries. During articulation, the motor plan for pronouncing the phonemes corresponding to the lexemes is created and executed. It seems unlikely that the production of representational gestures influences motor aspects of articulation, and, indeed, research has typically focused on the earlier stages of speech production (conceptualization and formulation) as the possible beneficiaries of gesture.

Work by Krauss and colleagues (Krauss, Chen, & Chawla, 1996) places the influence of gesture on speech as occurring primarily during the formulation stage. According to their view, referred to hereafter as the Lexical Access Hypothesis, gestures serve as a cross-modal prime to help speakers access specific items in the lexicon. In support of this view, a number of studies have shown that speakers produce more iconic gestures when the words of an utterance are more elusive (e.g. Hostetter & Hopkins, 2002; Morrel-Samuels & Krauss, 1992). For example, when speakers have time to verbally rehearse an utterance, they gesture less than when speaking completely extemporaneously (Chawla & Krauss, 1994). Similarly, speakers gesture more when describing ideas or shapes that are not readily named than when describing ideas or shapes that are easily named (Graham & Argyle, 1975; Morsella & Krauss, in press). Research with aphasic patients also suggests that gesture is involved in the formulation stage. Aphasic patients whose problems are primarily ones of lexical access use more gestures than age-matched controls (Hadar, Burstein, Krauss, & Soroker, 1998). Those whose problems are primarily not ones of lexical access produce fewer gestures than other types of aphasic patients (Hadar, Wenkert-Olenik, Krauss, & Soroker, 1998). Finally, prohibiting speakers from gesturing has been shown to negatively affect speech fluency, especially for speech that is spatial in nature (Rauscher et al, 1996).

Studies that induce tip-of-the-tongue states have yielded slightly less compelling and more contradictory findings about the facilitative effects of gesture on formulation.

Frick-Horbury and Guttentag (1998) found that preventing participants from gesturing increased retrieval failures, whereas Beattie and Coughlan (1998) found exactly the opposite pattern. Participants who were restricted from gesturing actually retrieved more words than those who were allowed to gesture in Beattie and Coughlan's study. In both studies, when participants actually produced gestures, they did not resolve their tip-of-the-tongue states more often than when they did not gesture. Thus, although gestures do tend to co-occur with speech that is spatial and with words that are difficult to find, the claim that gestures actually help the speaker to find the right words remains somewhat unwarranted at this time.

Because Levelt's (1989) speech production model is a stage model, each stage of the production process partly depends on input coming from the previous stage. Articulation cannot begin without at least a minimal amount of characteristic input from the formulator; a word cannot be uttered until it has been decided what word should be uttered. Likewise, formulation and lexical access depend on the output from the conceptualizer. The ideas and propositions that are to be expressed must be available before the correct lemma and lexical affiliate can be searched for and accessed. It would seem therefore that facilitation in the conceptualization process would translate into some facilitation at the lexical level as well. A concept that is clear in the speaker's mind is more readily lexicalized than a concept that is unclear and vague. Thus, the fact that gestures tend to co-occur with words that are spatial and somewhat elusive could also be explained as facilitation at the conceptual level. Gesture may help speakers to clarify or organize their ideas, and this may make the output of the conceptualizer more readily accessible to the formulator. Such a view of gesture is referred to as the Information Packaging Hypothesis (Kita, 2000). According to this hypothesis, gestures help speakers organize knowledge that is spatio-motoric in nature and put it into a verbalizable form. Gesture is thus a mode of thinking, an aid in translating spatio-motoric knowledge into linguistic output. By activating the appropriate bodily representations of spatio-motoric ideas, the ideas can be more fully realized.[1]

Evidence for the Information Packaging Hypothesis comes from studies that have attempted to manipulate the difficulty of conceptualization while holding constant the difficulty of lexical access. Alibali, Kita, and Young (2000) did this with children using a conservation task. They found that children used more representational gestures when they were asked to explain why two items (e.g. two balls of play dough) were different amounts than when they were asked simply to describe how the two items looked different. The words used by the children were highly similar across the two tasks; however, the explanation task required more complex conceptualization than did the description task. The authors argued that children used representational gestures more frequently in the explanation task because of the increased demands on the conceptualizer. Melinger and Kita (2001) found that adults were more likely to gesture in instances where there was a greater choice of what to say, despite the fact that the actual words being spoken in both situations were nearly identical. Again, the authors argued that gesture arises as a result of taxing the conceptualizer rather than the formulator.

Although these studies provide suggestive evidence for the Information Packaging Hypothesis, their conclusions are far from definitive. Although Alibali et al. (2000) found a significant difference in gesture production based on the difficulty of conceptualization, they did not find an especially strong effect. Also, because the study investigated gesture production in children, it may not be appropriate to generalize the findings to adult gesturers. Melinger and Kita's (2001) results suggest an effect in adults. However, they did not report statistical analysis of their data, so it is not clear whether the differences they describe are reliable.

The present experiment was designed to further distinguish between the Information Packaging Hypothesis and the Lexical Access Hypothesis. If gestures do indeed help speakers to conceptualize a spatial situation rather than just helping them to find the right words to describe that situation, then speakers should produce more gestures in a task where there are multiple conceptual options. Similarly, when the task provides only one conceptual option, speakers should produce fewer gestures. However, as long as the words that are ultimately used to describe the situation are the same, the Lexical Access Hypothesis would predict no difference in gesture production regardless of conceptual difficulty.

In order to manipulate conceptual difficulty without affecting difficulty of lexical access, it was necessary to find a task that would result in similar verbal output regardless of the level of conceptual difficulty. For this purpose, we designed a dot description task in which participants were asked to describe patterns of dots to a listener. Patterns of dots were created that could be conceptualized in a number of different ways; that is, a number of different geometric patterns could be imagined to be drawn through each dot pattern. For example, the pattern in Figure 1 could be conceptualized as two triangles, one rectangle with a triangle on top, a five-pointed star, three straight lines, or two parallelograms. This scenario (the dots-only condition) should be more conceptually difficult than a scenario in which the same dot patterns are displayed with lines drawn through them to guide conceptualization (the dots-plus-shapes condition). In both conditions, however, the goal of the participants is to describe only the dots; thus, the words ultimately used by participants should be similar regardless of the conceptual condition in which the dot pattern is presented.

[1] It should be noted that when gestures aid the conceptualizer, they may or may not also aid the formulator. A speaker who is having difficulty finding a particular word may use gesture as a way of clarifying the idea in his or her mind; this may or may not add enough additional information for the formulator to successfully access the lexical affiliate.

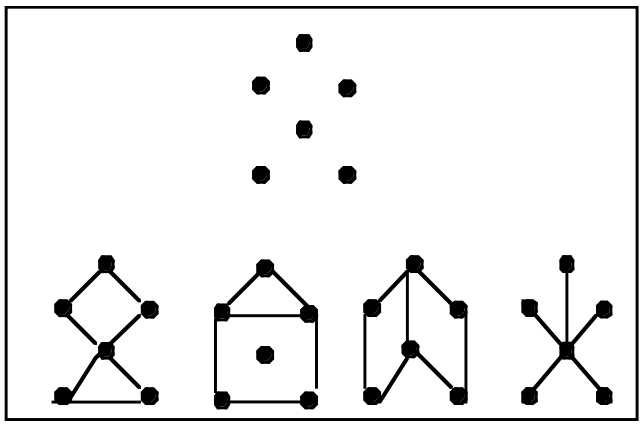


Figure 1: Sample Pattern (top) in Dots-Only Condition and 4 Possible Conceptualizations (bottom)

Method

Participants

Undergraduate students were recruited from the Psychology participant pool at the University of Wisconsin, Madison. A total of 63 individuals (32 males, 31 females) were screened for participation, and those who had not learned English in infancy were excluded. Additionally, one individual was not included because a wrist injury made it difficult for him to gesture. These eliminations resulted in a final sample of 48 individuals (24 males, 24 females) with a mean age of 19.4 yrs (SD = 1.75). Participants did not know that gesture was the focus of the study.

Stimuli

The stimuli were six dot patterns, each of which included 6 to 9 black dots on a white background. Each pattern was designed so that it did not represent the outline of any single geometric shape, but instead afforded a variety of different geometric shapes (see Figure 1). Patterns were created in AppleWorks 6 and loaded into PsyScope for experimental presentation in the conceptually difficult (dots-only) condition.

From each of these six dot patterns, the participants' natural responses in the dots-only condition were used to create patterns for the dots-plus-shapes condition. Each conceptualization provided by the participants in the dots-only condition was transposed onto the appropriate dot pattern by making lines through the dots to indicate the conceptualization pattern. These patterns were also made in AppleWorks 6 and loaded into PsyScope for experimental presentation in the conceptually easy (dots-plus-shapes) condition. All stimulus patterns in both conditions were presented on a Macintosh Powerbook G3 laptop with a 35-cm color screen.

Procedure

Participants were told that the focus of the study was their ability to remember dot patterns that are presented to them for a very short duration and to describe these patterns effectively to another participant. They were told that their descriptions would be audio-taped and played later for another participant who would try to recreate the pattern based on their descriptions. A hidden video camera was focused on the participant throughout the experiment providing a head-on view of the participant from the waist up, and their descriptions were never heard or seen by anyone other than the experimenters. Following their participation, each participant was debriefed regarding the true nature of the experiment and given the opportunity to have his or her videotape destroyed. All declined.

Each participant was brought individually into the testing room, which was divided by a wooden screen. On one side of the screen, a chair was placed in front of a small table (58.5 cm H x 71 cm W x 71 cm L) where the laptop computer was situated. The participant was told to sit in this chair and the experimenter knelt next to the computer and participant to give the instructions and practice trial for the experiment. During instruction, each participant was shown a sample dot pattern and told that patterns similar to it would appear on the computer screen for a very short duration. The participants were told that their task was to describe the pattern as clearly as possible so that the participant who would hear the description via audiotape would be able to successfully reproduce it. Furthermore, they were told that while the task was to get the listener to reconstruct the dots only, they should imagine the dot patterns in terms of geometric shapes and figures. Rather than saying, for example, that there is a dot at the top of the page and another dot about 2 cm below it with another dot directly to the left of that, they should describe the pattern as being a right triangle with dots on each corner with the right angle of the triangle facing toward the left.

Participants in both conditions received two sample patterns that were appropriate to their experimental condition (i.e., patterns shown to dots-only participants contained only dots while patterns shown to dots-plus-shapes participants showed dots with lines drawn to aid conceptualization). During the instructions and example presentations in both conditions, the experimenter produced some small, scripted gestures that were identical across conditions.

Following these instructions, participants in both conditions were asked to complete a practice trial while the experimenter was still present. The experimenter provided feedback as needed based on this practice trial to reemphasize the need to describe the pattern in terms of geometric shapes and to adequately describe the location of the dots within these shapes.

Because the purpose of this study was to manipulate conceptual difficulty without affecting lexical difficulty, measures were taken to assure that lexical access was as easy as possible for participants in both conditions. Following the examples and practice trial, all participants were presented with an alphabetical list of 16 words that

seemed likely to occur in descriptions of the patterns. This list included names of geometric shapes as well as spatial and relational prepositions. Each word was displayed in the center of the computer screen for 1200 ms, and participants were asked to pronounce each word out loud as it appeared on the screen. The goal was to have these 16 words primed and readily accessible to participants in both conditions.

After making sure that the participant understood all of the instructions, the experimenter pressed 'record' on the audio-recorder and went to the other side of the wooden screen, where she pretended to prepare for the next part of the experiment. On the experimenter's side of the screen, a table and chair were set up to face away from the participant so that in addition to vision being blocked by the wooden screen, the experimenter was not looking in the same general direction as the participant as he or she described the patterns. While it is difficult to ever definitively rule out the possibility that some gestures produced by the participants were intended for communicative illustration, the presence of the wooden screen, the relative positions of the experimenter and participant, and the participants' naivete regarding the hidden video camera make it highly unlikely that the participants perceived any direct visual audience for their descriptions. Thus the gestures produced by the speakers were most likely for purposes other than direct communicative illustration.

When the participant was ready to begin the first trial, he or she pressed a key on the laptop keyboard. At the beginning of each of the six trials in the experiment, a single black dot was displayed in the center of the computer screen for 2 s as a signal that the stimulus pattern was about to appear. The single dot was then replaced by one of the six dot patterns, which were presented to all participants in the same fixed order. The pattern remained on the screen for 3 s and was followed by a 1 s pause. After this brief pause, a short beep was heard which cued the participant that it was time to begin describing the pattern. When the participant was ready to proceed with the next pattern, pressing any key on the laptop keyboard prompted the beginning of the next trial.

Because it is crucial to the design of this study that the words used by participants in the dots-only condition closely match those used by participants in the dots-plus-shapes condition, each participant in the dots-plus-shapes condition was matched to a participant in the dots-only condition. The responses of the participant in the dots-only condition were used to create the stimuli shown to the matched participant in the dots-plus-shapes condition. Lines and shapes were drawn through each dot pattern to produce a replica of the conceptualization that was described by the dots-only participant. This was done in order to encourage the participants in the dots-plus-shapes condition to conceptualize the pattern in the same way as their dots-only counterpart, and consequently to use similar words to describe the pattern. Because this process necessitated that the stimuli be redesigned and the computer reprogrammed for each dots-plus-shapes participant, 8 participants were completed in the dots-only condition before the patterns were recreated for the dots-plus-shapes condition. Participants were then randomly matched to one of the dots-only participants with gender as the only criterion for pairing; males were always matched to males and females were always matched to females. This pairing procedure was then repeated for blocks of 8 individuals until the final sample of 24 gender-matched pairs (12 male and 12 female) was obtained.

Coding

The descriptions given by each participant were transcribed verbatim, and all iconic gestures were identified. Individual gestures were distinguished from one another by a change in hand shape or motion. For example, a motion straight across from left to right accompanying the words "the bottom line" was coded as one iconic gesture. If a similar movement occurred as the first motion of a sequence in which the hand moved diagonally upward and then diagonally downward without changing hand shape (to imply triangle), this entire sequence was coded as one iconic gesture.

The total number of gestures produced by each participant for each pattern was divided by the total number of words uttered during the participant's description of that pattern. This quotient was then multiplied by 100 to yield the iconic gesture rate per 100 words. Thus, each participant's gesture rate per 100 words was calculated for each of the six pattern descriptions.

Table 1: Frequency of Spatial Terms Used in Each Condition

Dots-Only	Spatial Word	Dots-Plus-Shapes
121	Line	148
116	Triangle	172
110	Top	122
100	Right	67
94	Bottom	124
90	Point	106
72	Middle	78
58	Parallelogram	89
58	Down	45
54	Left	40

Results

Analysis of Speech Produced

To address the question of whether conceptual difficulty affects gesture production separately from the effects of lexical difficulty, it is important that the lexical items being retrieved are similar across conditions. The 10 most commonly used spatial terms used by participants in the dots-only condition are shown in Table 1. Although there is some variation in the exact rank-order of the words, the correlation between the frequency of occurrence of each word in the two conditions is high and significant, $r(8) =$

0.833, $p < .001$, suggesting that the spatial words used by participants were similar in the two conditions.

Analysis of Gesture Production

The central question of interest in this experiment is whether or not frequency of gesture production is affected by difficulty of conceptualization. The Information Packaging Hypothesis predicts that gestures aid in conceptualization and should thus be used more frequently when conceptualization is more difficult. Alternatively, the Lexical Access Hypothesis predicts that so long as the accessibility of the lexical items being produced remains constant, gesture production should not be affected.

An inherent problem in analyses of gesture production between participants, however, is the fact that there is a large amount of individual variation in the amount of gesture produced by different individuals. Some individuals gesture a lot, while others gesture rarely or not at all. For example, in the current data set, two participants produced an average of more than 16 gestures per 100 words, while eleven others did not gesture at all. This large variation in gesture production between participants necessitates the presence of a very large difference between condition means before a significant effect can be detected. Because of this issue, we analyzed the data using items (i.e., dot patterns) rather than individuals as the unit of analysis. That is, rather than collapsing across patterns and comparing individuals in the two different conditions, data were collapsed across participants and gesture rates were compared for each of the six patterns.

A paired t-test revealed a significant effect of condition on iconic gesture rate ($t(5) = 2.84$, $p < .05$). Patterns in the dots-only condition elicited an average of 5.69 iconic gestures per 100 words ($SD = .81$) whereas the same patterns in the dots-plus-shapes condition elicited an average of 4.79 iconic gestures per 100 words ($SD = 0.26$) (see Figure 2). Thus, gesture rate varied as a function of difficulty of conceptualization, with more gestures produced in the condition with more difficult conceptualization. The distribution of high and low gesturers in each condition suggests that this effect was not driven solely by the presence of a few high gesturers in the dots-only condition.

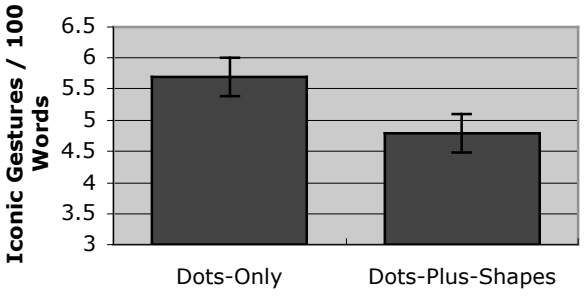


Figure 2: Average Rate of Iconic Gestures/100 Words Produced in Each of the Two Conditions

The dots-only condition included the four participants with the highest gesture rates, but also included four participants who did not gesture at all. The dots-plus-shapes condition also included participants at both extremes, with six who gestured more than 10 times per 100 words and seven who did not gesture at all.

Discussion

The Information Packaging Hypothesis (Kita, 2000) holds that gesture helps refine and organize spatio-motoric concepts so that they can be readily translated into verbalizable units. Alternatively, the Lexical Access Hypothesis holds that gesture primarily aids speech production by priming the appropriate items in the lexicon. The present experiment sought to distinguish between these two hypotheses by varying the conceptual difficulty of a task and observing the extent to which gestures are produced under the two levels of conceptual difficulty.

The difference in gesture rates in the two conditions supports the Information Packaging Hypothesis. Although the difference was small, participants gestured at a higher rate in the dots-only condition, in which they had to produce their own conceptualizations of the stimuli, than in the dots-plus-shapes condition, in which they had the conceptualizations given to them. This finding suggests that gesture does indeed occur not only when lexical access is more difficult, but also when the situation is more conceptually difficult. The present findings do not disprove the Lexical Access Hypothesis, but they do suggest that gesture can serve an earlier stage in the speech production process, above and beyond any benefits it may have at the lexical access stage.

This finding is consistent with current views about the embodied nature of cognition (e.g., Glenberg, 1997). Briefly stated, the central claim of embodied accounts of cognition is that the ways in which we are able to interact bodily with the world profoundly affect the way we think. From an embodied perspective, symbolic representations such as language are grounded or assigned meaning via their links to bodily experiences and actions. It has been shown, for example, that sentence comprehension is affected by how easy it is for the comprehender to mentally simulate him or herself actually performing the actions implied by the sentence (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Glenberg & Kaschak, 2002).

If we *understand* language in terms of how we can interact bodily with the world, it seems likely that this same embodied knowledge may also be integral to our ability to *produce* language. Indeed, the present work regarding the role of gesture in conceptualizing and formulating speech seems to point to a role for embodiment in language production. We suggest that the spontaneous gestures produced in the act of speaking are a manifestation of embodied knowledge. Borrowing a phrase from Schwartz (1998), who argued that gestures reflect "physically instantiated mental models", we suggest that gestures reflect *bodily* instantiated mental models. According to the

Information Packaging Hypothesis, such gestures enhance speakers' abilities to think and speak about those concepts. Thus, when speakers activate their embodied knowledge through gestures, they are better able to express that knowledge in the linear, symbolic system of language.

In conclusion, then, it may very well be that a picture is worth a thousand words; however, the pictures we make with our hands are not only worthwhile for our listeners, but also for ourselves.

Acknowledgments

We thank Arthur Glenberg, Charles Snowdon, and Maryellen MacDonald for their insightful comments on the design of this project.

References

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15,* 593-613.

Alibali, M. W., Heath, D. C., & Meyers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44,* 169-188.

Beattie, G., & Coughlan, J. (1998). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology, 90,* 35-56.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47,* 30-49.

Chawla, P., & Krauss, R. M. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology, 30,* 580-601.

Frick-Horbury, D., & Guttentag, R. E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *American Journal of Psychology, 111,* 43-63.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20,* 1-55.

Glenberg, A. M. & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin, & Review, 9,* 558-565.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology, 10,* 57-67.

Hadar, U., Burstein, A., Krauss, R. M., & Soroker, N. (1998). Ideational gestures and speech: A neurolinguistic investigation. *Language and Cognitive Processes, 13,* 59-76.

Hadar, U., Dar, R., & Teitelman, A. (2001). Gesture during speech in first and second language: Implications for lexical retrieval. *Gesture, 1,* 151-165.

Hadar, U., Wenkert-Olenik, D., Kuass, R. M., & Soroker, N. (1998). Gesture and the processing of speech: Neuropsychological evidence. *Brain & Language, 62,* 107-126.

Hostetter, A. B., & Hopkins, W. D. (2002). The effect of thought structure on the production of lexical movements. *Brain & Language, 82,* 22-29.

Kita, S. (2001). How representational gestures help speaking. In D. McNeill (Ed.), *Language and Gesture.* Cambridge, UK: Cambridge University Press.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7,* 54-60

Krauss, R. M., Chen, Y., & Chawla, P. (1996) Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology, 28,* 389-450.

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology, 31,* 533-552.

Krauss, R. M. & Hadar, U. (2001). The role of speech-related arm/hand gestures in word retrieval. In R. Campbell & L. Messing (Eds.), *Gesture, Speech and Sign.* Oxford: Oxford University Press.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61,* 743-754.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-38.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

Melinger, A. & Kita, S. (2001?). Does gesture help processes of speech production? Evidence for conceptual level facilitation. *Proceedings of the 27th Berkeley Linguistics Society Meeting.*

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18,* 615-622.

Morsella, E., & Krauss, R. M. (in press). Movement facilitates speech production: A gestural feedback model.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7,* 226-231.

Schwartz, D., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science, 20*, 457-497.

Cognitive Constraint Modeling: A Formal Approach to Supporting Reasoning About Behavior

Andrew Howes (HowesA@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff, Wales, UK CF10 3YG.

Alonso Vera (avera@mail.arc.nasa.gov)
NASA Ames Research Center, MS 262-3 Moffet Field, CA 94035.

Richard L. Lewis (rickl@umich.edu)
Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1109.

Michael McCurdy (mmccurdy@arc.nasa.gov)
NASA Ames Research Center, MS 262-3 Moffet Field, CA 94035.

Abstract

Cognitive Constraint Modeling (CCM) is an approach to reasoning about behavior that (1) provides a framework for investigating the hypothesis that skilled behavior is the optimal solution to a constraint satisfaction problem defined by objective, environmental, knowledge, and architectural constraints, (2) derives predictions of behavior from formal specifications of theory, (3) supports reasoning using both dependency-based and cascade-based ontologies for expressing temporal relationships between processes. A software tool that demonstrates the potential advantages of CCM is described. The tool, called CORE, can be used to partially automate the generation of behavioral predictions given a specification of the constraints. We explore the application of CORE to dual-task data previously modeled with EPIC and ACT-R.

Introduction

When people acquire a skill they are able to adapt their behavior so as to incrementally improve the value of some utility function. With practice, the scope for improvement attenuates and performance asymptotes. It may asymptote at a level that is consistent with constraints imposed by the environment or perhaps at a level determined by the knowledge that is brought to the task. The bounds may instead be imposed by the human cognitive architecture. More plausibly, the asymptote may be determined by a combination of constraints, including the stochastic and temporal profiles of the task environment *and* the human cognitive, perceptual, and motor systems. The approach to the asymptote is bounded by a multiplicity of constraints (Simon, 1992).

There has of course been much work aimed at modeling skilled behavior and its acquisition (e.g. Anderson and Lebiere, 1998; Meyer and Kieras, 1997; Taatgen and Anderson, 2003). The purpose of the current paper is to provide an initial demonstration of how models of skilled behavior can be generated by the formal derivation of behavior descriptions from multiple constraints, and in particular, how this approach supports reasoning about asymptotic bounds on skilled behavior. The specific objectives of the paper are:

(1) To introduce the hypothesis that skilled behavior is the optimal solution to a constraint satisfaction problem defined by architecture, task environment, and knowledge constraints.

(2) To introduce a formal modeling approach, called CCM, that directly supports reasoning about the optimal bounds on skilled behavior. By using deductive inference and constraint satisfaction algorithms, CCM computes the necessary consequences of the constraints imposed by the task environment, by strategic knowledge, and by the cognitive architecture. These constraints may determine, for example, which cognitive and environmental processes can execute in parallel and which have sequential dependencies.

(3) To specify two ontologies which provide alternative information processing vocabularies for the cognitive and task theory, and the resulting descriptions of behavior. The first is a straightforward formalization of temporal dependencies, implicit in existing work based on CPM-GOMS. The second is a richer ontology that permits specifying sets of communicating information processes, where both the processes, inter-process communication channels and buffers are subject to resource constraints. This framework has much in common with McClelland's cascade model (McClelland, 1979). Both ontologies are formally defined by a set of declarative axioms that are part of the model specification.

The paper has the following structure. We first introduce the background to our work on CCM and then describe a reasoning tool, called CORE (Constraint-based Optimal Reasoning Engine). We describe the application of CORE, using the temporal dependency axioms, to reasoning about a dual task experiment first reported by Schumacher, Lauber, Glass, Zubriggen, Gmeindl, Kieras, Meyer (1999) and subsequently modeled by Byrne and Anderson with ACT-R/PM (Byrne and Anderson, 2001). In doing so we show that CORE is flexible enough to support inference about the implications of both central and peripheral bottleneck theories of dual task performance. CORE requires 42 simple, universally quantified, declarative statements to

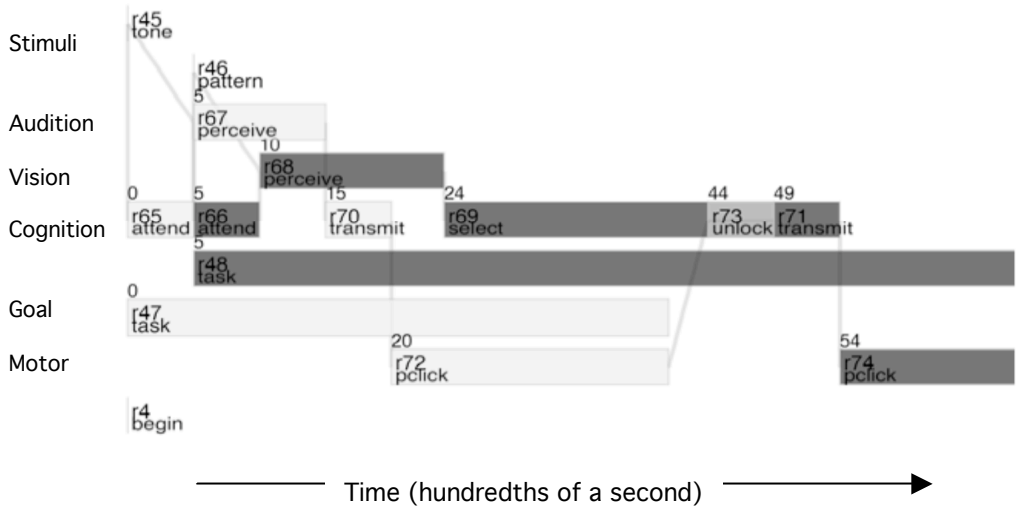


Figure 1: A CCM-d prediction of performance on Schumacher et al.'s (1999) Experiment 3 task.

specify the task, strategies, architecture, and axioms required to reason about the dual task. We also report that the optimal schedule for Schumacher et al.'s task suggests that participants may have been using a strategy not previously considered. Lastly, we introduce, and describe the benefits of, the cascade-based axioms.

Background

Predicting how long it will take people to perform a task is difficult, but important. It is difficult because human performance depends on a multiplicity of complex interacting constraints that derive from the environment, from human psychology, and from knowledge that people bring to the task. Skilled performance of a routine task usually involves the execution of a number of parallel but interdependent streams of activity: For example, one hand may move to a mouse; while the other finishes typing a word; and the eyes begin to fixate on a menu while the required menu label is retrieved from memory. Each of these processes takes a few hundred milliseconds, but together they form behaviors that take many seconds. Importantly, the details of how processes are scheduled, of how they are ordered, and of the implications of their inter-dependencies, has significant consequences for the overall time requirement.

There are a number of scientific and engineering tools that support the prediction of skilled performance time. Many of these tools share a common intellectual origin in the Model Human Processor (MHP: Card, Moran & Newell, 1983). Card et al. introduced the MHP as an engineering model in which human cognition was described as a set of communicating processors each of which had parameters (e.g. for cycle time) derived from human experimental psychology. More recent engineering tools, particularly CPM-GOMS (Gray, John, Atwood, 1993), have also utilized the processor and process framework. EPIC, a production system architecture that synthesizes more recent results in cognitive and perceptual psychology (Meyer and Kieras, 1997) was also influenced by the work of Card et al. (1983). Most recently, ACT-R/PM (Byrne and Anderson, 1998) extended the ACT-R architecture with a set of EPIC-like perceptual-motor modules.

The strength of these strands of work is evident in the range of experimental findings for which explanations can be offered; for the successful efforts at delivering scientifically validated tools to applied practitioners (e.g. Gray, John, Atwood, 1983); and for the rigor that is evident in the insistence that theory be expressed computationally (e.g. Byrne and Anderson, 2001; Meyer and Kieras, 1997). However, there are issues. Below we have listed three that were significant in motivating the work reported in this paper.

(1) It is difficult to inspect and modify architectural assumptions (Cooper and Shallice, 1995). Cognitive architectures embody architectural assumptions in underlying code, and are not easy to change. This would not be a problem if the details of an architectural theory were stable and comprehensive enough to be applied to a wide range of tasks. But in the forseeable future the modeler will find it valuable to easily manipulate and add architectural assumptions that are still under debate in the field.

(2) Model predictions can be a function of theoretically-irrelevant or implicit assumptions. Current approaches force modeling at certain fixed levels of abstraction. In general, computational cognitive architectures force computational completeness in order to incrementally simulate behavior. But one consequence is that modelers must specify the details of procedural knowledge and the representations used in long term and short-term memory, which may not be intended as theoretical commitments.

(3) It is difficult to predict the asymptotic bound on skilled behavior. Though learning architectures such as ACT-R/PM could in principle automatically asymptote to the appropriate skilled behavior, the mechanism is an open research problem and puts a robust learning theory on the critical path to efficiently modeling skilled behavior.

CORE: A tool to support reasoning about behavior

CORE takes as input a set of mathematically stated constraints on behavior and outputs a prediction. In this respect it shares some similarities to the work of Duke and Duce (1999). One of the formats for the output, a CPM-GOMS-like Pert chart, is illustrated in Figure 1. The prediction in the Figure is for a dual task behavior studied by Schumacher et al. (1999) and subsequently modeled in ACT-R/PM by Byrne and Anderson (2001). Each box represents a process. Task 1 (light gray processes) is to respond to a tone (high or low) with a left-finger key press, and task 2 (dark gray) is to respond to a pattern with a right-finger key press. In the Figure, time is represented on the horizontal access and each row represents a different resource or processor, perception at the top, through cognition and goal, to motor actions. The task processes represent the temporal extent of the representation of each task on the goal.

Following Card et al. (1983) constraints are described in terms of the temporal and resource properties of a distributed set of processors, each with its own processing capabilities. Each processor is defined in terms of a set of parameters and a defined set of processes. Each process has parameterized limits (min, max) on its duration. The duration of motor movements can be automatically determined by a calculation of Fitts's Law. The duration of cognitive and perceptual processes may be directly determined from the empirical literature (e.g. estimates of the time required to switch attention), or by functions that, for example, model hypotheses about the behavior of retrieval mechanisms.

The relationships between the processes represented in Figure 1 are an attempt to reproduce the assumptions adopted by Byrne and Anderson (2001). The processing sequence for each stimulus is: attend to the stimulus, perceive the stimulus, select a response, transmit a command to the motor system. The duration of the select process is determined by ACT-R/PM's retrieval function.

The lines between processes in Figure 1 represent temporal *dependencies*. A dependency is a type of constraint that specifies that one process must be scheduled after another has finished. While the particular prediction illustrated in Figure 1 has been constrained by dependencies, cognitive constraint modeling is not limited to dependency-based representation of theories. The constraints that specify the meaning of dependencies are the essence of a set of CCM axioms that we call CCM-d (CCM-dependency). CCM-d provides a formal specification of the CPM-GOMS modeling framework (Vera et al., 2004). (An alternative set of axioms, called CCM-c, for CCM-cascade, is described later in the current article.)

Representing a theory

Constraints on behavior are specified to CORE in terms of relationships between events in the environment, tasks, and psychological processes.

The semantics of the language for expressing statements is a subset of second-order predicate calculus. An entity is represented as a set of elements where each element is either an ordered pair, or a triple where the first element is '++'. For example, the following reads, there exists a cognitive process called initclick that must be scheduled after process U_j.

$$\exists P_i: \{ (isa,process)\ (name,initclick)\ (resource,cognition)\ (++,after,U_j) \} \subseteq P_i \qquad (1)$$

Each pair consists of an *attribute* and a *value*. A set must only contain a single element with a particular attribute (e.g. there must only be, at most, one pair that matches the pattern (name,_)). Each triple consists of the symbol '++', an attribute, and a value. For triples, there are no restrictions on the attribute or value. Triples support the expression of sets in which an attribute can have multiple values. The features in (1) are specified as a subset of P_i ($\subseteq$). Further features may complete the specification of this process.

Sets that represent processes, must have a start attribute and a duration attribute. This can be represented with the statement that all P_is that contain the pair (isa,process), must also contain a start time S_i and a duration D_i.

$$\forall P_i: \{ (isa,process) \} \subseteq P_i \rightarrow \{ (start,S_i), (duration,D_i) \} \subseteq P_i \qquad (2)$$

Relationships between the start times and durations of processes are represented with simple integer-arithmetic constraints. The following represents the assumption that a motor process is a necessary consequence of an initialization process, that a motor process cannot occur before its initialization process, and that the maximum temporal gap between the two processes is 300ms. This constraint must hold irrespective of the task.

$$\forall P_j: \{ (isa,process)\ (name,initclick)\ (start,S_j)\ (duration,D_j) \} \subseteq P_j \rightarrow \exists P_i: \{ (isa,process)\ (name,click)\ (start,S_i) \} \subseteq P_i \wedge S_j + D_j \leq S_i \wedge S_i - (S_j + D_j) \leq 300 \qquad (3)$$

Given a set of axioms, statements of this form can be used to represent theoretical assumptions about the task environment, about instruction taking, about the strategies that people deploy, and about the human cognitive architecture.

Crucially, universally quantified constraints specified in a predicate calculus are not production rules. The constraints may appear to possess a similar surface form to production rules but, in fact, the semantics are very different. Most importantly, unlike production rules, these declarative statements of theory are statements of what *must* be true irrespective of context. They are not elements of a

procedure that generates the description. The constraint must hold for every circumstance where its antecedent is met. The generation of a model with these constraints is entirely monotonic and the order of expansion can be (and often is) different to the predicted order of behavior.

Generating a prediction

Given constraints on behavior, CORE can be used to generate a prediction. This is a two-phase process.

Phase 1. CORE derives the necessary implications of the theory. For example, given a P_i (as defined in statement 1) above and rule (2), CORE would derive that the initclick process must have a start and duration:

$$\exists\, Pj: \{ (isa,process)\ (name,initclick)\ (resource,cognition)\ (++,source,U_i)\ (start,S_j)\ (duration,D_j) \} \subseteq P_j \quad (4)$$

Subsequently, with rule (3), CORE can derive that there must be a motor click process, with a start time and duration constrained by the given equations.

Arithmetic constraints on the start time, duration, and costs of a process are posted to a constraint store that is implemented in a Sicstus Prolog variant of Constraint Logic Programming for Finite Domains (CLP FD: Jaffar & Lassez, 1987). Much of the power that CORE provides is a direct consequence of CLP FD functionality (Vera et al., 2004). Importantly, the scheduling algorithms provided by CLP FD make it possible for an analyst using CORE to focus on the declarative specification of theory without worrying about the theory-irrelevant algorithms by which behavioral implications will be derived.

At the end of phase 1 the values of the start, duration, and other parameters, such as cost, are constrained by the posted equations, but their values are not yet uniquely determined.

Phase 2. Phase 2 involves making a prediction by finding a particular behavior that is consistent with the set of constraints, i.e. phase 2 must identify a consistent set of values for variables that were posted to the CLP FD constraint store (e.g. start time, duration, cost). This is achieved by calling a function that uses constraint satisfaction to achieve variable assignment. This function can be configured to use a range of scheduling algorithms. Two are particularly important for the purposes of reasoning about cognition: greedy scheduling and optimal scheduling.

Greedy scheduling. Scheduling proceeds with the tick of a clock. On each tick, a process is selected that can be scheduled immediately. The process is assigned the appropriate start time. Greedy scheduling can be used to model ACT-R/PM and EPIC. A greedy scheduling algorithm is not guaranteed to give an optimal schedule.

Optimal scheduling. Using CLP FD, a branch-and-bound algorithm can be used to generate a schedule with the greatest utility. We have used a utility function that is maximal when cost is minimized. Cost is defined as the sum of the total duration of the schedule and the durations of the buffers required to store information. As we illustrate below the ease with which CORE can be used to generate predictions of the optimal behavior, given the theoretical assumptions, is one of its key advantages.

Reasoning about dual task performance

In Schumacher et al.'s (1999) experiment (Experiment 3) participants were required to respond to a tone and a visual pattern with key presses that depended on whether the tone was high or low and whether the pattern contained a particular feature. The tone and the pattern were presented with a small gap of between 50 and 1000ms (Stimulus Onset Asynchrony). Participants were asked to prioritize the tone task. The tone task response times were, on average, unaffected by SOA. In contrast, the mean pattern task response time, at a short SOA (50ms), was less than the sum of the tone task and pattern response times at long SOAs (> 500ms). This finding has been taken as evidence that some elements of tone and pattern task were performed in parallel at short SOAs. Byrne and Anderson were interested in modeling Schumacher's data using ACT-R/PM in order to demonstrate that cognitive parallelism is not required to explain these results. They argued that the results can be modeled with either the EPIC or ACT-R/PM assumptions and that Schumacher's data provides evidence for strategic deferment of the pattern task response.

Specification and Inference

We demonstrate that the theoretical assumptions of Byrne and Anderson (2001) and separately of Schumacher et al. (1999) can be precisely expressed as a small set of predicate calculus constraints, and that CORE can be used to support reasoning about their behavioral consequences.

We used 42 universally quantified constraints to capture the theoretical assumptions underlying the architecture and strategies deployed in Byrne and Anderson's model (see www.cf.ac.uk/psych/howesa/ccm). Together with the CCM-d axioms, these constraints are a mathematically-complete specification of the theory underlying Figure 1. They capture constraints on the task environment, the strategy, the architecture, and in addition the axioms of CCM-d. (It would in fact be possible to use many fewer constraints but we attempted to write them in a way that was general enough to enable reuse.) We selected parameters to fit the performance time and ran the model with a greedy scheduling algorithm to check its performance. It produced the same pattern of results as reported by Schumacher and modeled by Byrne and Anderson (2001).

One of the constraints that is particularly important for the predictions made by Anderson and Byrne's model states that the duration of retrieval is sensitive to whether the retrieval request is issued when the tone task overlaps in time with the pattern task. This relies on an implementation of the ACT-R retrieval time function, *retrieval_time(B,S,T),* where B is the base level activation, S is the source activation, and T is the returned retrieval duration. The source activation is lower per task when there are multiple concurrent tasks.

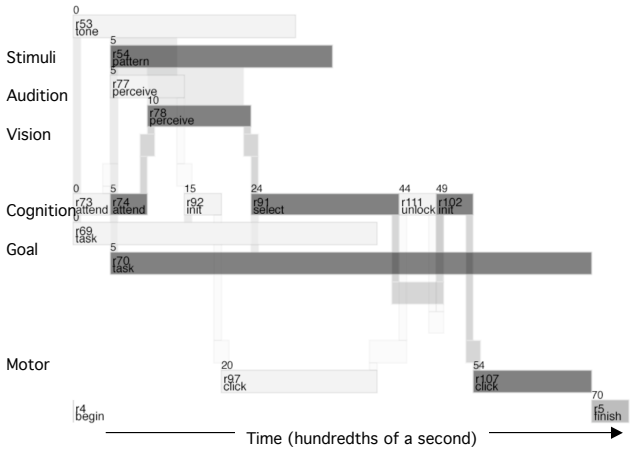


Figure 2: A CCM-c prediction of performance on Schumacher et al.'s (1999) experiment 3 task.

Subsequently, we modified the specification in order to remove the ACT-R/PM assumption that there is a central bottleneck on human cognition and permit EPIC-like concurrent cognitive processing. Only a handful of changes were required to make this alteration, demonstrating the claim that CORE facilitates reasoning about the consequence of different architectural assumptions. First, the tone task and pattern task cognitive processes were assigned to separate resource streams, and second, an unlock process was introduced and its duration adjusted to fit the data.

The fact that ACT-R and EPIC resource assumptions can be captured with such similar sets of constraints is unsurprising given their shared intellectual history (though it would be interesting to compare the Lisp code). In general, the space of theories that can be represented with CCM is determined by the requirement that theories are described in terms of processes, processors, the relationships between them, and by the axioms (CCM-d or CCM-c).

Skill and optimal constraint satisfaction

In order to explore the hypothesis that skilled behavior is the optimal solution to a constraint satisfaction problem, we switched from greedy scheduling to optimal scheduling (a parameter change in the input specification). CORE generated a novel prediction. In the behavior for a task with a 50ms SOA and a simple pattern, rather than choose to schedule the pattern selection process so that it was concurrent with the tone task, CORE chose a schedule in which selection (i.e. retrieval) is deferred until after the tone task has finished. The benefit is that by deferring selection, the overall time requirement is slightly reduced. This is because even though the selection process starts later (after the tone click in Figure 1) it has a much shorter duration (only 30ms compared to the 250ms). The resulting difference in the overall time cost of the schedules is marginal but the qualitative difference in the strategies is dramatic. The analysis exposes a necessary consequence of ACT-R's retrieval function and the assumption that people adapt strategies to reduce time cost.

The example illustrates the way in which CORE can facilitate the exposure of a logically required implication of a set of theoretical assumptions. Byrne and Anderson's ACT-R model does not make this prediction because it does not optimize over the total cost of the behavior. Optimal scheduling exposes the possibility that participants strategically defer retrieval so as not to incur the costs of concurrent processing.

Our analysis also raises a question about a fundamental assumption embedded in the ACT-R retrieval function: That retrieval time is not dynamically adjusted with changes in source activation occurring during retrieval. I.e. the sensitivity of the retrieval time function to source activation is limited to the value of the source activation at the time of the retrieval request. An alternative assumption would be that retrieval could take advantage of increases in source activation that occur after a retrieval request is made but before a chunk is delivered. With this alternative assumption, deferred retrieval in the Schumacher task would carry no advantage. Which assumption provides a better model of human retrieval is an empirical question that is not answered by Schumacher's data.

Cognition as cascading information processing

CORE is flexible enough to accept theory specifications expressed relative to a range of different sets of axioms. For the work reported above we used axioms that were based on the notion of a dependency (CCM-d). However, there are intrinsic limitations of dependency-based axioms (Vera et al., 2004) and we have therefore been working on the specification of a set of axioms that is based in part on the idea of a cascade as a mechanism for representing overlapping, communicating processes. Our formalization builds on McClelland's (1978) original cascade assumptions to explicitly include the declaration of resource-limited communication channels and buffers between processes.

Figure 2 illustrates a prediction derived by CORE using CCM-c axioms. The start times and durations of the processes are the same as in Figure 1. The difference between the figures is that the relationships between processes are expressed in terms of cascades. These capture the resource requirements and temporal constraints on the inter-process communication channels. The cascades are represented in Figure 1 as the lightest gray bars that run between the processes. One advantage of cascades is that they prevent cognitively implausible process orderings that would be legal using CCM-d. For example, the process ordering init(x), init(y), click(y), click(x) is legal in CCM-d but cognitively implausible because it assumes no cost to buffering information between the cognitive intention and the motor action.

Discussion

We have introduced a framework and a tool for making inferences about the implications of formally specified theories of skilled behavior. The tool uses a constraint logic-programming environment to support the inference of the

asymptotic bound on skilled behavior given a specification of the constraints on the task environment, on perception, on cognition, and on action.

Our investigations are at an early stage. We have so far explored the potential of dependency and cascade axioms on only a handful of tasks. In addition to the dual task described in the current paper we have also used CORE to generate predictions for a range of applied tasks including a call-center accounts advise task, and a laboratory version of an Automated Teller Machine.

Our aim in conducting this work was not to recast ACT-R/PM and EPIC in a formal language. The aim was to provide a tool that could assist in the prediction of the asymptotic bound on skilled behavior given constraints on, not only, objective and environment but also on strategies and architecture. We concur with Simon (1992) that an analysis of the optimal adaptation given all of these sources of constraint provides a more accurate estimation of behavior. While the extent to which we can achieve our aim is yet to be determined, we have presented arguments for the scientific merit of deductive inference in exploring the asymptotic bound on skilled behavior. We have shown that by deriving the optimal schedule of behavior for these constraints, logically implied but previously unexplored predictions of behavior can be exposed. The fact that a novel prediction was generated for a task that has been the subject of a number of published studes illustrates that the benefits of cognitive constraint modeling go beyond redescription of existing theory.

One potential concern is that if we were to write a set of constraints to capture the range of behaviors exhibited by, for example ACT-R, we would generate a set that was as large and formidable as ACT-R's Lisp code. Our response is twofold. First, we note that CCM is not a simple subset of ACT-R, it includes functionality, particularly optimization, that is not present in simulation architectures. Second, we point out, again, that our aim is not to recast ACT-R or EPIC in a formal language. More particularly, our aim, at present, is not to build a simulation architecture, rather it is to provide a tool for supporting reasoning about psychological theory. Much of the complexity of the ACT-R and EPIC implementations may be related to the simulation-based framework in which they are cast.

Our current work is aimed at further developing the generality of CORE. Most importantly, we need to take full advantage of the constraint satisfaction engine, CLP FD, that is used for the calculation of arithmetic parameters. In the present implementation of CORE, this engine is not used to reason about the symbolic inter-process constraints. We also need to work on using constraint satisfaction techniques that support reasoning about statistical distributions rather than just integer values.

In conclusion, we have introduced a constraint-based framework for reasoning about human behavior and argued for the utility of a specific tool called CORE. Our investigation suggests that partially automatic algorithms can be used to generate predictions of optimal human behavior from concise, theory-relevant, and readily modifiable, specifications of psychological theory.

Acknowledgments

This work was supported by ONR Grant number N0001404IP2002. Bonnie John, Mike Byrne, and Duncan Brumby made a number of valuable comments on this work.

References

Anderson, J.R. (1990). *Rational Analysis*. LEA.

Anderson, J.R. and Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Byrne, M.D. & Anderson, J.R. (1998). Perception and action. In J. R. Anderson and C. Lebiere (Eds.) *The Atomic Components of Thought* (pp. 167-200). Mahwah, NJ: Erlbaum.

Byrne, M.D. & Anderson, J.R. (2001). Serial modules in parallel: The psychological refractory period and perfect time sharing. *Psychological Review, 108*, 4, 847-869.

Card, S.K., Moran, T.P., Newell, A. (1983). *The Psychology of Human Computer Interaction*. NJ: Erlbaum.

Cooper, R. & Shallice, T. (1995). Soar and the case for unified theories of cognition. *Cognition, 55*(2), 115-149.

Duke, D.J. and Duce, D.A. (1999). The formalization of a cognitive architecture and its application to reasoning about human computer interaction. *Formal Aspects of Computing, 11*, 665-689.

Gray, W.D., John, B.E. and Atwood, M. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human-Computer Interaction, 8*, (3), 237-309.

Jaffar, J., & Lassez, J.L. (1987). Constraint logic programming. In *Proceedings of the ACM Symposium on Principles of Programming Languages*, ACM Press.

McClelland, J.L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86*, 287-330.

Meyer, D.E. & Kieras, D.E. (1997). A computational theory of human multiple-task performance: The EPIC information-processing architecture and strategic response deferment model. *Psychological Review, 104*, 1-65.

Schumacher, E. H., Lauber, E. J., Glass, J. M., Zurbriggen, E. L., Gmeindl, L., Kieras, D. E., & Meyer, D. E. (1999). Concurrent response-selection processes in dual-task performance: evidence for adaptive executive control of task scheduling. *Journal of Experimental Psychology: Human Perception and Performance , 25*, 791-814.

Simon, H.A. (1992). What is an "explanation" of behavior? *Psychological Science, 3*, 150-161.

Taatgen, N.A. & Lee, F.J. (2003). Production Compilation: A simple mechanism to model Complex Skill Acquisition. *Human Factors, 45*(1), 61-76.

Vera, A., Howes, A., McCurdy, M., Lewis, R.L. (2004). A constraint satisfaction approach to predicting skilled interactive cognition. In *Procedings of the ACM Conference on Human Factors in Computing Systems CHI'04*, Vienna, April 24-29, 2004.

Connectionist Modelling of Chinese Character Pronunciation Based on Foveal Splitting

Janet Hui-wen Hsiao (h.hsiao@sms.ed.ac.uk)
Richard Shillcock (rcs@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Abstract

We describe a connectionist model designed to reflect some of the anatomy of the visual pathways, notably the precise division of the human fovea and its subsequent contralateral projection to the cortex. The model was trained on a realistically large-scale problem, mapping between Chinese orthography and phonology. This split-fovea model replicated the interaction between character regularity and frequency that has been found in Chinese phonetic compound naming tasks. It also provided cross-language support for the hemispheric desynchronization account of dyslexia. Finally, the model predicted different regularity effects between characters with different phonetic radical positions.

Introduction

Cognitive scientists aim to understand language processing universals. Seidenberg and McClelland's "triangle model" of the reading of monosyllabic English words has been substantially developed (e.g. Harm & Seidenberg, 1999). However, there is still little application to languages other than English. The cognitive modelling of the processing of Chinese orthography suffers from an input representativeness problem (cf. Chater & Christiansen, 1999) due to its complexity; there is ongoing debate as to how to represent Chinese characters in a psychologically realistic way. Most of the proposed computational models of Chinese character reading either have not been computationally implemented (e.g. Perfetti & Tan, 1999), or have employed relatively small-scale training data (e.g. Chen & Peng 1994). Cognitive modelling research in Chinese reading thus has lagged behind research in the reading of English.

Chinese has a radically different orthography from any alphabetic language. The basic writing units of Chinese are characters, which usually contain meaningful morphemes, instead of the letter-based representations of speech segments found in alphabetic languages. In general, there are four different ways of composing Chinese characters: *pictographs*, *indicatives*, *ideographs*, and *semantic-phonetic compounds*. A semantic-phonetic compound (or simply a *phonetic compound*) contains both semantic and phonological information. Such compounds comprise about 81% of the 7,000 most frequent characters in the Chinese dictionary (Li & Kang, 1993). Hence, understanding how Chinese readers recognize these phonetic compounds is an important goal in psycholinguistic cognitive modelling.

A phonetic compound can be decomposed into two major components: a semantic component, which bears information about the meaning of the character, and a phonetic component, which may have partial information about the pronunciation of the character. Most phonetic compounds have their semantic radicals on the left and phonetic radicals on the right (SP characters). For example, the character "沐" means "taking a bath" in English and is pronounced as "mu4" in Pinyin[1]. It consists of a semantic radical on the left, which means "water", and a phonetic radical on the right, which is pronounced the same as the character. We call these characters *regular phonetic compounds*. Some characters may be pronounced slightly differently from their phonetic radicals, such as "柚". Its phonetic radical "由" is pronounced as "iou2" in Pinyin. However, "柚" has a different tone – it is pronounced as "iou4". These characters are referred to as *semi-regular phonetic compounds*. Finally, there are some characters pronounced very differently from their phonetic radicals, such as "洒" (sa3) and "西" (xi1). We call them *irregular phonetic compounds*. The opposite structure to an SP character exists, in which the phonetic radical is on the left and the semantic radical is on the right (PS characters). The ratio of SP characters to PS characters is about 9:1 (Hsiao & Shillcock, *in preparation*).

A *regularity effect* has been found in the processing of Chinese phonetic compounds: Chinese readers name regular characters faster than irregular characters. There is also a frequency by regularity interaction in Chinese, as in English (see, *.e.g.*, Hue, 1992; Liu, Wu & Chou, 1996; Seidenberg, 1985.)

Researchers have also studied Chinese character reading in brain-damaged patients (Yin & Butterworth, 1992) and found similar disorders as those found in

[1] The Chinese Pinyin system is a spelling system based on the Latin alphabet. The number at the end indicates the tone type.

English word reading. Chinese deep dyslexics were found to be able to pronounce irregular characters well but had difficulties pronouncing pseudo-characters with real semantic and phonetic radicals. On the other hand, Chinese surface dyslexics tended to regularize irregular characters and were able to pronounce about 50% of pseudo-characters according to their phonetic radicals (Zhou, 1999).

There is clear evidence that the human fovea is split precisely about the vertical midline: the left and right visual hemifields are projected contralaterally to the right and left hemispheres respectively (see, e.g. Fendrich & Gazzaniga, 1989). On the basis of anatomical and other evidence, a "split-fovea model" of English word reading has successfully captured several reading phenomena (see, *e.g.*, Shillcock Ellison & Monaghan, 2000; Shillcock & Monaghan, 2001). Chinese phonetic compounds provide opportunities not available in alphabetic languages for examining the plausibility of this split-fovea model, since phonological information only comes directly from half of a character. In other words, the split fovea architecture seems to correspond fortuitously to the major functional division in the structure of Chinese phonetic compounds; the model "carves the problem at its joints". Also, when an input character is irregular, the model faces an XOR-like problem, which makes interaction between the two halves necessary. Here we report our results of applying this split-fovea architecture to the modelling of Chinese character pronunciation.

Simulations

Phonological Representations

The sound system of Chinese differs from that of English. One of the most salient differences is the four distinct tones in standard Chinese (i.e. Mandarin)[2]. The pronunciation of each character has only one syllable, and every syllable has a nucleus and a tone associated. Characters with the same nucleus but different tones are usually not related in their meanings or orthography. In addition to a nucleus and a tone, there are three optional components associated with a syllable: a consonant at the beginning, a glide in the middle, and a glide or a consonant from a restricted class at the end (Wang, 1973). In total, syllables in Chinese have eight possible forms.

In Chinese syllables, all consonants can appear in the initial consonant position, and all vowels can appear in the nucleus position. On the other hand, there are only three possible vowels in the medial glide position, and five possible consonants and vowels in the ending position. According to the phonetic features of the Chinese Pinyin system ("Mandarin Consonants and Vowels"), there are 14 features for consonants: *bilabial, labiodental, dental, alveolar, palatal, velar, stop-aspirated, stop-unaspirated, nasal, fricative, affricative-aspirated, affricative-unaspirated, glide, and liquid*. Hence, we encoded every consonant in terms of these 14 features. Vowels were encoded with 8 features: *front, central, back, high, mid, low, unround, and round*.

In our phonological representation, the two major parts were the initial consonant, which consisted of 14 nodes for the 14 consonant features, and the nucleus vowel, which consisted of 8 nodes for the 8 vowel features. The glide was represented together with the vowel features in the nucleus vowel section. The same applied to the vowel features in the ending position. After 8 vowel feature nodes, we used 3 nodes to represent the features of the consonant in the ending position (*nasal, dental,* and *velar*). Notice that there are only two consonants (n and ng) possible in the final position. The last 2 nodes represented high and low tones respectively. 4 different tones in Chinese were represented with different combinations of the high and low tones (Yip, 2002). In total, the distributed phonological representation consisted of 27 nodes (see Figure 1).

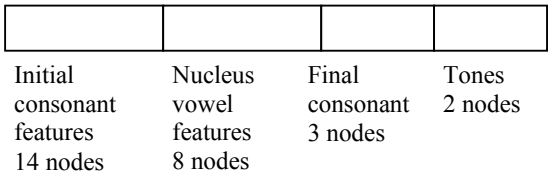


Figure 1: The phonological representation.

Orthographic Representations

Chinese characters consist of several individual strokes. There are some 20+ distinct strokes in Chinese orthography. Together, a few strokes may comprise a "stroke pattern", a recurrent orthographic unit of Chinese characters. Some stroke patterns can be characters by themselves. Units can be constructed recursively to form another composite unit. Those units that are integral stroke patterns and cannot be further decomposed into other units have been referred to as *single bodies* (Chen et al, 1996).

Researchers have long believed that Chinese character recognition starts from an analysis of features and the number of individual strokes (e.g., Seidenberg, 1985), in contrast with letters in alphabetic writing systems. In recent years, researchers have found evidence that this recognition by skilled readers is based upon well-defined orthographic constituents, instead of individual strokes (Chen, Allport, and

[2] Some dialects in China, such as Cantonese or Southern Min, may have more than four different tones.

Marshall, 1996; Zhou & Marslen-Wilson, 1999). Hence, in the orthographic representation, we used the basic stroke patterns defined in Cangjie, a Chinese transcription system developed by Ban-fu Chu in 1978. From a database analysis, there are 179 such stroke patterns comprising the radicals of all left-right structured Chinese phonetic compounds (Hsiao & Shillcock, *in preparation*). Hence, we used these 179 stroke patterns to encode the orthographic representation of the Chinese characters whose pronunciation we modelled.

Training and Test Corpora

The training corpus contained all left-right structured Chinese phonetic compounds and all their radicals which exist as characters on their own. During training, each character was presented according to its log token frequency, taken from a Chinese lexical database (Hsiao & Shillcock, *in preparation*). The database contains about 3,000 of the most frequent Chinese phonetic compound characters. Among them there are 2,159 left-right structured phonetic compounds and 880 radicals that are also existing characters. The test corpus contained the same phonetic compounds, but not the radicals on their own.

Network Architecture

Anatomical evidence has shown that the human fovea is precisely split about a vertical midline: when an alphabetic word or a Chinese character is fixated, the parts to the left and right of the fixation point are directly projected contralaterally. In modelling Chinese character recognition, we initially abstracted from real fixation behaviour and assumed that a character consisting of a semantic and a phonetic radical side by side could receive three possible fixations (see Figure 2). Characters were presented in the three fixation positions equally frequently during training. The task for the model, as for the reader, was to coordinate the information across the hemifields/hemispheres (Shillcock *et al.*, 2000).

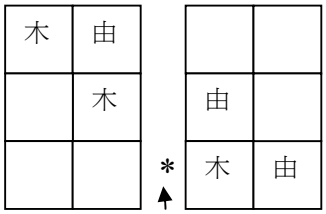
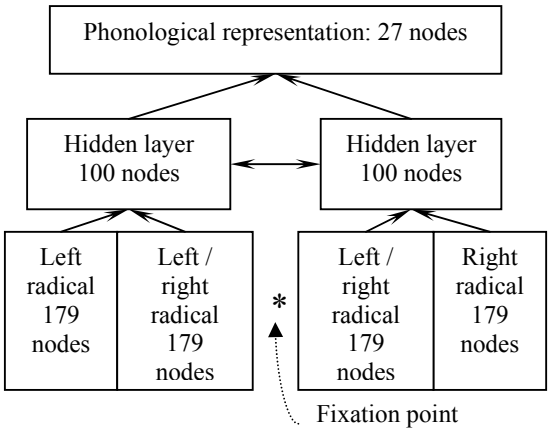


Figure 2: The complete pattern of inputs.

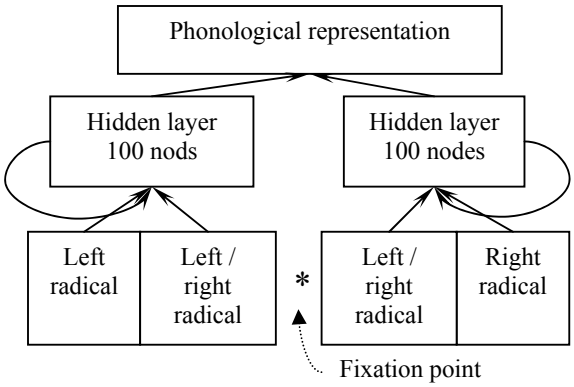
The network consisted of three layers. Adjacent layers were fully connected. Input units were localist representations of stroke patterns, capturing the claim that stroke patterns are functional units of character recognition. The characters were all represented in each of the three positions necessary to accommodate the input schema shown in Figure 2. Each position represented each of the 179 possible stroke patterns. The input was mapped, via a hidden layer, onto a feature-level phonological output. For characters with more than one pronunciation, only the most frequent pronunciation was employed.

The model is shown in Figure 3. To solve the task, "interhemispheric" communication is necessary, in the form of "callosal" connections between the two sets of hidden units.



Total number of links: 97,000

Figure 3: The split-fovea model for mapping Chinese orthography to phonology, with callosal connections.



Total number of links: 97,000

Figure 4: The model with no callosal connections.

Figure 4 shows a comparison model with no callosal connections in the hidden layers, which was trained on

the same task. In order to compare the performance of the two different architectures, we equalized their computational power by putting recurrent links on the hidden layers of the model with no callosal connections. Hence, both models had identical parameters and numbers of weighted connections. Thus, the principal difference between the models was the network architecture. We report elsewhere the more comprehensive comparison with a non-split model. The learning algorithm was discrete back propagation through time (Rumelhart, Hinton, and Williams, 1986).

Results

We ran each of the two different models three times and analyzed their average performance. Figure 5 shows the performance of the two models on regular and irregular characters, in terms of summed square error (SSE) at different stages during training. Neither of the two models had difficulty learning this task well. The split architecture encouraged the model to discover the formal similarities within the radicals in the two halves of the characters; that is, that most phonological information came from the right half of the characters. The divided visual system fortuitously mirrored this distinction in the orthography.

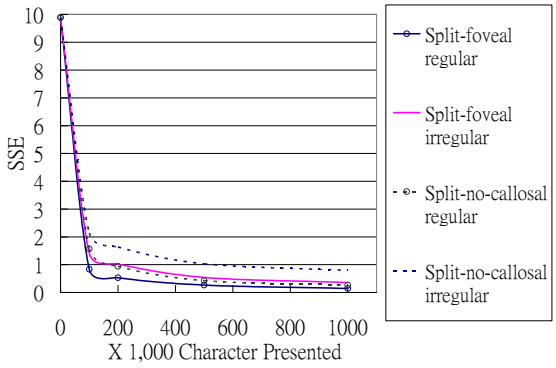


Figure 5: Performance of different models on regular and irregular characters.

The implemented split-fovea model provides an approach to understanding dyslexia in terms of hemispheric desynchronization (Shillcock & Monaghan, 2001). In the current simulations, the split-fovea model with callosal connections outperformed the model with no callosal connections (equivalent to extreme hemispheric desynchronization) on both regular and irregular characters; it especially exhibited more difficulty learning irregular characters, which constitute an XOR-like problem for the model with no callosal connections. Chinese surface dyslexics demonstrate reading impairments similar and analogous to those of dyslexics in alphabetic languages: poorer performance reading irregular characters (Yin & Butterworth, 1992). Hence, the implemented split-fovea model provides cross-language support for the hemispheric desynchronization account of dyslexia.

The model with no callosal connections made regularization errors on irregular characters, as we might predict from the nature of the problem it faced. Table 1 shows some examples of such regularization errors. As can be seen, most characters were mistakenly pronounced exactly like their phonetic radicals; some were given the same pronunciation but with a different tone. Interestingly, we found some which were pronounced as other irregular characters with the same phonetic radical (e.g., 俗 in Table 1). This shows that the pronunciation of an irregular character was not only affected by its phonetic radical, but also by orthographic "neighbours" which share the same phonetic radical.

Character	Correct pronunciation	Generated pronunciation	Phonetic radical pronunciation
猜	cai1	qing1	qing1 (青)
帖	tie3	zhan4	zhan4 (占)
横	heng2	huang2	huang2 (黄)
俗	Su2	yu4 (欲, 裕)	gu3 (谷)
沙	sha1	shao2	shao3 (少)
冶	ye3	tai2	tai2 (台)
杯	bei1	bu4	bu4 (不)

Table 1: Examples of regularization errors generated by the split model with no callosal connections.

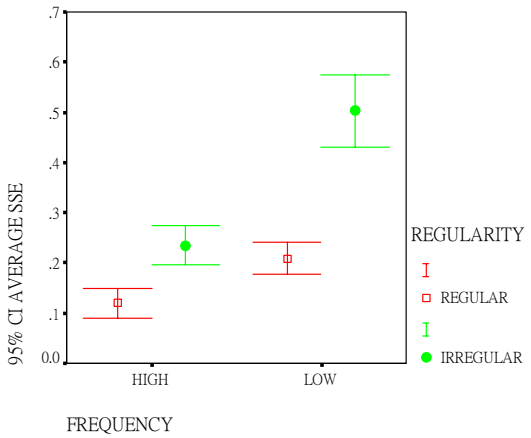


Figure 6: Interaction between frequency and regularity in the model with callosal connections after two million character presentations.

Figure 6 shows the interaction between frequency and regularity effects in the split-fovea model with callosal connections, after two million character presentations. This same interaction has been shown in experiments on Chinese character recognition (see, e.g., Shu et al, 2000; Hue, 1992; Liu, Wu & Chou, 1996; Seidenberg, 1985.). The model also produced this behaviour: the regularity effect was clearer among low frequency characters; there was a significant interaction between regularity and frequency (ANOVA analysis, $F(1,1075) = 16.296$, $p < 0.001$). The same significant interaction was also found in the version of the model with no callosal connections ($F(1,175) = 6.809$, $p < 0.01$).

We also examined the model's behaviour on SP and PS characters. It showed that there was no significant difference in the average SSE between the two groups in both split models with and without callosal connections ($F(1,2155) = 1.730$, $p > 0.05$; $F(1,2155)=2.117$, $p > 0.05$). A significant interaction between position of the phonetic radical (i.e. SP or PS characters) and regularity was also found in both models ($F(1,2155) = 4.719$, $p <0.05$; $F(1,2155) = 5.479$, $p <0.05$. See Figure 7 and 8). In the split model with callosal connections, there was a significant regularity effect among SP characters ($F(1,1940) = 127.486$, $p < 0.001$), but not among PS characters ($F(1,215) = 3.048$, $p > 0.05$). This may reflect the fact that only 24% characters are regular in the PS group, compared with 39% in the SP group (Hsiao & Shillcock, *in preparation*). On the other hand, the split model with no callosal connections did not exhibit the same behaviour: there were significant regularity effects among both SP characters ($F(1,1940) = 140.654$, $p < 0.001$) and PS characters ($F(1,215) = 6.493$, $p < 0.001$. See Figure 8). Here the modelling makes a testable prediction regarding human behaviour. Elsewhere we verify this prediction (Hsiao & Shillcock, *submitted*).

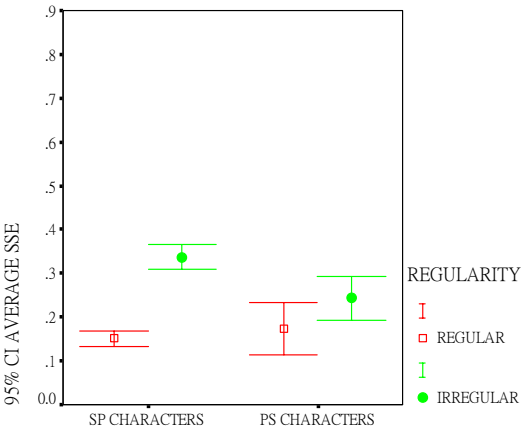


Figure 7: Interaction between position of phonetic radicals and regularity of characters in the split model with callosal connections.

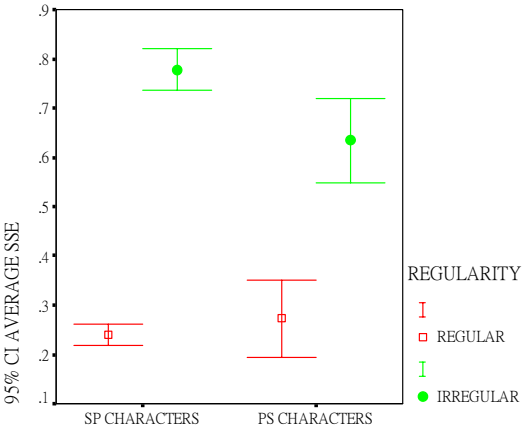


Figure 8: Interaction between position of phonetic radicals and regularity of characters in the split model with no callosal connections.

Conclusion and Discussion

We have presented a connectionist model of Chinese character recognition, an extension of the anatomically based split-fovea model, and we have compared the processing of Chinese phonetic compounds in architectures with and without callosal connections. We have incorporated several simplifications concerning the nature of the orthographic input and fixation behaviour, but several dimensions of our modelling have been of a psychologically realistic scale and the modelling has succeeded in capturing a number of behaviours and also in making experimentally testable predictions.

On the task of orthography to phonology mapping, the split-fovea architecture facilitates the network's discovery of the relationship between character substructure and pronunciation. The split architecture fortuitously corresponds to the major functional division in the stimuli we have used. This modelling further demonstrates the potential value of incorporating the anatomical constraints of the visual pathways into the computational modelling of reading: the requirement of a staggered input (Figure 2) effectively parses the stimuli (a process that is more apparent in modelling the reading of alphabetic inputs).

Also, we have examined the performance of the model with no callosal connections and found behaviour similar to that of Chinese surface dyslexics. The performance of the "callosally impaired" model is worse than the split-fovea model especially on irregular characters. A further examination showed that most

errors made were regularization errors, which matches the behaviour of surface dyslexics. The modelling hence provides cross-language support for the hemispheric desynchronization account of surface dyslexia.

The model also has made some testable predictions from its performance. It shows that the regularity effect is more salient among characters with their phonetic radicals on the right than on the left. This interaction reflects a statistical fact that there is greater regularity among characters with phonetic radicals on the right. Hence, these phonetic radicals become better cues for pronunciation.

References

Chen, Y., and Peng, D. (1994). A connection model of chinese character recognition and pronunciation. *Advances in the study of chinese language processing*, pp.211-240.

Chen, Y. P., Allport, D. A., & Marshall, J. C. (1996). What are the functional orthographic units in Chinese word recognition: The stroke or the stroke pattern? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 49(4), 1024-1043.

Coltheart, M. (1981). Disorders of reading and their implications for models of normal reading. *Visible Language, 15*, 245-286.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences, 5*, 82-88.

Fendrich, R. & Gazzaniga, M.S. (1989). Evidence of foveal splitting in a commissurotomy patient. *Neuropsychologia*, Vol. 27, No. 3, pp. 273-281.

Goswami, U., & Bryant, P.E. (1990). *Phonological skills and learning to read.* Hillsdale, NJ: Erlbaum.

Chinese characters: A Genealogy and Dictionary by Harbaugh

Harm, M. W. & Seidenberg, M.S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psych. Rev.*, *106*, 491–528.

Hsiao, J. H. & Shillcock, R. (2004). Analysis of A Chinese lexical database (*in preparation*).

Hsiao, J. H. & Shillcock, R. (2004) Regularity Effect in Naming Chinese Phonetic Compounds with the Phonetic Radical on the Left or Right (*submitted*)

Hue, C. W. (1992). Recognition processes in character naming. In *Language Processing in Chinese*, H.C. Chen and O.J.L. Tzeng (Eds.) 1992 Elsevier Science Publishers B.V.

Li, Y. & Kang, J. S. (1993). Analysis of phonetics of the ideophonetic characters in Modern Chinese. In Y. Chen (ed.), *Information analysis of usage of characters in Modern Chinese*, pp.84-98. Shanghai: Shanghai Education Publisher. (In Chinese)

Liu, I. M., Wu, J. T., & Chou, T. L. (1996). Encoding operation and transcoding as the major loci of the frequency effect. *Cognition*, 59, 149-168.

Mandarin consonants and vowels. (n.d.). Retrieved May 13, 2004, from http://personal.cityu.edu.hk/~cttomlai/doc/teach/stuff/mancon.htm

Perfetti, C.A. & Tan, L. (1999). The constituency model of Chinese word identification. In J. Wang, A. Inhoff & H. Chen (Eds.) *Reading Chinese Script*, Erlbaum: London, 115-134.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the microstructure of cognition; Vol. 1: Foundations*, Cambridge, Massachusetts. The MIT Press.

Seidenberg, M.S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*, 1-30.

Shillcock, R., Ellison, T. M., and Monaghan, P. (2000). Eye-Fixation Behavior, Lexical Storage, and Visual Word Recognition in a Split Processing Model. *Psychological Review, 2000,* Vol. 107. No. 4, 824-851.

Shillcock, R. C. & Monaghan, P. (2001). Connectionist modelling of surface dyslexia based on foveal splitting: Impaired pronunciation after only two half *pint*s. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Edinburgh: LEA. pp.916-921.

Shu, H., Anderson, R.C., and Wu. (2000). Phonetic Awareness: Knowledge of Orthography-Phonology Relationships in the Character Acquisition of Chinese Children. *Journal of Educational Psychology*, Vol. 92, No. 1, 56-62.

Wang, W. S. Y. (1973). The Chinese language. In Freeman (Ed.), *Readings from Scientific American: Language, Writing and the Computer* (pp. 50-60): Scientific American.

Yin, W. & Butterworth B. (1992). Deep and Surface Dyslexia in Chinese. In H.C. Chen and O.J.L. Tzeng (Eds.), *Language Processing In Chinese*, pp. 349-366.

Yip, M.J. (2002). *Tone*. Cambridge University Press.

Zhou, X. & Marslen-Wilson (1999). Sublexical Processing in Reading Chinese. In J. Wang, A. Inhoff & H. Chen (Eds.) *Reading Chinese Script*, 37-63, Erlbaum: London.

How speech processing affects our attention to visually similar objects: Shape competitor effects and the visual world paradigm

Falk Huettig (f.huettig@psych.york.ac.uk)
M. Gareth Gaskell (g.gaskell@psych.york.ac.uk)
Philip T. Quinlan (p.quinlan@psych.york.ac.uk)
Department of Psychology, University of York
York, YO10 5DD, United Kingdom

Abstract

It was investigated how spoken language is mapped onto the mental representations of objects in the visual field. Specifically, the visual world paradigm was used to test the hypothesis that during 'passive' listening tasks attention is directed more towards objects in the visual field that match the physical shape of the concept of the word concurrently heard than towards objects that do not match on physical shape. Participants listened to sentences containing certain critical target words of concepts with a typical shape (e.g. '*snake'*) while concurrently viewing a visual display of four objects. We found that participants tended to fixate conceptually unrelated objects with a similar physical shape (e.g. cable) as soon as information from the target word (e.g. '*snake'*) started to acoustically unfold. The results indicate that (contrary to some priming studies, e.g. Moss et al., 1997) shape information is accessed long before the offset of the spoken word. We discuss the findings with respect to the applicability of the visual world paradigm for the investigation of the access of lexical representations and theories of active vision.

Humans 'translate' between spoken language and concurrent visual input in such a natural way that we are hardly ever consciously aware of the processes involved. Surprisingly, there has been little research that has attempted to explore explicitly the interaction of these processes. A notable exception has been the research within the 'visual world paradigm' (the measuring of eye movements around a visual scene or display of objects in response to concurrent speech: Cooper, 1974; Tanenhaus et al. 1995) using both linguistic and visual contexts. Cooper (1974) established that when participants were presented simultaneously with spoken language and a visual field containing referents of the spoken words, participants tended to spontaneously fixate the visual referents of the words currently being heard. For example, participants were more likely to fixate the picture of a snake when hearing part or the entire word '*snake'* than to fixate pictures of unrelated control words. Cooper (1974) also found that participants were more likely to fixate pictures showing a lion, a zebra, or a snake when hearing the semantically related word '*Africa'* than to fixate semantically unrelated control words. Cooper's (1974) early study thus established two main findings: first, during the acoustic duration of a spoken word participants show a strong tendency to fixate objects that the word refers to. Second, his study highlighted the influence of semantic relationships on language-mediated fixation behavior: participants are more likely to fixate a visual referent that has some semantic relationship with the word heard than a semantically unrelated visual referent (see Huettig & Altmann, 2004; Yee & Sedivy, 2001; for follow-up studies). The primary goal of research in the visual world paradigm following Cooper's (1974) pioneering study has been to use eye movements as a tool to shed light on linguistic processing. For example, Allopenna, et al. (1998) asked participants to '*Pick up the candy. Now put it …*' in the context of a visual display of objects including (among other things) a candy and a candle. They found evidence for a phonological competitor effect: eye movements to both the candy and the candle increased as the word '*candy'* acoustically unfolded but that soon after its acoustic offset, looks to the candle decreased while looks to the candy continued to rise. The Allopenna et al. (1998) study provided evidence for a standard competitor effect as predicted by theories of auditory word recognition such as TRACE (e.g. McCleland & Elman, 1986). Importantly the study demonstrated this effect in real-time as the speech stream was unfolding acoustically (see also Dahan et al., 2001; for more evidence that the visual world paradigm provides fine-grained measures of lexical processing).

However, far less attention has focused on examining the interaction of spoken language with directed attention and the *visual* properties of the presented objects. In this regard, Cooper (1974) also found that participants tended to fixate a picture of a snake when hearing the word '*wormed'* (in the context '*just as I had wormed my way on my stomach*'). This finding (although not discussed by Cooper) suggests that there may also be a strong link between lexical processing and the visual properties of an object such as an object's shape (although it cannot be ruled out that in Cooper's experiment participants mistook the snake for a worm and therefore directed their attention to the picture of the snake when hearing '*wormed'*).

The visual world paradigm and the access of lexical representations Importantly, Cooper's study (1974) is indicative that similar processes to semantic priming are taking place when people map spoken words onto related visual objects. The semantic priming paradigm (Meyer & Schvanefeldt, 1971) has proven to be particular useful for

the investigation of lexical representations. The currently dominant view is that a word's representation is composed of smaller units (or 'features') of different kinds that are accessed during spoken word recognition. Recent distributed models of spoken word recognition (e.g. Gaskell & Marslen-Wilson, 1997) assume that some aspects of a word's meaning may be activated more rapidly than others resulting in a dynamic pattern of changes in the semantic properties throughout the duration of the spoken word. Evidence supporting this notion comes, for example, from priming studies by Moss et al. (1997) that found a significant priming effect for functional properties of words early during the duration of the word but priming for perceptual targets (e.g. the shape overlap between *hook* and *curve*) only at the offset of the prime word. In order to investigate these time-course issues the visual world paradigm may be particularly useful because of the closely time-locked, fine-grained measures the method provides.

In the current study we explored how the physical shape of objects in a visual display interacts with language-directed attention. Participants' eye movements were measured, during the acoustic duration of certain target words (concepts with a typical shape, e.g. '*snake'*), to conceptually unrelated visual objects that have a similar shape (e.g. the image of a cable). If participants shift their attention to conceptually unrelated objects with a similar shape (e.g. cable) when the word '*snake*' unfolds during online speech, then the inspection of the time-course of fixation probabilities should shed light on the issue whether (lexical) shape information is accessed only at word offset or before. In other words, if (lexical) 'perceptual' information such as shape is not accessed before the offset of the spoken word then the prediction is that there should be no increased attention to shape competitors (e.g. cable) *before* the offset of the acoustic target words (e.g. '*snake'*).

The visual world paradigm and 'active vision' Our study explored the effect of 'shape overlap' between spoken words and visual objects on *overt* attention. Relevant in this regard is that most vision research has focused on what Findlay & Gilchrist (2003) term 'passive vision': the assumption that image interpretation is largely passive and that parallel processing occurs across the visual image with algorithms charting the "progress from a grey-scale retinal input to an internal representation in the head". Findlay & Gilchrist (2003) reject this view in favor of 'active vision': the notion that *overt* gaze orienting is an essential and crucial feature of vision. This approach emphasizes the importance of re-directing attention overtly (rather than covertly) by moving the gaze in order for the attended location to obtain the instant benefit of high-resolution foveal vision. These proposals are similar to the notion that the perceptual system offloads information by leaving it in the environment rather than just passively passing information on to the cognitive system for propositional representations to be created. According to this view, perceptual information in the environment is accessed when needed, with the visual world functioning as a kind of external memory (e.g. O'Regan, 1992). Objects in this situated memory are represented in a spatial data structure which contains 'pointers' to the real-world location of the object. Thus, the system need not store internally detailed information about the object, but can instead locate that information, when it has to, by directing attention back to that object in the environment. Essentially, the focus in active vision research is on understanding why and when gaze is re-directed. In other words active vision places vision in a context. And one important variable that impacts on (and guides) active vision is spoken language.

Importantly, there was one second preview of the visual display in our study and the target words unfolded approx. five seconds after the onset of the visual display. This means that all four objects were fixated (usually several times) before the onset of the target word. Therefore the prediction is that on 'passive vision' accounts, arguably, there is no need for an *immediate* shift in overt attention towards a conceptually *unrelated* object (e.g. cable) when the word '*snake'* unfolds. In other words on 'passive vision' accounts all relevant information has already been encoded and is available to the system for further cognitive processing. On active or situated vision accounts, however, the prediction is that on hearing the target word, overt attention will be re-directed to the shape competitor to retrieve more information about that object to establish its fit with the specification provided by the target word.

Methods

Participants 48 participants from the University of York student community took part in this study. All were native speakers of British English and had either uncorrected vision or wore soft contact lenses or glasses.

Materials The experiment made use of three conditions. 21 items were created consisting of two types of spoken sentences containing a target word (e.g. *'snake'*) for two sets of visual stimuli. In the 'target set' the visual stimulus was a picture depicting a fully matching referent (e.g. a snake) for the acoustic target word plus three distractors depicting objects from different conceptual categories. The stimuli in the 'shape competitor set' consisted of the same four pictures, in identical positions, except that the target picture (e.g. the snake) was replaced by a picture depicting an object with a similar shape as the target word (e.g. a cable, Figure 1).

The sentential stimuli were constructed for three conditions: neutral sentences with the pictures of the 'target set' (the neutral condition), sentences biasing the visual target referent (biasing snake) with the pictures of the 'target set' (the biasing condition), and the same sentences as in the biasing condition but where the picture of the target object (e.g. snake) was replaced by a picture of a shape competitor (the competitor condition, e.g. the picture of a cable). The rationale for presenting the shape competitor in a biasing context was simply to make it relatively unlikely that participants would anticipate, prior

to the target word, that the shape competitor would be the object of attention (even though it was not going to be referred to directly). The neutral sentences were included in order to establish a baseline against which the efficacy of the biasing context could be determined.

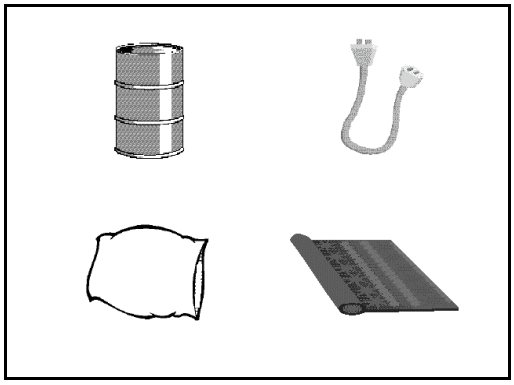


Figure 1. Visual stimulus in the competitor condition (depicting: physical shape competitor of the acoustic target word '*snake'*: cable, 3 distractors)

In sum, for the neutral condition, the sentence did not provide any contextual bias up until the target word that would favor any of the pictures depicted in the visual scene ('*In the beginning, the man watched closely, but then he looked at the snake and realized that it was harmless*'). In the biasing condition, the sentence was constructed to contextually bias towards the depicted target object (e.g. the snake): '*In the beginning, the zookeeper worried greatly, but then he looked at the snake and realized that it was harmless*'. In the competitor condition the sentence was identical to the biasing condition. However, the picture of the shape competitor (e.g. cable) was semantically unrelated to the target word (e.g. *'snake'*) and thus the sentential context did not provide any contextual bias towards the picture of the shape competitor. The target-competitor pairs were: anchor/arrow, apple/moon, banana/sword, bell/hat, button/coin, candle/tube, cigar/carrot, chimney/rocket, dice/ice cube, football/planet, globe/orange, horseshoe/magnet, lighthouse/flask, microphone/cone, mirror/frame, pencil/column, plate/wheel, racket/saucepan, scissors/chopsticks, snake/cable, wheelbarrow/sledge.

Norming study In order to determine the relative similarity in physical shape of the target concept activated by the acoustic target word with the depicted objects a norming study was conducted. Twelve participants provided normative data. Participants were presented with the written target word (e.g. *snake*) and the actual visual items. Participants were asked to judge how similar the typical physical shape of the target concept (*snake*) was with the physical shape of the depicted objects on a scale from 0 to 10 (zero representing: 'absolutely no similarity in physical shape', 10 representing: 'identical in physical shape'). The mean similarity for the shape competitors was 7.1 (SD = 1.8) and for the distractors 1.4 (SD = 0.7). These differences in the shape similarity judgments between the shape competitors and the visually dissimilar distractors were highly significant ($F1$ (1, 11) = 268.89, MSE = 0.07, $p < 0.01$; $F2$ (1, 20) = 200.35, MSE = 0.17, $p < 0.001$).

Procedure and Design There were 21 experimental items (counterbalanced across the three conditions). For 14 of the experimental items per participant the visual stimulus included a visual referent matching the full 'target specification' of the target word (e.g. the acoustic word '*snake*' and the picture of a snake in the 'neutral condition' and the 'biasing condition'). For 7 of the experimental items there was no picture matching the full 'target specification'. For these items there was only a 'physical shape match' between the acoustic target word and the shape competitor picture (e.g. the acoustic target *snake* and the picture of a cable in the 'shape competitor condition'). 15 additional filler items were added, which all included a fully matching visual referent of an acoustic target word. There were four practice trials. Thus 82% of the 40 trials included a *fully* matching target visual referent (e.g. hearing '*snake'* and seeing a snake). This design made it very unlikely that the participants were able to note the physical shape relationship and adopt a conscious strategy accordingly. In addition, participants consistently stated in self-report that they neither moved their eyes according to some kind of explicit strategy nor noticed the 'shape manipulation'.

Participants were seated at a comfortable distance in front of a 17' display and wore an SMI EyeLink head-mounted eye-tracker, sampling at 250Hz from the right eye (viewing was binocular). They were told that they should listen to the sentences carefully. They were also told that they could look at whatever they wanted but were asked not to take their eyes off the screen throughout the experiment. The onset of the visual stimulus was one second before the onset of the spoken stimulus. The onset of the acoustic target word was on average 4 seconds after the onset of the spoken sentence and thus the acoustic target word started to unfold on average 5 seconds after the onset of the visual stimulus. The entire experiment lasted approximately twenty minutes. Participants' eye movements were recorded as they listened to the sentences.

Results

Fixation probabilities: *p*(fix) The probability of fixating a type of picture at a defined moment in time, *p*(fix), will be reported. The visual display consisted of four quadrants, each with one object. Gaze positions were categorized by the quadrant in which an object was depicted. The *a priori* probability of fixating one of the four pictures in absence of any bias was thus 0.25.

Table 1: *P*(fix) at the acoustic onsets and offsets of the target words (e.g. '*snake'*) and difference scores per condition

condition	neutral		biasing		competitor	
type of picture	target (snake)	distractor	target (snake)	distractor	competitor (cable)	distractor
p(fix) at onset	.27	.24	.39	.19	.25	.24
difference score at onset	.04		0.21		.01	
p(fix) at offset	.50	.15	.52	.14	.33	.22
difference score at offset	.35		.38		.11	

Table 1 shows *p*(fix) for the type of picture at the time points of critical interest: the acoustic onset of the target word (e.g. the acoustic onset of '*snake'*), and the offset of the target word (e.g. the acoustic offset of '*snake'*). The probability to fixate the three distractors was averaged to obtain one distractor value. Note that we did not add any time to account for the time it takes to program a saccade. All measures and analyses were based on the *real* acoustic time points. Figure 2 shows the time-course of *p*(fix) in the three conditions from the acoustic onset of the target word for 1000 ms.

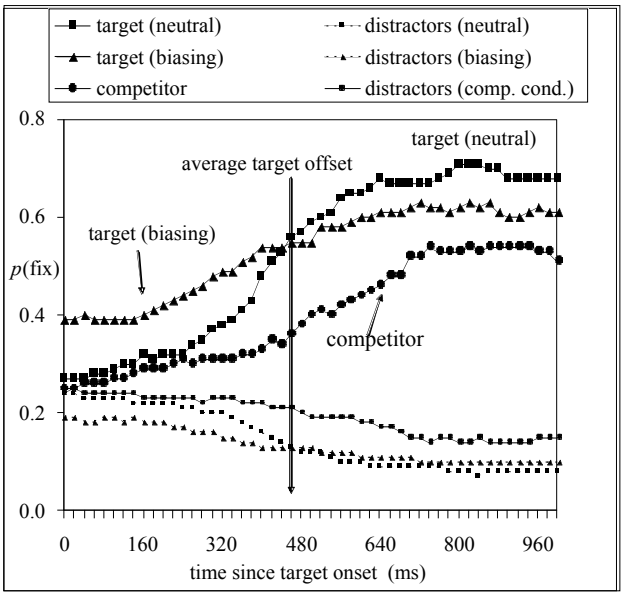


Figure 2. Time-course of *p*(fix) to the target in the neutral condition and the biasing condition, and to the shape competitor in the competitor condition (and averaged distractors of each condition).

Fixations of the different types of pictures at the acoustic onset of the target word is of interest in order to assess whether there were any biases in attention before information from the critical target word (e.g. '*snake'*) became available. It was predicted that at this point there would be no such bias in the neutral condition and the competitor condition if the context had been neutral with respect to directed attention to any of the pictures. However, it was predicted that at the acoustic onset of the target word there would be a bias towards the target picture (e.g. the snake) in the biasing condition because of the biasing sentential context. These predictions are apparently born out by the data. Table 1 shows that *p*(fix) at the onset of the target word was around 0.25 in the neutral and competitor conditions but that there was a strong bias towards the target object in the biasing condition. The acoustic offset of the target word reflects the point when the entire spoken target word has been heard by the participants. Fixations at this point are of interest in order to assess whether the acoustic unfolding of the target word resulted in changes in overt attention. Table 1 and Figure 2 show that the target and competitor fixations had increased in the neutral and the competitor conditions, whereas in the biasing condition the probability to fixate the target increased further. Nonetheless, Table 1 and Figure 2 suggest that *p*(fix) of the targets in the neutral and the biasing conditions was higher than *p*(fix) of the shape competitors in the competitor competition. In other words as acoustic information from the target words became available the probability to fixate the target picture *and* the shape competitor picture increased. However, *p*(fix) of the target pictures increased much more than *p*(fix) of the shape competitors.

Statistical analyses In order not to violate statistical assumptions (in particular that pertaining to the independence of observations), difference scores obtained in each condition are compared For instance to assess a bias to look at a critical picture (target or competitor vs. a distractor referent) the differences in fixation probabilities to these stimuli are considered. Such difference scores reveal both the magnitude and direction of the effects. In the current study *p*(fix distractor) was subtracted from *p*(fix target) and *p*(fix competitor). Any positive difference reveals a bias of looks towards the critical picture, a negative difference reveals a bias of looks towards the distractors, and difference scores close to zero reveals neither bias. The use of error bars in the form of 95% confidence intervals plotted around the sample means provides a quantitative visual representation of the faith that should be placed in the pattern of sample means as an estimate of corresponding patterns of population means. Figure 3 shows the mean of the difference scores for participants and items at the acoustic onset of the

target words (e.g. '*snake'*) in the three conditions. Error bars represent the 95% confidence intervals of the means computed individually for each difference score.

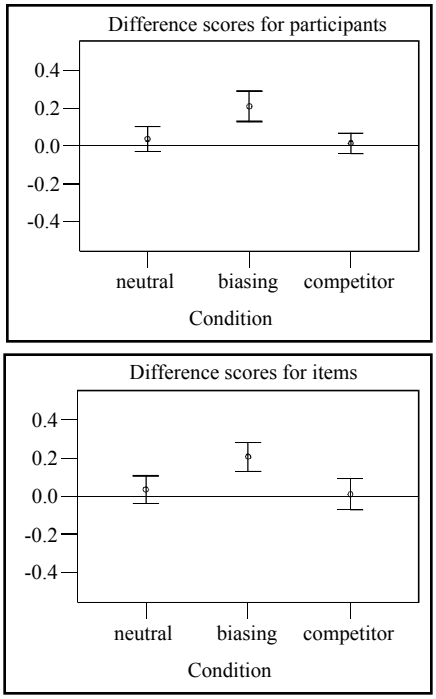


Figure 3. Means of the difference scores (participants and items) at the acoustic onset of the target words (Error bars represent the 95% confidence intervals)

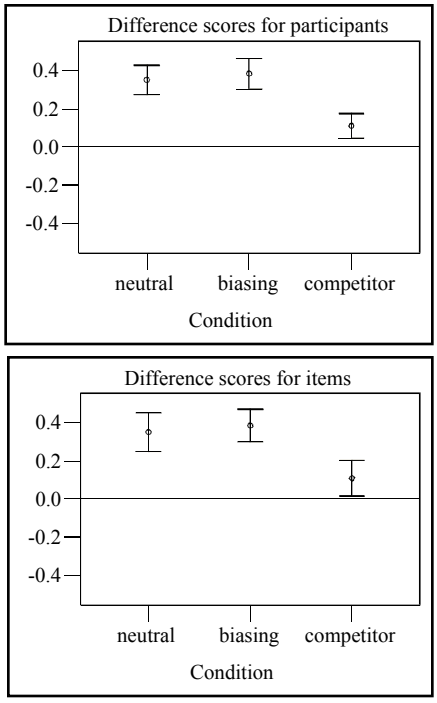


Figure 4. Means of the difference scores (participants and items) at the acoustic offset of the target words (Error bars represent the 95% confidence intervals)

Figure 3 shows that at target onset there were no reliable differences between looks to the critical pictures and the distractors in the neutral and the competitor conditions. In other words there were no differences in directed attention for type of picture at the acoustic onset of the target. However, there was a reliable bias in directed attention towards the target object in the biasing condition. Thus the contextual manipulation had been successful.

Figure 4 shows the mean of the difference scores for participants and items at the acoustic offset of the critical words. There was a reliable bias in directed attention to the critical pictures in all three conditions. Importantly, there was a statistically robust higher probability to fixate the shape competitor (e.g. the cable) than the distractors at the acoustic offset of the target word (e.g. '*snake'*).

General Discussion

Our findings directly link online conceptual processing during lexical access in speech to attentional behavior in the visual world. They extend Cooper's (1974) work by showing that during 'passive' listening tasks attention is also directed significantly more towards objects that match the shape of the word concurrently heard than towards objects that do not match on shape.

Importantly, the competitor effect was significant at the offset of the acoustic target words. This means that the shape competitor effect started to occur as soon as information from the target word acoustically unfolded given that the average duration of the target words was 447ms and that the *minimum* latency to program and initiate a saccade is 150 to 200 ms (e.g. Saslow, 1967). This result is contrary to priming studies (Moss et al., 1997) that found activation of perceptual (including shape) information only at the offset of the prime word. The current study suggests that the visual world paradigm is particularly sensitive for capturing the access of conceptual and perceptual information during lexical processing. Note that Moss et al. (1997) included five different types of 'perceptual' properties in their study. A reason for the discrepancy to our results thus may be that Moss et al. (1997) did not distinguish between different 'perceptual' properties such as color and shape. In other words they may have found delayed priming for perceptual targets because of the differential properties of the items they selected for their perceptual condition. An alternative explanation is that the information (from the spoken words) provided for the attentional system to visual objects involves such a tight 'loop' that other means of observing the access of (lexical) perceptual representations (e.g. the lexical decision task) can only do so at some delay. Similarly the activation in our study may have partly originated through spreading activation from shape information portrayed within the visual display. The shape information may have activated concepts sharing those features resulting in an earlier access of shape information during spoken word recognition. Our data do not show to what extent this activation may have originated from the visual display.

Notably, the present results are in line with the predictions derived from active or situated vision accounts that on hearing the target word, *overt* attention may be re-directed *immediately* towards only partly matching objects in order to retrieve more information and to establish their fit with the target specification as provided by the target word. Arguably, our findings coupled with the fact that there was one second preview of the visual display and that the target words unfolded approximatly five seconds after the onset of the visual display, cannot be as *easily* incorporated in 'passive vision' accounts.

Pickering, McElree, & Garrod (submitted) have recently proposed that participants may engage in a covert naming strategy in visual world experiments. Pickering et al. state that "many effects in this paradigm may be partly the result of participants' regularly naming the objects covertly… further research is needed to determine the extent to which visual world results depend upon linguistic recoding (covert naming) of the objects". The current data do not rule out that our participants on occasion named an object covertly. However, the immediate and robust shift in directed attention to a conceptually unrelated and clearly identifiable shape competitor with a different name (e.g. cable) on hearing the target word (e.g. '*snake'*) does *not* fit comfortably with their suggestion. The current study thus casts doubt on the claim that participants *regularly* name objects in visual world studies.

The shape competitor effects are unlikely to be limited to the passive listening task we employed. Dahan & Tanenhaus (2002) recently presented evidence that similar visual form competitor effects also occur when participants are required to engage in an explicit physical task (moving the objects mentioned in spoken sentences using a computer mouse). Our procedure of a 'passive' listening task is strong evidence that these perceptual competitor effects are not limited to certain specific 'goal-directed' task demands.

In sum, the findings are best compatible with the notion of a rich mapping process between spoken language and concurrent visual input. On a methodological note, the visual world paradigm promises to be a valuable research tool for investigations into the access of (lexical) perceptual and conceptual representations as well as into issues in visual perception.

Acknowledgments

FH would like to thank Gerry Altmann and Graham Hitch for discussion of this research.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419-439.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, 813-839.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time-course of frequency effects in spoken word recognition: Evidence from eye movments. *Cognitive Psychology*, *42*, 317-367.

Dahan, D. & Tanenhaus, M. K. (2002). Activation of conceptual representations during spoken-word recognition. Poster presented at the 43rd Annual Meeting of the Psychonomics Society, Kansas, USA.

Findlay J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford University Press.

Gaskell, M.G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, *12*, 613-656.

Huettig, F. & Altmann, G.T.M. (2004). The online processing of ambiguous and unambiguous words in context: Evidence from head-mounted eye-tracking. In M. Carreiras & C. Clifton (Eds.). *The On-line Study of Sentence Comprehension: Eyetracking, ERP and Beyond*. New York, NY: Psychology Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing words: Evidence of a dependence upon retrieval operations. *Journal of Experimental Psychology, 90*, 227-234.

Moss, H. E., McCormick, S. F., & Tyler, L. K. (1997). The time course of activation of semantic information during spoken word recognition. *Language and Cognitive Processes*, *12*, 695-731.

O'Regan, J. K. (1992). Solving the 'real' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, *46*, 461-488.

Pickering, M. J., McElree, B., & Garrod, S. (submitted). Interactions of language and vision restrict "visual world"interpretations.http://ila.psych.nyu.edu/users/bdm/Dept/index.html

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*, 1030-1033.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Yee, E., & Sedivy, J. (2001) Using Eye Movements to Track the Spread of Semantic activation during spoken word recognition. Paper presented to the 13th Annual CUNY Conference, Philadelphia, USA.

The Importance of Temporal Information for Inflection-type Effects in Linguistic and Non-linguistic Domains

Julie M. Hupp (hupp.34@osu.edu)
Center for Cognitive Science
1961 Tuttle Park Place
The Ohio State University
Columbus, OH 43210, USA

Vladimir Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
1961 Tuttle Park Place
The Ohio State University
Columbus, OH 43210, USA

Peter Culicover (culicover.1@osu.edu)
Department of Linguistics and Center for Cognitive Science
1712 Neil Avenue
The Ohio State University
Columbus, OH 43210, USA

Abstract

One of the important tasks of language acquisition is the ability to distinguish between an inflectional derivation from a target word, which is a variant of this word (e.g., *tool* → *tools*), and a completely new word (e.g., *tool* → *stool*). In an attempt to explain the ability to solve this problem, it has been proposed that the beginning of the word is its most psychologically salient portion. However, it is not clear whether this phenomenon is specific to language or whether it stems from a more general cognitive mechanism, with beginnings of sequences being more salient than endings. The three reported experiments were designed to answer this question. In these experiments, participants judged the similarity of test sequences to target sequences across three domains: linguistic, musical and visual. The test items were judged as more similar to an original target item if information was added to the end of that item rather than to the beginning of the item across all three domains. This suggests that there may be a more general cognitive mechanism underlying the well-documented suffixation preference, according to which changes in the end of the word are more readily interpreted as inflectional derivations from the target word.

Introduction

One of the important tasks of language acquisition is the ability to distinguish between an inflectional derivation from a target word, which is a variant of this word (e.g., *tool*→ *tools*), and a completely new word (e.g., *tool* → *stool*).

There are multiple types of inflections that exist across languages, including prefixation (e.g., adding a morpheme before the stem), suffixation (e.g., adding a morpheme after the stem), infixation (e.g., adding a morpheme inside the stem), and nonconcatenative devices (e.g., interleaving a string of vowels with a string of consonants).

Two types of inflections are frequently present in many European languages, prefixes and suffixes, and it has been established that suffixes are easier to acquire (e.g., interpret suffixation as an inflectional derivation) than prefixes. This finding is not specific to the English language. Cross-linguistically, the suffixing preference results in stems generally being ordered before the added morpheme because language users prefer to process stems before the added morpheme (Hawkins & Cutler, 1988). Overall, suffixing is more frequent than prefixing (Hawkins & Gilligan, 1988). A number of explanations have been proposed, although any single explanation alone may not fully account for this phenomenon.

First of all, there are positional differences with the addition of a morpheme at the beginning and the end of the word, which is important because words take place in time (Gasser, 1994). For example, words are often recognized before they are completed (Tyler, Marslen-Wilson, Rentoul, & Hanney, 1992). The information that reaches the ear first may be the key to the identification of that piece of information. If this is the case, then the temporal aspect of language may be the underlying reason for suffixation preferences in language.

Similarly, the psychologically most salient part of any word is its beginning portion (Clark 1991; Hawkins & Cutler, 1988). This is to say that the effect of distorting a word is more severe if the distortion is at the beginning of the word (e.g., prefix) rather than the end (e.g., suffix). This is true in both comprehension and production. In comprehension, adding a morpheme to the end of a word does not affect the recognition of the word, and in production, it is easier to produce a familiar sequence

followed by a modification in the form of a suffix than the reverse, a modification first and then the familiar word (Clark, 1991).

These contentions seem to be supported by the literature on language acquisition. In particular, Slobin (1985) claimed that children use procedures or strategies called Operating Principles (OP) in their linguistic development. He proposed many different principles that children use, but the one of importance for this area is OP (ATTENTION): BEGINNING OF UNIT. This principle states that children pay attention to the first syllable of an extracted speech unit. They store it separately and in relation to the unit with which it occurs. If a child were specifically attending to the beginning of a word, then adding a morpheme to the end of the word would be less detrimental to the recognition of that word than adding the morpheme to the beginning of the word.

Clark has done much research on the area of children and inflections that add support to suffixation preference from a developmental perspective. She has found that children acquire inflections from their earliest word use and continue to comprehend and produce them throughout their linguistic development (1995). In general, children begin to add noun and verb inflections between 18 and 24 months; however, they consistently learn suffixes before prefixes, even when these inflectional forms express equivalent information (1995; 1998). In addition, children aged 5 to 7 find nonsense suffixes are easier to imitate than nonsense prefixes (1998).

Overall, children seem to find it easier to process information added to the ends of words than to the beginnings, and it has been argued that the beginning of the word is its most psychologically salient portion. This would explain why children are better at learning suffixes rather than prefixes. These findings map quite well onto the adult research on inflectional morphology.

However, it is not clear whether this phenomenon is specific to language or whether it stems from a more general cognitive mechanism.

We suggest that this ability may reflect a more general property of processing of temporally organized information: changes in the beginning of a sequence are easier to detect than at the end of the sequence. If this is the case, then non-linguistic information that has temporal structure may also give rise to inflection-type effects, such that changes at the end of the sequence would more likely be considered variants of the original string than changes at the beginning of the sequence.

To test this hypothesis, we conducted three experiments, using language (Experiment 1), music (Experiment 2), and visual sequences (Experiment 3).

Experiment 1: Inflections in the Linguistic Domain

Method

Participants There were 17 participants in this experiment. The participants were undergraduate students from The Ohio State University who participated to fulfill a psychology course requirement. Five participants failed to correctly respond to at least 70 % of the catch items and were excluded from this experiment.

Design and Materials The stimuli consisted of 42 sets, with each set consisting of a 2-syllable artificial Target word followed by two Test words. One of the Test words was the Target with a morpheme added to the beginning (Test-Pre). The other Test word was the Target with a morpheme added to the end (Test-Post).

The Target words were constructed by randomly connecting discrete syllables (e.g., *Ta-Te*) with .06 sec between syllables (see Johnson & Jusczyk, 2001; Saffran, Aslin, & Newport, 1996, for details of stimuli creation). The Test words were created by either adding a syllable to the beginning of the Target word (Test-Pre: *BE-Ta-T*e), to the end of the Target word (Test-Post; *Ta-Te-BE*), adding nothing to the Target word (Test-Identical: *Ta-Te*), or changing the Target word completely (Test-Different: *Pu-La-Fi*).

On each trial, participants received a Target word, followed by two Test words (the order of each of the Test words was counterbalanced), and their task was to determine which of the Test words was more similar to the Target.

There were six types of sets determined by pairing of the types of Test words: Pre-Post, Pre-Identical, Post-Identical, Pre-Different, Post-Different, and Identical-Different. The first type was the focal interest (e.g., 25 Pre-Post sets), whereas the remaining 5 conditions were catch trials (3 sets for each condition, and 2 additional Identical-Different sets for the practice trials). The set types varied within participants.

Procedure Each participant received 2 randomly presented practice trials with a break to ask the experimenter any questions, and then the remaining 40 trials were presented randomly. *Presentation* software was used to deliver the instructions, present the stimuli and record the responses.

The participants were instructed that they would hear a 2-syllable Target word followed by two Test words, and they were to decide which of the Test words was more similar to the initial Target word. If the first Test word was most similar, they were to press "F" on the keyboard, and if the last Test word was most similar, they were to press "L". To start each new trial, they were instructed to press the space bar.

There was 1 sec in between each word, and the order of the Test words was counterbalanced across sets. The Target word was heard from both of the computer speakers

while the first Test word was heard only from the left speaker and the second Test word was heard only from the right speaker.

Results and Discussion

Overall, participants were accurate on catch trials, exhibiting over 90% accuracy (M = 94.90%), above chance, one-sample t (16) = 25.43, p < .001.

However, the analysis of participants' responses to Pre-Post items was of considerable interest. Data analyses focused on the percent of participants' responses in which the Test-Post item was considered more similar to the Target than the Test-Pre item. Overall, in more than 85% of responses (M = 88.00%) participants deemed the Test-Post item to be more similar to the Target than the Test-Pre item, above chance, one-sample t (16) = 9.64, p < .001. Thus, as expected there was a clear tendency to choose the Test-Post items as more similar to the original Target word than the Test-Pre words.

Having established that the procedure captures the effect in the domain of language, we conducted Experiments 2 and 3, using the same procedure with music tones and visual patterns.

Experiment 2: Inflection-type Effects in the Domain of Music

Method

Participants There were 18 participants in this experiment. The participants were undergraduate students from The Ohio State University who participated to fulfill a psychology course requirement.

Design and Materials The design was the same as in Experiment 1, except the sets were made up of a 2-note Target melody and two Test melodies. The Test items were created by adding notes to either the beginning (Pre) or the end (Post) of the Target melodies.

Procedure The overall procedure was identical to Experiment 1. The main exception was that instead of hearing words, the participants were instructed that they would hear a small Target musical melody followed by two Test melodies. From this, they were to decide which Test melody was the most similar to the original Target melody.

Results and Discussion

Overall, participants were accurate on catch trials for this experiment as well, exhibiting over 90% accuracy (M = 91.85%), above chance, one-sample t (17) = 20.35, p < .001.

Similar to Experiment 1, data analyses of central interest focused on the percent of participants' responses in which the Test-Post item was considered more similar to the Target than the Test-Pre item. Once again, the participants were more likely to choose the Test-Post items as more similar to the Target than the Test-Pre items (M = 71.56%), above chance, one-sample t (17) = 4.03, p = .001.

Experiment 3a: Inflection-type Effects in the Visual Domain

Method

Participants There were 17 participants in the visual domain. The participants were undergraduate students from The Ohio State University who participated to fulfill a psychology course requirement. Using the same exclusion criterion as in Experiment 1, 2 participants were eliminated from this experiment.

Design and Materials The design was the same as in Experiments 1 and 2. The stimuli in this experiment consisted of object sequences. There were a total of 25 objects that were randomly connected to form the Target sequences. The Target sequences were composed of either all red, blue, green or orange shapes. Each set consisted of a Target sequence made of two simple objects that flashed for 1 sec each while centered at the top of the computer screen (e.g., Cross, Heart).

Then, 1 sec later, the first of two Test sequences appeared at the bottom of the screen. There was 1 sec in between each Test sequence, and the order of the Test sequences was counterbalanced across sets. The first Test sequence appeared on the bottom left of the computer screen, and the second Test sequence appeared on the bottom right of the screen. The Test items were created by adding an object (e.g., Diamond) for 1 sec either at the beginning of the Target sequence (Test-Pre; Diamond, Cross, Heart), at the end of the Target sequence (Test-Post: Cross, Heart, Diamond), no change at all to the Target sequence (Identical: Cross, Heart) or changed the sequence completely (Different: Star, Light Bulb, Lock). The object that was added was of a different color than the Target sequence: a red Target sequence would have a blue object added (and vice-versa), and green and orange were similarly paired.

Procedure The overall set up of the experiment was similar to Experiments 1 and 2. In this experiment, the participants were instructed that they would see a Target sequence of objects on the top of the screen followed by two Test sequences on the bottom of the screen. They were to decide which Test sequence was more similar to the initial Target sequence.

Results and Discussion

Participants were accurate on catch trials with an overall accuracy over 95% (M = 98.04%), above chance, one-sample t (16) = 63.23, p < .001.

Similar to Experiments 1 and 2, participants were more likely to choose the Test-Post items as more similar to the Target than the Test-Pre items (M = 91.53%), above chance, one-sample t (16) = 13.72, p < .001.

Having established that this effect is present in the visual domain, it was important to investigate the effect of temporal information in this domain. Therefore,

Experiment 3b was conducted as a control experiment for the visual domain without the addition of temporal information.

Experiment 3b: The Visual Domain without Temporal Information

Method

Participants There were 18 participants in the visual control condition. The participants were undergraduate students from The Ohio State University who participated to fulfill a psychology course requirement.

Design and Materials The design was the same as in the previous experiments. The stimuli in this experiment consisted of the same object sequences that were used in Experiment 2 without the addition of temporal information. That is to say that the participants viewed a row of stationary shapes instead of a dynamic sequence of shapes.

The Target appeared at the top of the screen while the Test items simultaneously appeared at the bottom of the screen. Once again, the positioning of the Test sequences was counterbalanced across sets. One Test sequence appeared on the bottom left of the computer screen, and one Test sequence appeared on the bottom right of the screen. Similar to Experiment 3a, the Test items were created by adding an object (e.g., Diamond) either to the left of the Target sequence (Test-Pre; Diamond, Cross, Heart,), to the right of the Target sequence (Test-Post: Cross, Heart, Diamond), no change at all to the Target sequence Identical: Cross, Heart) or changed the sequence completely (Different: Star, Light Bulb, Lock).

Procedure The overall set up of the experiment was similar to the previous experiments. In this experiment, the participants were instructed that they would see a Target sequence of objects on the top of the screen and two Test sequences on the bottom of the screen. They were to decide which Test sequence was more similar to the initial Target sequence.

Results and Discussion

Participants were accurate on catch trials with an overall accuracy over 95% (M = 95.55%), above chance, one-sample t (17) = 42.26, $p < .001$.

Similar to Experiments 1, 2 and 3a, participants were more likely to choose the Test-Post items as more similar to the Target than the Test-Pre items (M = 80.89%), above chance, one-sample t (17) = 7.52, $p < .001$.

However, when this control condition is compared to the original visual domain experiment, it appears that the absence of temporal information attenuates this effect. There is a higher propensity to choose the Test-Post items as more similar to the Target than the Test-Pre items when there is the addition of temporal information, independent-samples t (33) = 2.07, $p < .05$.

General Discussion

The results of the three reported experiments clearly indicate that across the three domains, the beginning of the sequence was more salient than the end of the sequence, and as a result, the addition of a single element to the beginning of the sequence was perceived as a greater change than the addition of a single element to the end of the sequence. Presence of this tendency across the three domains indicates that this tendency is not limited to language. More specifically, the "suffixation preference" found in the linguistic domain appears to be analogues for sequences of musical tones and visual patterns, all having a temporal component.

Results of Experiment 3b indicate that the temporal component is fundamental: once the temporal component is removed and stimuli are presented simultaneously, the effect is diminished.

Therefore, the suffixation preference, which is often considered a useful linguistic bias for solving the inflectional problem, does not appear to be specific to language, but rather it stems from processing of temporally organized information.

To better understand this phenomenon, it is important to investigate this bias in native speakers of languages that do not have the same dominant suffixing preference (e.g., Thai). This may further reveal the direction and strength of this tendency.

In addition, there are several important questions that are to be answered in future research. First, it is unclear whether this tendency to consider the beginning of a sequence as more salient than the end of a sequence appears as a domain-general attentional bias or whether it first manifests itself in the domain of language, and then gets extended to other temporally organized domains. Although the former possibility seems more likely, a developmental study using the same set of stimuli is necessary to answer this question.

Given the fact that the effect exists in the visual domain even without temporal information, it is important to investigate possible explanations. Since the visual images without temporal information were presented in a manner that resembles the structure of written material, it is possible that this structure brought about the positional biases similar to those in the linguistic domain. For example, while reading English, one visually scans from left to right; therefore, this same mechanism could account for the effect in Experiment 3b even without temporal information. To better understand the phenomenon, it is necessary to structure the visual information so that a left to right scanning pattern does not temporally constrain the information (e.g., vertical presentation of the stimuli). In addition, this explanation may be further expanded if the effects would vary not only according to the presentation of the stimuli, but also developmentally. This explanation could be ruled out if children who do not yet read readily show this effect in the visual domain without temporal information.

Another important question is how flexible is this tendency in non-linguistic domains. The suffixation preference is very flexible in the domain of language

(otherwise people would not be able to acquire various kinds of inflectional morphology), and if it stems from a general mechanism, this tendency has to exhibit flexibility in other domains as well. These issues are currently under investigation.

In sum, the results suggest that when information is added to the end of a Target sequence, it is perceived as more similar to the original Target than if the same information was added to the beginning of this Target. This was true across all three domains investigated, suggesting that there might be a general cognitive mechanism of processing of temporal information that may underlie the suffixation preference, which is prominently present in the linguistic domain.

Acknowledgments

We would like to thank Zach Schendel for the creation of the musical stimuli. This research is supported by grants from the National Science Foundation (REC #0 208103 and BCS # 0078945) to Vladimir M. Sloutsky.

References

Clark, E. V. (1991). Acquisitional principles in lexical development. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development*. Cambridge, MA: Cambridge University Press.

Clark, E. V. (1995). Language acquisition: The lexicon and syntax. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition, 2nd ed.: Speech, language, and communication*. New York, NY: Academic Press.

Clark, E. V. (1998). Morphology in language acquisition. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology*. Malden, MA: Blackwell.

Gasser, M. (1994). Acquiring receptive morphology: A connectionist model. *Annual Meeting of the Association for Computational Linguistics*, 32, 279-286.

Hawkins, J. A., & Cutler, A. (1988). Psycholinguistic factors in morphologicalasymmetry. In J. A. Hawkins (Ed.), *Explaining language universals*. New York, NY: Basil Blackwell.

Hawkins, J. A., & Gilligan, G. (1988). Prefixing and suffixing universals in relation to basic word order. *Lingua*, 74, 219-259.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory & Language*, *44*, 548-567.

Slobin, D. I. (1985). Crosslinguistic evidence for the language-making capacity. In D. I. Slobin (Ed.), *The Crosslinguistic study of language acquisition, Vol. 1: The data; Vol. 2: Theoretical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Tyler, L. K., Marslen-Wilson, W. D., Rentoul, J., & Hanney, P. (1992). Continuous access processes in spoken word recognition. *Journal of Memory and Language*, *27*, 368-381.

How Copying Artwork Affects Students' Artistic Creativity

Kentaro Ishibashi (k.ishibashi@cc.nagoya-u.ac.jp)
Graduate School of Education and Human Development,
Nagoya University, Nagoya 464-8601, Japan

Takeshi Okada (j46006a@cc.nagoya-u.ac.jp)
Graduate School of Education and Human Development and
Institute for Advanced Research,
Nagoya University, Nagoya 464-8601, Japan

Abstract

30 undergraduates participated individually in a three-day-drawing experiment. It was explored whether an experience copying others' drawing facilitated subjects' artistic creativity. Results showed that drawings by subjects who previously had copied others' drawings were rated more creative than the drawings of subjects who had not copied. Two further analyses revealed how subjects could produce creative drawings. First, in the examination of constraint relaxation processes, subjects were initially constrained by a belief that they should draw things realistically. Then, they relaxed this constraint by means of copying abstract pictures. Second, according to protocols of the copying process, copying forced subjects to explore their original expression through a comparison with other artwork. It seemed that copying enabled them to generate new drawing ideas.

Introduction

It is often said that people cannot produce original works through the imitation of others. In the domain of art, many art educators believe that copying others' work inhibits people's, particularly children's, artistic creativity. They claim that artistic expression should be as free as possible from copying (Lowenfeld, 1957). It is well known, however, that artists of impressionism created their original paintings by means of imitating Japanese prints, *Ukiyoe*. In addition, some famous painters, e.g., van Gogh and Picasso, created their original paintings through copying the work of old masters (Galassi, 1996; Homburg, 1996). The question of whether copying inhibits or facilitates creative art has been controversial among artists, art researchers, and art educators (Duncum, 1988).

In some modern cultures, including Japanese culture, many art lay people (i.e., nonartists) seem to think that representational and realistic paintings have higher value than other forms of painting. That may be due to the content and methods of art education in school settings. Especially in Japanese elementary and middle schools, students spend the majority of their time in art class sketching. This may lead them to believe that drawing is primarily to represent objects in the real world on paper (Kozawa, 2001). In other cultures, it is also reported that people prefer realistic paintings to abstract or other types (Cupchik & Gebotys, 1988; O'Hare, 1976). Such beliefs might limit the range of students' means of expression. It is predicted that subjects would create new drawing styles if their constraints become relaxed. Therefore, we focused on copying others' work as a candidate for an intervention that could relax constraints and investigated its effect on creative drawing.

Method

Subjects. 30 undergraduates participated in this study. None of them had special training in drawing since at least middle school.

Experimental Design. A three-day-experiment (pre-treatment-post design) was conducted. All of the subjects were initially required to create two original drawings in the pretest phase. In the treatment and posttest phases, subjects were divided into three groups. In the Experimental Group (EG), subjects were asked to copy two pieces of an artist's drawings, then to create their own original drawing. In the Reproduction Group (RG), subjects were also asked to copy, then to draw a new picture using the artist's style. In the Control Group (CG), subjects were asked to draw their own original drawings in every session.

Materials. Subjects were required to draw pictures using as subject matter the materials displayed in Table 1. A4-sized Kent paper and a black ballpoint pen were offered to subjects for each drawing. The pictures copied by subjects in the two groups were abstract paintings by a Japanese modern artist (Figure 1).

Procedure. Each subject participated individually in a three-day-experiment; one session per day, each lasting approximately 90 minutes. Subjects were asked to draw two pictures in each of the pretest and treatment phases. The second picture in each phase was presented three minutes after the first one was completed. In the posttest, subjects drew a picture and then were asked to

Table 1: Materials presented to subjects.

	Experimental Phases		
	Pretest	Treatment	Posttest
1st drawing	a cocktail glass	a shell (Venus Comb Murex)*	an orange and a shell (Common Spider Conch)**
2nd drawing	a paprika and a pinecone**	a potted plant*	-

* For EG and RG, the pictures to copy were drawn with each of these materials by an artist and presented alongside the materials.
** These sets of materials were counterbalanced among subjects.

complete a questionnaire (described later in detail) and were interviewed about their drawings. Thus, each subject drew five pictures in total during three days.

Subjects in CG were instructed as follows in all phases: "Draw your own ORIGINAL picture using this (these) material(s) as subject matter." Subjects in EG were instructed in the same way in the pre and posttest phases. But, they were told in the treatment phase: "A painter drew this picture using this material as subject matter. Please copy the picture onto a blank piece of paper while imagining the painter's intention." Subjects in RG were instructed in the same way as EG in the pretest and treatment phases. However, they were told in the posttest phase: "Recall the previous day's experience of copying a painter's picture and then draw a picture with these materials in the painter's style. How would you represent the subject matter if you were the painter?"

We asked subjects to talk aloud while drawing, and recorded their verbal protocols and behavior with three videocassette recorders. Except for this procedure, we placed upon the subjects' activities as few restrictions as possible in order to promote maximum spontaneity (e.g., They were not told explicitly that there was a time limit on their drawing).

Results and Discussion

Preliminary Analysis

In the posttest phase, drawings in RG were quite different from those in EG and CG in terms of content and number of elements in each picture (Figure 2). The mean number of elements in a drawing was significantly greater in RG (23.8) than in EG and CG (10.2 and 5.4, respectively) [for group by phase interaction, $F(2,27)=5.05$, $p<.05$]. All drawings by subjects in RG consisted of much repetition of simple geometrical elements, but those by subjects in EG had no such characteristics. Thus, although subjects in EG

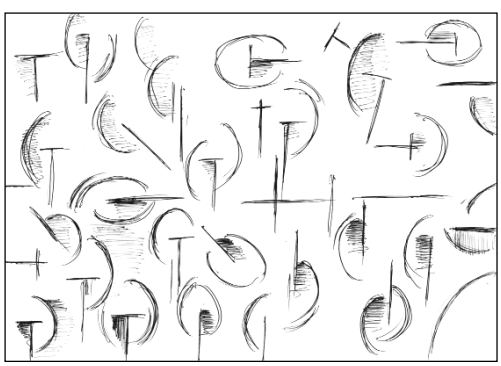


Figure 1: Example of the artist's drawings that subjects in the EG and RG saw.

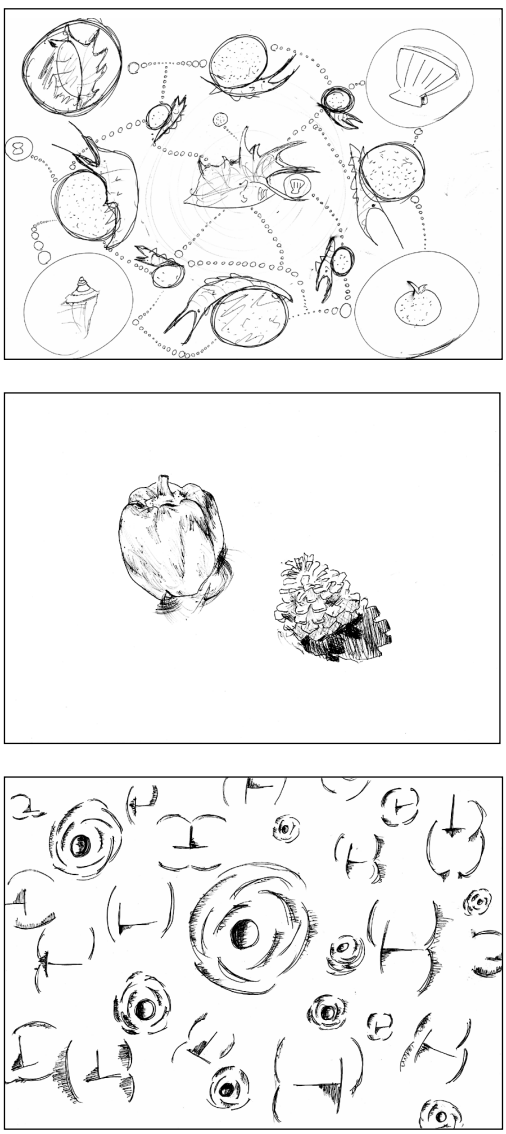


Figure 2: Examples of drawings in the posttest phase (EG; CG; RG, respectively from the top).

and RG copied pictures in the same manner, subjects in EG did not reproduce the artist's style of pictures, but created their own styles.

Analysis of Products: Rating Creativity of the Drawings

In order to compare the creativity of drawings in EG with that in CG, a new scale was constructed that included three aspects of artistic creativity: six items of aesthetic attractiveness (e.g., "vitality of expression"); nine items of originality (e.g., "originality of her or his view or sense of value"); and two items of technical skills (e.g., "technical skill in picture composition"). Thus, in total, 17 items were included in the scale with all items ranging from 1 to 5. Because our critical question was to reveal whether or not the artistic creativity of subjects who copied others' art works was superior to that of subjects who did not copy, the comparison of the two groups would be sufficient to answer the question. Thus, we excluded the drawings by subjects in RG from this analysis.

Two professional modern artists separately rated subjects' pre and posttest drawings using the scale. They were not informed of which drawings belonged to which condition. A result of factor analysis with Principal Component Analysis showed that the scale has one factor construction and was adequate for the evaluating creativity of drawings (eigenvalues for the first three factor were 10.58, 2.06, and 1.05, and the first factor accounted for 62.2% of the total variance). Since the coefficient alpha for internal coherence was .96 for all 17 items, we regarded the simple sum of the 17 items as the creativity score for each drawing.

A three-way ANOVA (two experimental groups X two raters X two expositional ordering of drawing materials) was performed for post-pre subtracted scores.

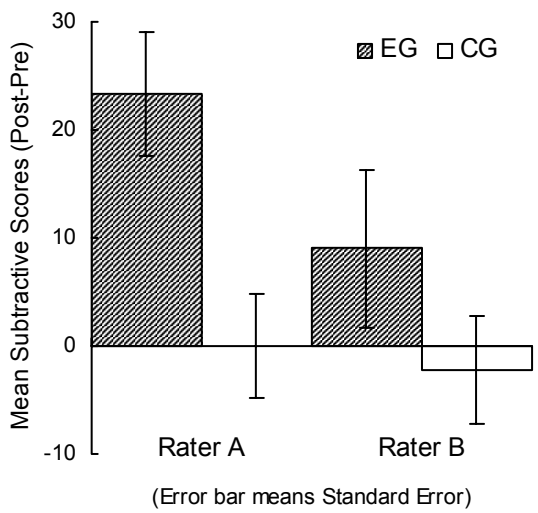


Figure 3: Comparison of creativity rating between subjects in EG and in CG

Drawings by subjects in EG were rated significantly higher than those in CG (Figure 3) [$F(1,16)=5.54$, $p<.05$]. The fact that scores were significantly different between the two raters suggested that norms of artistic creativity would vary among artists [$F(1,16)=4.65$, $p<.05$]. However, it was important to note that the two raters evaluated the posttest drawings by EG subjects in the same way. There was no interaction among the three factors. Findings suggest that copying other's drawings provided the subjects opportunities for creating new styles of drawing.

Analysis of Process 1: Relaxation of Students' Constraints

Why could subjects who had copied other's works produce more creative drawings? Note that the pictures the subjects had to copy (abstract style) were fairly different from typical pictures that subjects normally encounter (representational style). If subjects were constrained by their beliefs that drawing must follow a representational expression style, copying drawings in an abstract style might relax the constraint by making them aware of other stylistic possibilities.

In order to conduct further analyses of the process of creation, we focused on the following three aspects: (1) number of pictures that included realistic contents; (2) strength of the subjects' realistic intention; (3) number of subjects who reported a failure of creative drawing.

Number of Pictures that Included Realistic Contents. If subjects were constrained by their beliefs that drawing had to be realistic, such beliefs would affect the content of their drawings. In this study, we coded a drawing as constrained by such beliefs if it contained at least one of the following aspects: (1) drawings that designate a specific scene made up of either realistic elements or stylized ones (e.g., one similar to an illustration of a storybook); and (2) drawings in which subjects sketched only the materials presented (example of drawing from CG condition in Figure 2). We took these two types of drawings to indicate that subjects drew without their own figurative interpretations.

The numbers of drawing which contained the contents described above were approximately equal in the pretest phase in both conditions (70% in EG and 80% in CG) [$p=1.00$ with Fisher's exact test]. The frequency of that in EG, however, significantly decreased compared to that in CG in the posttest phase (20% and 90%, respectively) [$z=-2.21$, $p<.05$ with test by standardized scores].

Strength of Subjects' Realistic Intention. Did subjects actually intend to draw pictures so realistically? In order to capture their intention, we investigated how much they paid attention to technical viewpoints related to realistic sketch-like drawing. We assumed that the more subjects thought they had to draw realistic and photo-like pictures, the more strongly

they would attend to such technical viewpoints. If copying pictures in an abstract style relaxes such a constraint, then the degree of EG subjects' attention to such technical aspects would decrease in the posttest.

After posttest drawing, subjects were asked to answer a questionnaire intended to measure their realistic intention during both the post and pretest drawing. This questionnaire consisted of 11 items on five-point scale (ranged from 1 to 5, including two inverse items) that covered a variety of aspects of realistic intention (e.g., "I paid attention to capturing the materials' form exactly"; "I tried to express the quality of the materials' surface"). For the 11 items, the coefficient alpha was .86.

A two-way ANOVA (three experimental groups X two phases) for the sum of item scores revealed a significant interaction [$F(2,27)=9.82$, $p<.001$]. Further analysis revealed that, while there was no significant difference on groups in pretest phase [$F(1,54)=0.01$, *n.s.*], in posttest phase, scores in EG and RG significantly decreased compared to scores in CG (Figure 4) [$F(1,54)=7.93$, $p<.001$; $p<.05$ with Steel's multiple comparison for difference scores of post-pre test]. In addition, subjects' scores in the pretest phase were on average about 70 % or more of the maximum score and thus seemed to show their strict intention to use a realistic drawing style. Hence, we can conclude that the subjects did, in fact, have representational constraints in the beginning of the study and that the constraints were then relaxed by means of copying pictures with an abstract style.

Number of Subjects Who Reported a Failure of Creative Drawing. We asked subjects to report what they devised in the posttest drawings. Their answers were divided into categories and the contents and the number of responses were analyzed.

Characteristically, half of the subjects in CG reported

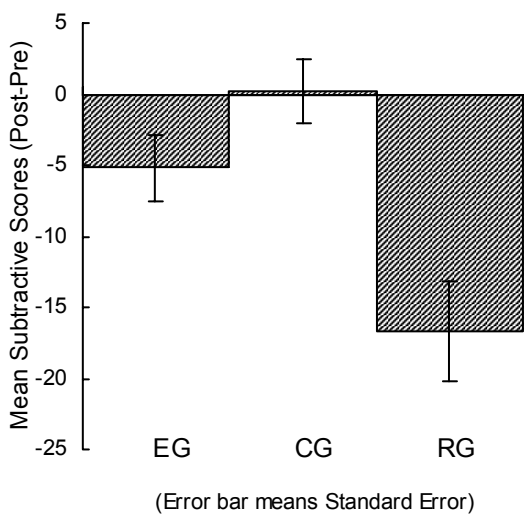


Figure 4: Subjects' intention to draw realistically.

that they could not come up with any new ideas and just sketched what they saw (e.g., "I thought that I could not draw well if I pay too much attention to originality. So, I decided to draw the materials as they are"). In EG, however, no subject reported such a comment [$p<.05$ with Fisher's exact test]. This result implies that subjects in CG were kept constrained by their beliefs and could not produce new ways of drawing.

Analysis of Process 2: Generation of New Ideas

The previous section revealed that copying relaxes subjects' constraint. However, even if their constraints are relaxed, it is insufficient for production of a new style of drawing. Because, in order to create a new style of drawing, subjects need to generate concrete ideas for drawing. In order to reveal how subjects in EG came up with new ideas when their constraints were relaxed, we focused on subjects' copying process in the treatment phase.

We presumed that thought processes during copying include two aspects: (1) understanding others (in this case, a creator who produced the artwork) and (2) understanding oneself. The former aspect is an effective one in order to reproduce others' artworks. The knowledge about the pictures would be deepened by means of inferring the creator's art making process. In this point of view, however, copying can be risky since people may lose their own originality. Thus, many people have claimed that copying might be harmful to creation. As we pointed out, this is a well known argument.

In contrast, the second aspect, understanding oneself, is not so well known. In this aspect of thought processes, the copiers' own expression may become clarified by means of comparisons with others' artworks. Thus, people's generation of new ideas might be facilitated through their searching for originality. This aspect may be particularly important for creativity, because it might promote the copier's ability to produce her/his own original artworks.

It is hypothesized that subjects in EG experienced these two aspects of thought processes when copying and were able to generate new ideas to draw. In the rest of this section, we will focus on the protocols by EG and RG subjects during copying an artist's artworks and describe whether or not the protocols include evidence of these two aspects. Of course, these aspects are double-faced, and one cannot work without the other. In this study, however, we will pragmatically separate them into two aspects and examine each.

Copying to Understand Others. In this aspect, getting to know the processes by copying could deepen knowledge of the products. For example, F. Natsume (cartoon artist; 1992) copied a famous Japanese cartoonist's work. He found that the lines of this cartoon give a very round and centripetal impression. This characteristic of the lines has an important role in

creating this cartoonist's characters' special features such as bravery and cuteness. The case shows that he deepened his knowledge about the cartoon through copying its lines.

In addition, understanding others' works requires changing one's standpoint. In order to really copy, it is necessary for us to understand the other's underlying intention of the procedure. When copying, we are forced to infer the underlying intention of the other's works. This process makes us switch our standpoint from an observer to a creator.

Thus, copying facilitates understanding a creator. Did actual copying processes include this aspect? We focused on subjects' protocols during the copying phase (treatment phase in EG and RG) and analyzed whether or not their protocols exemplify understanding others.

In copying, subjects noticed concrete features of elements/parts in the artist's drawings.

- *Why did he draw this horizontal line?* (subject =ID3)
- *He doesn't draw outlines, does he?* (ID3)
- *I copy it paying attention to the distance with other parts.* (ID14)
- *I must use stronger lines. My lines were not clear at all.* (ID19)

Subjects also tried to understand the artist's intention.

- *I think, the thick parts of the leaves indicate this plant's vitality.* (ID16)
- *Each element in the picture may not represent each leaf of the real plant.* (ID7)

In this way, copying process did have an aspect of understanding others. Subjects in RG could reproduce the new picture in the artist's style because they would engage such a process and deepen their knowledge about the artist.

Copying to Understand Oneself. It seems that understanding others also facilitates understanding oneself. Consider the following case. Even if you had no opinion about an issue at first, you may often form your own opinion while listening to others'. In the domain of art, there would exist such a case that the deeper you understand someone's artworks, the more you become aware of your originality.

There are many such examples in art history. For example, Picasso and van Gogh copied old masters' artworks, exploring their own original style rather than keeping the styles of the artworks exactly (Galassi, 1996; Homburg, 1996). Picasso talked about his copying (Sabartés, 1959):

> *Suppose one were to make a copy of The Maids of Honor (Las Meninas); if it were I, the moment would come when I would say to myself: suppose I moved this figure a little to the right or a little to the left? At that point I would try it without giving a thought to Velázquez. Almost certainly, I would be tempted to modify the light or to arrange it differently in view of the changed position of the figure. Gradually I would create a painting of The Maids of Honor sure to horrify the specialist in the copying old masters. It would not be The Maids of Honor he saw when he looked at Velázquez's picture; it would be my Maids of Honor.*

This case shows that Picasso actively explored his own expression through copying Velázquez's work. It is a different aspect from the one that focuses on learning particular techniques or expressions (i.e. understanding other's works).

Why does copying facilitate self understanding? We propose the following two reasons. At first, when copying other's works, you constantly compare other's expression with your own. This "comparison" process forces you to actively interpret the differences between the other's works and your own. This is the first step in searching for your own original expressions. Secondly, particularly in copying artworks, you can externally compare a model with your copy. This encourages you to notice differences between the two.

Our protocol data show subjects' self understanding. First, subjects' own visions emerged. They interpreted the figures in their own way as well as inferred the artist's intention.

- *It looks like fossil fishes are swimming.* (ID7)
- *It looks like insects are flying.* (ID12)
- *They look like ribs or fish bones.* (ID3)

Some subjects felt uncomfortable with the other's works. It seemed that such feelings prompted them to explore their comfortable expressions.

- *Why did he/she draw such cross marks? I cannot find them in this material* [= a shell]. *I don't understand it.* (ID19)
- *I don't like patterned figures like this. Because it's monotonous.* (ID1)

Some subjects became aware of their own expression by means of the comparisons with other's work.

- *His lines end smoothly, but mine stopped tightly.* (ID1)
- *In the previous copy, I failed to draw pictures well, because I drew the elements too big and lost balance.* (ID14)

- *This picture reminds me of previous day's paintings* [Subject's own drawings in the last session]. *I now understand that mine were not so original.* (ID1)

These findings enable subjects to understand what kinds of expressions they usually use and what kinds of expressions they want to create.

General Discussion

This study revealed that an experience copying others' drawings facilitated subjects' artistic creativity. It was also showed that at least two underlying processes affected this performance. First, constraint relaxation processes enabled subjects in the EG to explore drawing styles beyond the familiar realistic and representative style. Second, generating new ideas through comparison with other's works prompted subjects to notice their own original expression. Based upon these findings, we propose a model about copying to creation (Figure 5). It is suggested that constraint relaxation and generation of new ideas (including two aspects of copying) together can facilitate a new style of drawing.

Some recent studies investigated the effect of experimenter-presented examples on a creative generation task. For instance, Smith, Ward, & Schumacher (1993) found that people unconsciously tend to incorporate features of the examples in their creation (conformity effect). This effect varied with conditions; for example, the effect was enhanced with a delay between exemplar presentation and creation test (Marsh, Landau, & Hicks, 1996). In addition, it was related to inadvertent plagiarism because people fail to monitor their source of novel knowledge appropriately. These studies suggest that examples may negatively affect creation. However, it is well known that no idea is completely original; all forms of creation are strongly affected by already existing things. Thus, the question that we want to answer here is how people create new ideas even if they have a tendency to be heavily influenced by old ideas, as previous studies suggest. Although this research is still in an early stage, we propose that the process of understanding oneself in comparison with others works is a key mechanism of creation.

One reason why previous studies did not focus on this aspect is perhaps that the subjects in these investigations saw exemplars for only a few minutes and thus did not have enough time to involve themselves in the process of understanding. In contrast, in our experiment, the subjects spent about forty minutes copying pictures. This long, active exposure to examples may have forced them to engage in the process of understanding themselves.

Despite this, copying others' works may not be the only means of making people more creative. If they were just told verbally to consider other forms of drawing or presented others' works as exemplars without copying, then they might also be able to draw more creatively. We are currently conducting another experiment to test these possibilities.

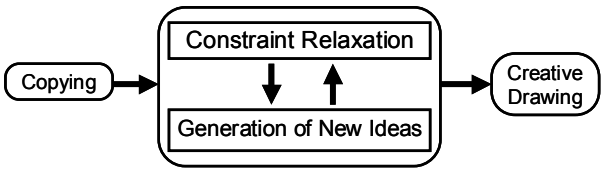


Figure 5: Interactive processes in copying.

Acknowledgments

This study was supported by a *Grant-in-Aid of Scientific Research* #C-14510135 from the Japan Society for the Promotion of Science to the second author.

Since the experiment was carried out over summer vacation, we indeed want to show gratitude to all subjects for their participation. We also want to acknowledge three modern artists for their generous support and suggestions.

References

Cupchik, G. C. & Gebotys, R. J. (1988). The search for meaning in art: Interpretive styles and judgments of quality. *Visual Arts Research*, *14*, 38-50.

Duncum, P. (1988). To copy or not to copy: A review. *Studies in Art Education*, *29*, 203-210.

Galassi, S. G. (1996). *Picasso's variations on the masters*. New York: H. N. Abrams.

Homburg, C. (1996). *The copy turns original: Vincent van Gogh and a new approach to traditional art practice*. Amsterdam; Philadelphia: John Benjamins.

Kozawa, M. (2001). *Kaiga no seisaku: Jiko hakken no tabi*. [Art making: The journey to discover oneself.] Tokyo: Kaden-sha.

Lowenfeld, V. (1957). *Creative and mental growth* (3rd ed.). New York: Macmillan.

Marsh, R. L., Landau, J. D., & Hicks, J. L. (1996). How examples may (and may not) constrain creativity. *Memory & Cognition*, *24*, 669-680.

Natsume, F. (1992). *Natsume Fusanosuke no Manga-gaku*. [A Theory of Cartoon by Fusanosuke Natsume.] Tokyo: Chikuma-shobo.

O'Hare, D. (1976). Individual differences in perceived similarity and preference for visual art: A multi dimensional scaling analysis. *Perception and Psychophisics*, *20*, 445-452.

Sabartés, J. (1959). *Picasso: variations on Velázquez' painting "The maids of honor"*. New York: H. N. Abrams.

Smith, S. M., Ward, T. B., & Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition*, *21*, 837-845.

Sensorimotor Contingencies, Event Codes, and Perceptual Symbols

Jason Jameson (j-jameson@northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Road
Evanston, IL 60208-2710 USA

Abstract

Cognitivism, the traditional approach to understanding cognition, has argued for the essential role of symbolic computations over internal mental representations. But this view has been criticized on a number of grounds, one in particular being the assumption of amodality: that the symbols involved in processing are arbitrarily related to their referents. An opposing view—the framework of Perceptual Symbol Systems—holds that the elements of thought should be treated not as amodal symbols, but rather as modality specific, analog representations that simulate particular aspects of perceptual experience. Though this approach has been gaining in popularity from intuitively appealing theoretical accounts, and suggestive empirical support, it has suffered from a lack of specificity for key constructs. To address this problem, this paper presents a more detailed study of the foundational concept of *perceptual symbol*. The proposal builds from recent work on the skill-based nature of visual perception (the Sensorimotor Contingency Theory), and research that provides tools for representing the inseparable link between perception and action (the Theory of Event Coding). From these two sources, the characterization of a perceptual symbol as a selective re-enactment of perceptual experience, treated as a unit, will be elaborated and defended.

Introduction

Cognitive science, for much of its short history, has been dominated by a view of cognition that emphasizes the necessary role of computation, and which holds that cognitive processing is rule-governed manipulation of internal mental representations (Fodor, 1975, 1983; Fodor & Pylyshyn, 1988; Johnson-Laird, 1989; Minsky, 1975; Newell & Simon, 1972; Pinker, 1998; Pylyshyn, 1984). The symbols that comprise these representations are what codify knowledge, and indeed *are* knowledge. Under this interpretation, symbols possess several key properties, the most important of which, for the purposes of this paper, is amodality: that a symbol is arbitrarily related to the thing it represents (Barsalou, 1999; Markman & Dietrich, 2000). In addition, specific psychological theories that adopt this framework "generally assume that knowledge resides in a modular semantic system separate from episodic memory and modality-specific systems for perception, action, and emotion" (Barsalou, et al., 2003). This view of cognition has undoubtedly met with much success (for accessible overviews, see Johnson-Laird (1989) and Pinker (1998)).

There are critics, however, who have challenged this framework (Barsalou, 1999; Carlson, 1997; Clancey, 1997; Clark, 1997; Damasio, 1994; Dourish, 2001; Dreyfus, 1972; Gibson, 1979; Glenberg, 1997; Harnad, 1990; Hutchins, 1995; Lave & Wenger, 1991; Rumelhart & McClelland, 1986; Searle, 1980; Smith & Thelen, 2003; Suchman, 1987; Thelen, 1994; Thelen, Schoner, Scheier, & Smith, 2001; Thomas, 1999; Van Gelder, 1998; Varela, Thompson, & Rosch, 1992). In particular, one approach questions the requirement that the symbols used in cognitive processing should be amodal. Instead, in a Perceptual Symbol System (PSS), the symbols are modality-specific representations that do bear a principled resemblance to the things represented (Barsalou, 1999; Barsalou, et al., 2003). Specifically, these symbols are perceptual in the sense that they re-enact selective aspects of experience. But this view, though promising, remains underspecified in important ways. This paper is an attempt to clarify a fundamental construct of the PSS approach, the *perceptual symbol.*

The specific goals of this paper are: (1) to present an account of conceptual representation that is at odds with the traditional view in one important respect—that the symbols used in thought are amodal; (2) to review theoretical arguments and empirical evidence that suggest that PSS should be taken seriously as a plausible alternative to Cognitivism; (3) to show that there are certain respects, however, in which the PSS framework is underspecified, specifically with respect to the foundational concept of a *perceptual symbol;* (4) based on the assumption that to clarify the concept of *perceptual symbol* requires some understanding of what perceptual experience is, to present one type of skill theory of perception that provides a comprehensive account of how perception *and* action interact to support perceptual experience; (5) to connect this account to perceptual symbols by adopting representational structures called *event codes* that possess key properties required by a PSS; (6) to review consistent empirical evidence that the properties of event codes that hold at the fine level of basic sensorimotor interaction might also hold during higher level cognitive processing; and (7) to suggest limitations and remaining questions for future study.

Why a Perceptual Symbol System?

The framework of Perceptual Symbol Systems (PSS) is a perceptually-based approach to conceptual representation that has gained in popularity for many theoretical and empirical reasons. On the theoretical side, the view is more sophisticated than its empiricist predecessors. First, it appears that the rejection of perceptually oriented approaches was too hasty (Barsalou, 1999). For example, the criticism that perceptual symbols are just holistic records of perceptual experience (like internal pictures that lack any interpretation) is based on the assumption that perceptual

symbols could not also be treated as discrete, componential units; and in this regard, be used productively in forming multimodal symbolic structures. But under this new construal, perceptually-based conceptual systems can acquire the power to represent a variety of concepts ranging in abstractness, and thereby possessing the desired flexibility shown by humans (Barsalou, et al., 2003).

Second, because of the analog relationship with their referents, perceptual symbols provide a great deal of implicit information about the things they represent (Zwaan, 1999). This information can then be made explicit through perceptual processes like scanning and selective attention (Goldstone & Barsalou, 1998). An amodal system requires that knowledge be expressed in terms of syntactically well-formed sentences, usually expressed in a first-order predicate calculus, or LISP-type language. This requirement places a heavy burden on a system if it must represent *all* knowledge explicitly (even the most mundane kind—for example, that cars have four wheels).

Third, the foundations of amodal symbol systems are not without problems, specifically in terms of how the symbols are acquired—the transduction problem (Barsalou, 1999)—and how the symbols relate back to the world—the symbol grounding problem (Harnad, 1990). If perception and cognition are realized by fundamentally different cognitive processes, then how does one representational language get *transduced*—that is, translated—into the other? Conversely, how does the output of cognitive processing connect with the world to enable purposeful interaction? In other words, how do the symbols become *grounded*?

Fourth, though presumed to be flexible by virtue of the (amodal) form of the symbols (such as using the symbol CAT to stand for all cats), amodal symbol systems lack the flexibility of human cognition. To get around this problem, amodal representations have been supplemented with specific episode information (Markman & Dietrich, 2000).

Fifth, amodal conceptual systems are well known to be able to account for numerous findings after the fact (Anderson, 1978; Solomon, 2001), but of far greater importance is the power of a theory to make *a priori* predictions. The amodal view cannot easily predict perceptually-motivated effects, whereas a perceptually-based view can do so with ease (Barsalou, et al., 2003).

In addition to the theoretical support, a growing body of empirical research suggests a strong influence of perceptually-based knowledge on conceptual processing. For example, in a property-listing task, people will generate response that depend on the nature of the perceptual variables involved in the simulation (e.g., listing "roots" for "rolled-up lawn", rather than for "lawn", because roots are less occluded in a rolled-up lawn)(Barsalou, Solomon, & Wu, 1999). From studies in text comprehension, in an effort to comprehend a text passage, people appear to construct online simulations of the situations described in the text. In other words, comprehension of language becomes "preparation for situated action" (Richardson & Spivey, 2000). The situation models (Zwaan, 1999) that underlie this comprehension process are fundamentally experiential, and not surprisingly, derive much motivation from the PSS framework. In research motivated by these ideas, people have been found to recognize a picture of an object (for example, a nail) more quickly if the object is in the same orientation, vertical or horizontal, implied by a text passage read earlier ("The nail was hammered into the floor/wall")(Stanfield & Zwaan, 2001). In related work, sentences such as "Open the drawer" are judged as sensible more quickly if at the same time people move in a manner consist with the implied motion (in this case, pulling rather than pushing)(Glenberg & Robertson, 2000). People also appear to re-enact the eye movements that accompanied earlier perceptual processing (Laeng & Teodorescu, 2002; Mast & Kosslyn, 2002; Richardson & Spivey, 2000). These results are just a few of the many that have provided support for the PSS framework (for a more detailed review, see Barsalou, et al., 2003)

The Structure of a Perceptual Symbol System

A PSS is a conceptual system composed of an integrated set of simulators (which in practice can be interpreted as the concepts). The simulators are composed of frames, which integrate perceptual symbols, and provide structure for event sequencing. Moreover, each simulator implies a simulation competence—the potential for producing an indefinite number, and limitless variety of perceptual simulations. Finally, processes of selective attention and memory integration provide the requisite representational power for the system to act as a fully functional conceptual system in the classical sense (Barsalou, et al., 2003; Fodor & Pylyshyn, 1988).

For a PSS to function as a conceptual system, it should possess certain properties. (1) The conceptual system should be able to *interpret* novel experience. This is what fundamentally distinguishes a conceptual system from a simple recording system (Barsalou, 1999). A conceptual system is selective and is able to bind tokens (perceived individuals) to knowledge of types stored in long-term memory. A record (e.g. a picture), on the other hand, is an undifferentiated—uninterpreted—mass. (2) A conceptual system should allow the thinker to go beyond the information given, to use stored knowledge to make inferences. (3) Conceptual systems should have the potential for generating an indefinite number of thoughts; that is, they should be productive.

For a conceptual system to do this, it must be composed of things that have special properties. What exactly these properties are is contested, but Markman and Dietrich (2000) have provided an illuminating analysis of the issue, and their general approach will be adopted here. Specifically, they have argued that *internal mediating states* can possess certain characteristics: (1) they may be enduring; (2) discrete; (3) abstract (amodal); (4) rule-governed; or (5) they may possess a compositional structure (for more detail on these properties, see Markman & Dietrich, 2000). Internal mediating states in the cognitivist

tradition hold all five. Those in the dynamical systems approach hold fewer. Relaxing one or more of these constraints can affect the representational capacity of a conceptual system.

There are two important ingredients that support the interpretive capabilities of a simulator, and as a result, the representational power of a PSS: the frames, which integrate and organize perceptual symbols; and the "potentially infinite set of simulations that can be constructed from the frame" (Barsalou, 1999). Thus, to understand what CAR means is to know, not only the perceptual symbols that comprise the representation for cars, but it is to know also how to interact with cars—to be able to organize the complex action sequences involved in effective interaction with cars. This means that the frames would be composed of perceptual symbols from several different modalities, and that a simulation could thus be considered a multimodal, selective re-enactment of perceptual experience

The importance of simulators and simulations for supporting the conceptual functions of the PSS cannot be overstated. However, given the role of perceptual symbols in supporting simulations, much work remains to be done in specifying their properties (such as how the symbols are encoded, stored, and used). There are difficulties, however, in getting a clear sense of what a perceptual symbol is, and how it fits in with the functional architecture of a PSS.

Most definitions of a perceptual symbol tend to emphasize the neural substrate, and specifically, that perceptual symbols are "records of the neural states that underlie perception" (Barsalou, 1999). But a limitation of this approach is that it captures just one aspect of the information contained within perceptual symbols. Though much current research tries to incorporate properties of classical cognitive architectures into neural networks, what is needed is a more explicit account of the functional structure to complement the neural description. What is needed, then, is a better sense of what information goes into a perceptual symbol and how that information is stored so that it can support the conceptual functions of a Perceptual Symbol System. In other words, we need a theory of perception, and a theory of how the products of perception are represented.

Perception as a skill

The perspective in this paper holds that perception is a skill: that it is the ability to engage in purposeful and effective interaction with the world (Ballard, 1983; Clark, 2002). The approach outlined is just one of many types of "skill" theories of perception, but it is one of the most elegant and best developed. Specifically, this Sensorimotor Contingency Theory holds that to perceive is to engage in skilled exploration of an environment, with the exploration mediated by the implicit knowledge of the lawful dependencies that hold between actions and sensory consequences (O'Regan & Noë, 2001). Thus, these lawful dependencies are assumed to play an essential role in providing content for perceptual symbols.

The SCT

The main goal of the Sensorimotor Contingency Theory (SCT) is to provide an answer to the so-called "hard problem" of visual consciousness: to explain how physical or informational processes could give rise to the qualitative character of experience (Chalmers, 1996). The solution to the problem is framed in terms of an interpretation of visual perception as a "mode of exploratory activity that is mediated by knowledge of sensorimotor contingencies" (O'Regan & Noë, 2001). This idea is in opposition to views, such as Muller's Doctrine of Specific Nerve Energies, in which what makes one sensory modality different from another is due to the nerve pathways that gather information. Rather, what makes modalities differ is that each is supported by different sets of sensorimotor laws: the dependencies between motor outputs, and the sensory consequences of those actions. In other words, the laws are the implicit, *procedural* knowledge of the expectancies derived from an agent's interaction with an environment. For more details, and supporting evidence, see O'Regan & Noë (2001),

According to this view, sensorimotor contingencies are a key ingredient in most, if not all aspects of cognition. As the authors describe it, "To see is to explore one's environment in a way that is mediated by one's mastery of sensorimotor contingencies *and* to be making use of this mastery in one's planning, reasoning, and speech behavior" (O'Regan & Noë, 2001). How to scale up to these behaviors remains to be seen, however. It is this role that the PSS should fill—to account for the emergence of abstract thought from this fundamental perception-action interface.

But there are a number of problems to be overcome in attempting to extend these principles to the PSS framework. The main limitation is that no indication is given for how sensorimotor contingencies should be represented, or even whether they should be represented at all. For reasons given in Markman and Dietrich (2000), it is too early to abandon representation as an explanatory construct in theories of cognition. So, to be able to characterize a perceptual symbol as a selective reenactment of perceptual experience, to be treated as a unit, not only must there be some sense of what perception consists in, there must also be a way to represent the information that supports both perception and cognition. In addition, the representation should possess the right properties to support the functional requirements of a conceptual system. The Theory of Event Coding (TEC), and specifically, the *event code,* is proposed to fill this role.

The TEC

The Theory of Event Coding (TEC) addresses the relationship between perception and action planning. In opposition to traditional approaches, the TEC does not assume independence between the two processes, but instead emphasizes that both functions are supported by a common representational medium. At the heart of the TEC is the notion of an *event code*, which "consists of the codes that represent the distal features of an event" (Hommel, et

al., 2001). These codes, then, are what underlie perception-action dependencies, but also, more generally, cognitive processes. In their words, "The theory holds that cognitive representations of events (i.e., of any to-be-perceived or to-be-generated incident in the distal environment) subserve not only representational functions (e.g., for perception, imagery, memory, reasoning) but action related functions as well (e.g., for action planning and initiation)" (Hommel, et al., 2001). The principle of common coding is a core assumption of the theory. The other—the principle of effect cause of actions—places special emphasis on the roles of specific types of feature codes within the event code; that is, the feature codes for actions are initiated by the resultant changes in sensory input caused by the actions.

As representational constructs, the event codes are discrete, compositional, and the individual elements retain an identity, even when they participate in more complex structures (Hommel, personal communication, Feb. 2004). In all, these considerations implicate an important functional role that event codes may play as representational constructs within the PSS.

Scaling up

Because the TEC has focused on fine sensorimotor interactions, such as "arrows or circles that come and go on a screen, or hand and fingers that go up and down on a key pad" (Hommel, et al., 2001), how are these event codes integrated to produce more complex event structures? Why start with the TEC rather than, say, with CHREST (Lane, Cheng, & Gobet, 2000) or PLAN (Chown, Kaplan, & Kortenkamp, 1995)? There are several reasons. First, the TEC is compatible with CHREST (as a model of active perception) and PLAN (a model of navigation that uses cognitive maps), and indeed they may all share a common logic (Hommel, et al., 2001). Second, the TEC was adopted primarily for practical reasons: It seemed to bear most closely on the issues addressed in the SCT, and would prove the shortest leap to make from non-representational to a representational description of sensorimotor dependencies. Third, it appeared that the best bet was to start small, at the principled "bottom", and then to work up. But associated with this is a riskier bet: that because of the recursive nature of the event codes, the integrative mechanisms operating at the lower-levels worked also at the higher levels. Fourth, the connection from the TEC to PSS had already been suggested by other researchers, specifically Richardson and Spivey (2001). Finally, some recent evidence in word learning suggests that the principles operating in the TEC might also hold at higher levels. Specifically, Terry Regier and his colleagues have suggested that more attention is paid to the *endpoint* than to the beginning of a spatial event (Regier & Zheng, 2003). Furthermore, as evidence of this, Regier and Zheng found that finer semantic distinctions are present in spatial terms at the endpoints of spatial events than at their beginnings. Specifically, in a task that required participants to judge whether two events, presented very briefly, were the same or different, fewer errors were committed in a "joining" task (e.g., where a lid would be put either on or in a container) that required attention be devoted to the endpoint, than in a "separation" task, in which a lid would be taken either off or out of a container, and attention would be required at the starting point. These results are consistent with the potentially important role of *goals,* and the principle of effect cause of actions, in encoding event sequences at a more general level.

Discussion

Much talk is made in this paper about perceptual symbols being "selective", suggesting that not all information from perception is used during cognitive processing. But what exactly does this mean? That the components of cognitive processing are less vivid, in the sense that *all* information carried over is less definite, less certain? Or does it mean that only specific kinds of information are carried over, such as spatial information, but that the information is no less definite, no less certain than when it was originally processed?

What are the laws of sensorimotor contingency that describe particular core aspects of experience, such as our experience of space, or time? Already much fascinating work has addressed the problem of doing this for space (Philipona, O'Regan, & Nadal, 2003), but then how might that knowledge map onto cognitive psychological research on space? And perhaps of even greater interest, how does our more abstract notion of time map onto that (Boroditsky & Ramscar, 2002; Gentner, Imai, & Boroditsky, 2002)?

What implications does this view have for the problem of reference—determining how cognitive structures connect with the external world (Evans, 1982)? On the one hand, it seems that there might be no such problem, since in both cases, the thing representing, and the thing represented are one and the same. Truly, the "external world" is itself an assortment of mental entities. In other words:

> Despite the importance of realism in many philosophical theories of concepts and meaning, this assumption seems superfluous and unempirical, and it introduces a number of additional problems to be dealt with that could be avoided without it. Rather than making a realist assumption, it would be easier to adopt a *coherence-based* framework. That is, the only information that any person has about the outside world comes from perceptual representations, which are themselves mental entities. Thus, rather than being concerned with whether a particular concept correctly refers to all and only proper extramental entities, it would be better to generate a theory in which the use of the concept attempts to remain consistent with other representations in the system. (Markman & Stillwell, in press)

But to someone still wary of the potential problems this (apparently) neo-idealistic position might suggest, the claim isn't so risky: we need not take on the assorted difficulties a Berkelyian idealism might, since we wouldn't be making claims about ontological status of the world (and whether it still would exist even if one were not immediately

perceiving it), but only claims about what is psychologically efficacious—that is, about the cognitive structures that matter to the thinker. But is this true? If it is the case that only a *subset* of information is carried over from perceptual processing to cognitive processing, though derived from the same source, might there still be a problem of reference?

Also related: what information is carried over from perceptual to conceptual processing? Isn't this "selection/extraction" problem eerily similar to the transduction problem faced by amodal theories of cognition? What are the principles that determine which sets of sensorimotor contingencies will be exercised during, say, activation of the perceptual symbol for DOG, and deducing from a perceptual simulation that a dog, if it wags its tail, is happy?

What implications does this view have for understanding "meaning"? Might it be the case that "meaning" is a quale, much like the "redness" of red, or the "hotness" of hot? That is, could the meaning of, say, "apple" be the qualitative feeling of what-it-is-like to interact effectively with apples, based on the implicit, procedural knowledge of the sensorimotor contingencies that define the actual and allowable interactions with apples? Given that O'Regan and Noë (2001) claim that the knowledge of sensorimotor laws, and the current exercise of that knowledge, determine the qualia of experience, is it too much of a leap to claim that the meaning of objects arises from the whole stretch of competent engagement with them? Admittedly, the problem of meaning and reference is a difficult one—beyond the scope of this paper—but the hope is that this discussion will serve as an "intuition pump" for more detailed analyses.

Conclusion

The main goal of this paper has been to elaborate a link between three deeply related, and mutually enriching areas of study. However, much of the difficulty still remains in determining how exactly the (putative) underlying processes (such as how event features are integrated into event codes) give rise to the functional properties of a PSS. But clearly, the possibility that sensorimotor contingencies, event codes, and perceptual symbols comprise the fundamental components of thought further suggests a deeper result: that now we might be more assured that suggestive correlations between perception and cognition (Goldstone & Barsalou, 1998) could now be given a principled causal basis for their interaction. We might now be in a better position to understand just in what respects perceptual processes might hold at the conceptual level; and accordingly, how we as researchers, and possessors of minds, might broaden and enrich our investigations.

Acknowledgments

This work was supported by NSF-ROLE award 21002/REC-0087516. I thank Dedre Gentner and Sam Day for valuable discussions. Also, I would like to thank Larry Barsalou, Bernhard Hommel, Alva Noe, and Kevin O'Regan for the time they devoted to answering my questions.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*, 249-277.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Barsalou, L. W., Solomon, K. O., & Wu, L. L. (1999). Perceptual simulation in conceptual tasks. Paper presented at the Cultural, typological, and psychological perspectives in cognitive linguistics: The proceedings of the 4th conference of the International Cognitive Linguistics Association, Vol. 3, Amsterdam.

Barsalou, L.W., Simmons, W.K., Barbey, A., & Wilson, C.D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7, 84-91.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chown, E., Kaplan, S., & Kortenkamp, D. (1995). Prototypes, locations, and associative networks (PLAN): Towards a unified theory of cognitive mapping. *Cognitive Science, 19,* 1-51.

Clancey, W. J. (1997). *Situated Cognition.* Cambridge: Cambridge University.

Clark, A. (1997). Being There: Putting Brain, Body, and World Together Again. Cambridge, MA: MIT Press.

Clark, A. (2002). Is seeing all it seems? Action, reason and the Grand Illusion. *Journal of Consciousness Studies,* 9.

Damasio, A. (1994). *Descartes' Error.* New York, NY: The Grosset Putnam.

Dourish, P. (2001). *Where the action is: The foundations of embodied interaction.* Cambridge, MA: MIT Press.

Dreyfus, H. (1972). *What computers can't do: The limits of artificial intelligence.* New York: Harper and Row.

Evans, G. (1982). *The varieties of reference.* Oxford: Clarendon Press.

Fodor, J. A. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture - a Critical Analysis. *Cognition, 28*, 3-71.

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space > time metaphors. *Language and Cognitive Processes, 17,* 537-565.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Hillsdale, NJ: Lawrence Erlbaum.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20,* 1-19.

Glenberg, A. M., Robertson, D. A., (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language, 43,* 379-401.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65,* 231-262.
Harnad, S. (1990). The Symbol Grounding Problem. *Physica D, 42,* 335-346.
Hommel, B., Musseler, J., Aschersleben, G., Prinz, W. (2001). The theory of event coding: A framework for perception and action planning. *Behavioral and Brain Sciences, 24*.
Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.
Johnson-Laird, P. N. (1989). *The computer and the mind : An introduction to cognitive science.* Cambridge, MA: Harvard University Press.
Laeng, B., & Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science, 26.*
Lane, P. C. R., Cheng, P. C. H., & Gobet, F. (2000). CHREST+: Investigating how humans learn to solve problems using diagrams. *Artificial Intelligence and the Simulation of Behavior (AISB) Quarterly, 103,* 24-30.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.
Markman, A. B., & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences. 4,* 70-75.
Mast, F. W., & Kosslyn, S. M. (2002). Eye movements during visual mental imagery. *Trends in Cognitive Sciences, 6,* 271-272.
Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.
O'Regan, J. K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral & Brain Sciences, 24.*
Philipona, D., O'Regan, K., & Nadal, J. P. (2003). Is there something out there? Inferring space from sensorimotor dependencies. *Neural Computation, 15.*
Pinker, S. (1998). *How the mind works.* London: The Penguin Press.
Pylyshyn, Z. W. (1984). *Computation and Cognition: Towards a Foundation for Cognitive Science.* Cambridge, MA: MIT Press.
Regier, T., & Zheng, M. (2003). An attentional constraint on spatial meaning. In Proceedings of the 25th Annual Meeting of the Cognitive Science Society.
Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition, 76,* 269-295.
Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition, Vol. 1: Foundations.* Cambridge, MA: MIT Press.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417-457.
Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences, 7,* 343-348.
Smith, L. B., Thelen, E. (Ed.). (1993). *A Dynamic Systems Approach to Development.* Cambridge, MA: MIT Press.
Solomon, K. O., & Barsalou, L. W. (2001). Representing properties locally. *Cognitive Psychology, 43,* 129-169.
Suchman, L. (1987). *Plans and Situated Actions.* Cambridge: Cambridge University Press.
Thelen, E. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.
Thelen, E., Schoner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences, 24.*
Thomas, N. J. T. (1999). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science, 23,* 207-245.
Van Gelder, T. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences, 21.*
Varela, F., Thompson, E., & Rosch, E. (1992). *The Embodied Mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.
Zwaan, R. A. (1999). Situation models: The mental leap into imagined worlds. *Current Directions in Psychological Science, 8.*

The Influence of Goal-directed Activity on Categorization and Reasoning

Ben D. Jee (bendj@uic.edu)
Jennifer Wiley (jwiley@uic.edu)

Department of Psychology, 1007 W. Harrison St.
University of Illinois, Chicago
Chicago, IL 60607 USA

Abstract

Studies of experts with different specialties in a domain suggest that goals may play an important role in category learning and inductive reasoning. Little experimental research, however, has addressed the potential influence of goal-directed activity on such processes. In the current study naïve participants acquire and utilize knowledge about a novel natural kind domain over a number of experimental sessions. Different groups of participants utilize different goals in their interactions with the experimental items. Participants utilizing a goal that requires the use of a certain subset of item features are found to develop categories around these features. Moreover, in a subsequent inductive reasoning task, these same participants perform in a manner that is highly consistent with the use of a goal-related category structure.

Introduction

Several researchers have encouraged a broadening of the methodology that has long predominated the psychological study of human categorization (e.g., Barsalou, 1991; Markman, & Ross, 2003; Medin, Ross, Atran, Burnett, & Blok, 2002; Murphy, 2002; Shafto, & Coley, 2003; Solomon, Medin, & Lynch, 1999). It is argued that categorization is implicated in a number of cognitive processes, such as, problem solving (Chi, Feltovich, & Glaser, 1981; Ross, 1996), inductive reasoning (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Proffitt, Medin, & Coley, 2000), and communication (Solomon et al., 1999; Markman & Ross, 2003), hence the study of categorization ought to reflect its multifaceted character. One area of research that has exemplified this broadened approach is the study of experts. These are individuals with extensive knowledge and experience in a given domain, such as, chess, medicine, birds, or trees.

Traditionally, studies of expertise have attempted to design tasks that elicit superior performances from experts in laboratory conditions so it may be systematically analyzed (Ericsson, & Smith, 1991). In most cases the performance of experts is compared with the performance of non-experts, typically recruited from undergraduate populations. However, recent research suggests that comparisons *among* experts with different specialties in a single domain may also provide valuable insight into the nature of expertise. With regard to theories of categorization, such comparisons could help to identify factors that contribute to the organization of conceptual knowledge that may be masked when research is restricted to expert-novice comparisons.

Inter-expert research has found that experts with different specializations in a domain may categorize and reason about domain-related information in characteristically different ways. For example, Medin, Lynch, Coley, and Atran (1997) found that different types of tree experts (taxonomists, landscapers, and maintenance workers) organized the same set of familiar tree species in distinct ways. Of particular interest is their finding that both landscapers and maintenance workers formed categories of trees based on their goal-related properties, such as, "weediness," "landscape utility," and "aesthetics." Moreover, in a subsequent inductive reasoning task maintenance workers made a number of inductive inferences consistent with their goal-related categorizations of the tree species. Proffitt et al. (2000) found that tree experts would often rely on causal-ecological relations over taxonomic relations in making inferences in their domain of expertise. As noted by Medin et al. (1997), the categories utilized by some types of tree experts closely resemble the goal-derived categories described by Barsalou (e.g., 1991): They crosscut the scientific/taxonomic organization of the domain and appear to be assembled around ideal attribute values.

Exploring the Influence of Goal Use

The above studies suggest that the nature of a domain, like trees, is but one factor contributing to the category structures that individuals form for it. It follows that research focusing exclusively on the effects of category structure on category learning may be missing an important aspect of this learning process (see also Markman & Ross, 2003). It may be that the influence of goal-directed activity on category learning is quite extensive, perhaps veiled by the characteristic ways in which members of a culture interact with the world. Indeed, cross-cultural studies have identified variations in categorization that may be related to such differences (e.g., Medin et al., 2002).

It is clear that evidence of cross-cultural and inter-expert variation in categorization and reasoning pose substantial challenges to the prevailing paradigm in categorization research. However, it is notoriously difficult to identify the causes of such variability. Numerous differences clearly exist between cultural groups, and even studies of different expert specialists from the same culture cannot control for differing experiences with the members of a domain.

Moreover, the relations between domain structure and domain use are likely to be highly complex. For instance, morphological properties can be related to both the taxonomic status of an object as well as its utility. Because domain structure and use are parameters that may vary freely in the real world, it is challenging to interpret cross-cultural and inter-expert comparisons. For these reasons experimentation is needed to isolate the roles played by these different factors.

Ross' recent work (e.g., 1996, 2000) represents an attempt to utilize experimental tasks to investigate the influence of goals on categorization, focusing particularly on the relation between categories and problem solving (cf. Chi et al., 1981). This research has demonstrated that the use of a category can alter its representation by enhancing the perceived importance of use-related features. For instance, in the domain of medical diagnosis, participants that learned that certain disease symptoms were good predictors of a treatment also perceived these symptoms as more diagnostic of the disease, even though they were no more diagnostic than other features that accompanied them. This research has primarily investigated how use affects an already-learned category. Thus, it is an unsettled question as to how goal-directed interaction affects category formation, particularly in natural kind domains. Examining the effects of goal-directed activity on inductive reasoning would provide an interesting extension of this line of research.

Overview

To examine the effects of goal-directed activity on categorization and inductive reasoning, the present study exposes naïve participants to an artificial natural kind domain (the "Creatures"). Over a four-session learning phase, these participants acquire knowledge about the different types of features possessed by the Creatures and use this knowledge to assess the Creatures according to a particular goal. These goals are manipulated between groups: One group is given a goal that directs them to utilize a certain subset of the Creatures' features, while the other group is given a goal that does not direct them to utilize any subset of features over another. Changes in categorization are assessed by having participants sort the Creatures both before and after the learning phase. An induction task is used to assess how goal-directed activity influences reasoning about the domain.

It is hypothesized that the performance of goal-directed activities will influence participants' categorizations and reasoning, but only in the case where participants must utilize a certain subset of the Creatures' features to perform the goal-related tasks. If these participants learn to re-organize their domain knowledge around ideal (i.e., goal-relevant) features, then a category structure that is based on goal-related features should provide a good account of their categorizations and their inductive inferences.

Method

Participants

Eight undergraduates from the University of Illinois at Chicago participated in this study to fulfill a course requirement. Participants were run individually. Four participants were randomly assigned to each condition.

Materials/tasks

Experimental stimuli. A set of 16 artificial Creatures was designed such that it could be sorted into three structurally equivalent hierarchies on the basis of three different feature types: Features related to Avoiding Predators (AP), Features NOT related to Avoiding Predators[1] (NAP), and Pictorially-represented features (PR). The PR features were represented in an illustration of the Creature, while the AP and NAP features were written in point form below the illustration (see Figure 1).

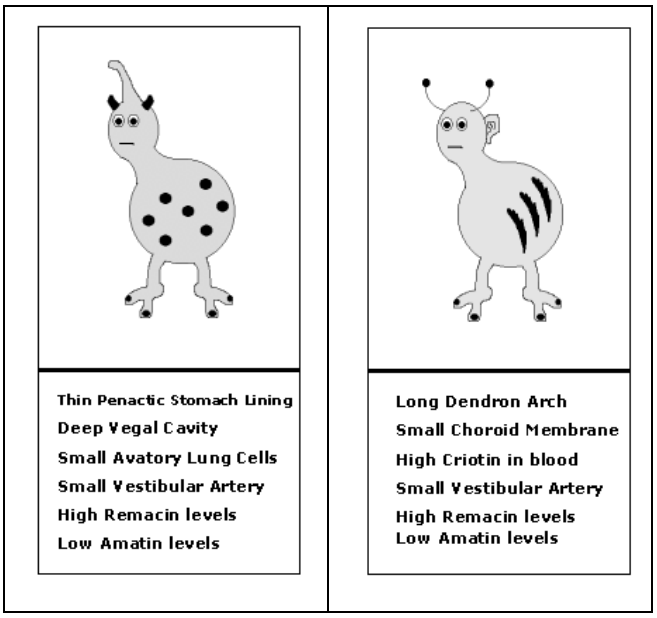


Figure 1: Examples of Creature cards. *Note.* Actual cards were in color.

Fourteen features of each type were distributed among the Creatures according to the structures in Figure 2. The dendrograms in Figure 2 show that for each feature type (AP, NAP, and PR), eight of the features are possessed by 2/16 Creatures, four of the features by 4/16, and two of the features by 8/16. These dendrograms can be understood as ideal category structures for each feature type – that is, they represent the three different category structures that would be formed by using each of the three different feature types. It is of interest to compare participants' categorizations with these ideal structures.

[1] These features were coherently related to another goal: assessing the Creatures according to how well they would serve as a source of nutrition. Participants in this experiment were not informed of this other goal.

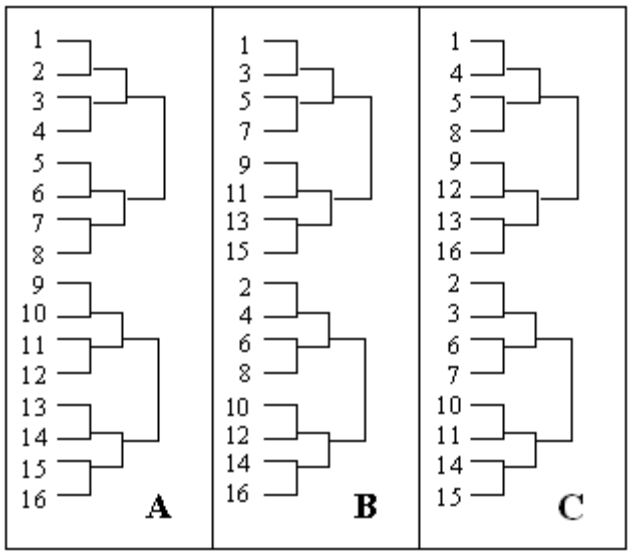


Figure 2: Dendrograms representing categories based on (A) Written features related to Avoiding Predators, (B) Written features NOT related to Avoiding Predators, and (C) Pictorially-represented features.

Goal Information. The independent variable in this study was whether or not participants were given a goal requiring them to utilize a certain subset of Creature features. All participants were informed that a large corporation had purchased an island habitat, and they want to introduce the Creatures to it. Participants in Group AP (Avoid Predators) were told that the corporation was interested in maximizing the survival of the Creatures on the island, and that they wanted to determine which Creatures would be best suited for avoiding predators. Participants in Group C (Control) were told that the corporation was interested in cataloguing the Creatures, and wanted to determine which Creatures would be easiest to monitor. Participants in both conditions were told that the corporation had hired them as a consultant, and that their job was to assess the Creatures according to either: (a) how well-suited they are for avoiding predators (Group AP) or (b) how easy they will be to recall (Group C). Participants were informed that the Creatures' features would be useful for making their assessments, and that they would learn more about the Creatures' features at the next session of the experiment.

Pre-training Tasks. Before engaging in goal-directed activities, participants had to be familiarized with the features that they would utilize. This pre-training is important in ensuring that differences between the two goal groups are not due to differences in familiarity with and knowledge of the features. The pre-training consisted of two parts, a learning phase and a test phase.

1. Feature Learning Task. In this task participants were required to study several computer screens of information about the Creatures' features and their relations to other properties. Half of these properties were coherently related to avoiding predators (AP) and half were not related to avoiding predators (NAP – see footnote 1). For example, participants had to learn that a "thin penactic stomach lining" means that the Creature cannot dig holes for hiding, and that a "small vestibular artery" means that the Creature has low vitamin C levels. Participants were permitted to study this information at their own pace.

2. Feature Memory Test. This test was designed to determine whether participants had learned the information about the features from the feature learning task. This computerized test consisted of 14 multiple choice questions about Creature features. Participants were required to re-take the test until they completed it without error. They were given up to 5 attempts per day. Four versions of this task were constructed. Each version contained the same questions, but with a different ordering of both questions and options.

Goal-directed Activity. The main task for the majority of the experiment was a **Creature Assessment task**. The task combined memory and decision making about the creatures. Students performed this task on days 2-5. At the beginning of each trial of this task, the subject was shown a Creature card on a computer screen for 10 s. Following this display the subject completed a 2-page assessment form. On page 1 the subject checked boxes next to the features possessed by the Creature that was just displayed. Participants in the AP condition were asked to check only the boxes next to the Creature's features that are relevant to avoiding predators. Participants in the C condition were asked to check the boxes next to all of the Creature's features. On page 2 of the assessment form the subject responded to 3 questions. Participants in the AP condition were asked: (1) Do you think the Creature that you were just shown will be good at avoiding predators in the habitat? (2) What about the Creature influenced your decision? (3) Rate the Creature on a scale of 1-10 according to how good it is likely to be at avoiding predators. Participants in the C condition were asked: (1) Do you think the Creature that you were just shown will be difficult to recall later on? (2) What about the Creature influenced your decision? (3) Rate your memory for the Creature on a scale of 1-10, according to how difficult it is likely to be to remember it. Following completion of the response form, the subject advanced the screen and the next item was displayed. Each of the 16 Creatures was shown in a different order once per day.

Dependent Measures. Two dependent measures were used to assess changes in categorization and reasoning about the Creatures. A sorting task (the Category Construction Task) was administered both before and after the training sessions. An Inference task was given on the final day of the experiment.

1. Category Construction Task. In this task, participants were asked to construct categories using the 16 Creature cards. Before sorting the cards, the subject was shown each card one-by-one for about two seconds. Following this initial exposure the Creature cards were randomly arrayed

on a table in front of the subject. The subject was instructed to find the "best way" to organize the cards. First, the subject was asked to form 8 pairs of Creatures and to provide justification for each group that they made. Second, the subject was asked to form 4 groups of 4 Creatures using the pairs that they constructed, and to provide further justifications. Finally, the subject was asked to combine the 4 groups of Creatures into 2 groups of 8 Creatures, and to provide justifications. This task was performed on both Day 1 and Day 6 of the experiment. On Day 6 the participants were instructed to use their knowledge of the Creatures when constructing categories, and they were not shown each card one-by-one as on Day 1. Otherwise, the procedure was the same on each day.

2. Inference Task. In this task participants were presented with a Creature on a computer screen and informed that the item has been found to possess a novel feature, "Sarca." The original item was removed after 10 s and 2 different Creatures were then shown. The subject was then asked to decide which of them is most likely to also possess the novel feature attributed to the first item. Following their response a blank screen was displayed for 2 s, followed by the next trial. Participants received two types of trials in which a more pictorially similar option (the PR option) was competed with either a more AP-related or NAP-related option. In AP vs. PR trials one option shared one more PR feature with the original item than did the other option. This other option, however, had one more AP-related feature in common with the original. Importantly, both options had exactly the same NAP-related features in common with the original. Thus, these trials tested whether participants would infer that the novel property would be more likely to generalize to a Creature that is more pictorially similar or one that is more similar with respect to AP properties. NAP vs. PR trials were analogously designed to compete an option sharing more PR features with one sharing more NAP-related features. The number of AP vs. PR and NAP vs. PR trials was matched. Trial order was randomized to create a single task list that was administered to all participants.

Procedure

The experiment took place on six consecutive weekdays (participants were not run on weekends). On day 1, participants performed the category construction task and received information about their goal. On day 2, participants learned about the properties related to the Creatures' features and subsequently completed the first version of the feature test. Participants then completed assessments of the Creatures, either with respect to how well suited they would be for avoiding predators (Group AP) or with respect to how easy it would be to recall them later on (Group C). Days 3-5 were the same in format. On these days, participants completed different versions of both the feature test and Creature assessments. Finally, on day 6 they completed the category construction task followed by the inference task.

Results

Feature Memory Test. Figure 3 displays the mean number of blocks (test attempts) to criterion across the four days that the test was administered. A 2 (Group) x 4 (Day) repeated measures ANOVA was conducted on the mean number of blocks to criterion. Participants that did not reach criterion on a given day were given a default score of 6 (five being the maximum number of attempts permitted). Across the first 3 days there were five instances in which a subject did not reach criterion, but by day 4 all participants reached criterion in 4 trials or less. The ANOVA revealed a significant effect of Day, $F(3, 18) = 11.61$, $MSE = 1.09$, $p < .01$, with no main effect of Group $F(1, 6) = 1.06$, $MSE = 6.62$, *ns*, and no interaction between Day and Group, $F(3, 18) = 1.95$, $MSE = 1.09$, *ns*. A follow-up comparison revealed a significant difference between the number of blocks to criterion required on Day 1 ($M = 4.6$) and Day 4 ($M = 1.9$), $F(1, 18) = 27.5$, $MSE = 1.10$, $p < .01$. These findings indicate that participants required fewer attempts to reach criterion by the final day.

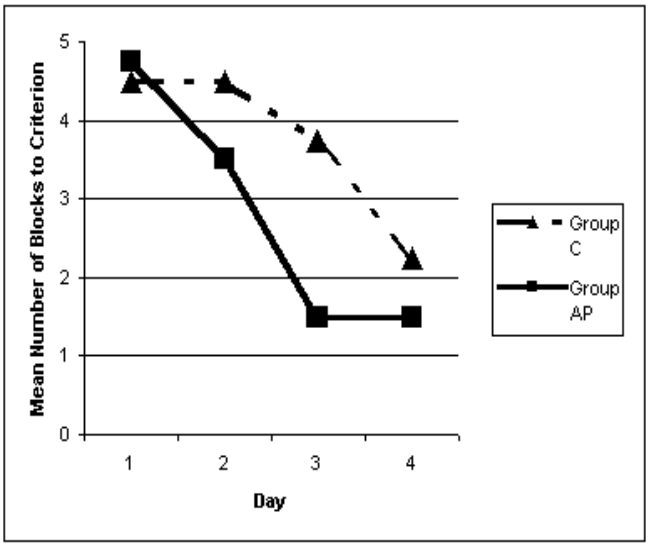


Figure 3: Mean number of blocks to criterion by day on the Feature Memory Task

The pattern of change in the number of blocks to criterion is consistent with the decline in error rates across days (see Figure 4). A 2 (Group) x 4 (Day) repeated measures ANOVA was conducted on mean error rates. This ANOVA revealed a significant effect of Day, $F(3, 18) = 9.16$, $MSE = 0.002$, $p < .01$, with no main effect of Group $F(1, 6) = 1.62$, $MSE = 0.005$, *ns*, and no Day x Group interaction, $F(3, 18) = 0.68$, $MSE = 0.002$, *ns*. A follow-up comparison revealed a significant difference between the error rates on Day 1 ($M = 0.13$) and Day 4 ($M = 0.04$), $F(1, 18) = 18.0$, $MSE = 0.002$, $p < .01$. These findings indicate that participants' accuracy on the Memory Test improved across the four days it was administered. Examination of the decrease in both blocks to criterion and error rate across days provides

evidence that participants in both groups learned and retained information related to all features.

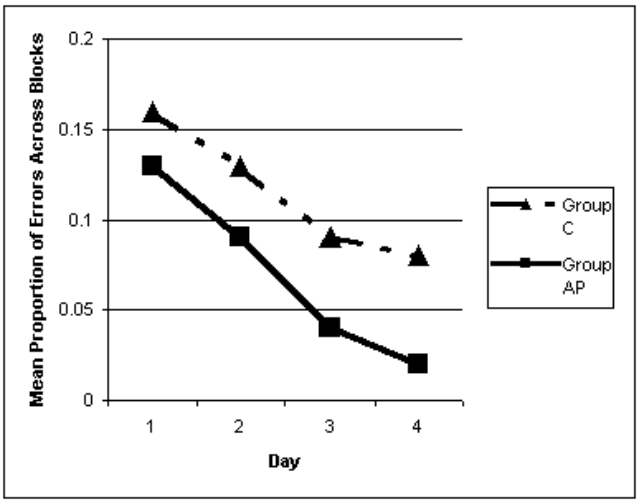


Figure 4: Mean error rate per block by day on the Feature Memory Task

Category Construction Task. Treatment of the category construction data closely follows Medin et al. (1997) and Shafto and Coley (2003). Participants' categorizations of the Creatures were converted to dendrograms (see Figure 2). Each subject's dendrogram was then used to derive a 16 x 16 pairwise distance matrix. A number was assigned to each cell of the matrix corresponding to the level in the dendrogram at which the pair of Creatures was combined. Creatures paired together at the lowest level (i.e., when the subject was instructed to make pairs of Creatures) were assigned a distance of 1; Creatures paired at the next level (groups of 4) were assigned a distance of 2; Creatures at the highest level (groups of 8) were assigned a 3; Creatures NOT paired at any level were assigned a default distance of 4. Only the 120 unique cells above the diagonal were used in subsequent analyses because of their redundancy with the cells below and because the cells on the diagonal (the distance between a Creature and itself) are irrelevant. Distances were averaged across participants in each group to create four different matrices representing initial and final category construction for Group C (C-1 and C-2) and AP (AP-1 and AP-2). Initial and final matrices for each group were correlated with one another and with the matrices derived from the ideal category structures (Figure 2).

Table 1 displays the correlations between the ideal category structures (APi, NAPi, and PRi) and the initial and final structures for each group (C-1, C-2, AP-1, and AP-2). The first thing to note in Table 1 is the relatively low intercorrelations among the ideal structures. This implies that a high correlation between a group structure and an ideal structure cannot be attributed to covariance among the ideal structures. Turning to the performance of the two groups, it is evident that on average participants in Group C initially sorted the Creatures in ways that were highly correlated with both the APi and PRi structures. In their final performance of the task, these participants produced category structures that are still significantly correlated with the APi structure, but are not significantly correlated with the PRi structure. Group AP's categories also have a high correlation with PRi at the initial stage, but unlike Group C they do not evidence a significant correlation with APi. However, at the final stage Group AP shows a high correlation with APi and a nonsignificant correlation with PRi. In summary, both groups appear to shift away from using PR features in their final category constructions, but only the participants in Group AP show an increased relation to APi.

Table 1: Correlations between ideal and subject category structures

Structure	NAPi	PRi	C-1	C-2	AP-1	AP-2
APi	0.15	0.15	0.50**	0.45**	0.13	0.91**
NAPi		-0.16	-0.08	0.03	-0.17	0.12
PRi			0.55**	-0.07	0.92**	0.12
C-1				0.41	0.49**	0.55**
C-2					-0.02	0.48**
AP-1						0.10

** $p < .01$ (two-tailed). *Note.* C-1 = Group C, initial sort; C-2 = Group C, final sort; AP-1 = Group AP, initial sort; AP-2 = Group AP, final sort

Inference Task. Participants' choices were scored according to whether they were consistent with the use of non-pictorial information. Thus, in the AP vs. PR trials the participant's choice was scored as a 1 if they chose the AP-related option and a 0 if they chose the PR option. In NAP vs. PR trials, NAP choices were scored as a 1, PR choices as 0. The mean proportion of non-PR choices was calculated for participants for each trial type. Figure 5 displays the mean proportion of non-PR choices for each trial type. A 2 (Group) x 2 (Trial Type) repeated measures ANOVA revealed no main effect of Trial Type, $F(1, 14) = 2.0$, $MSE = 0.05$, *ns*, and a main effect of Group, $F(1, 14) = 36.7$, $MSE = 0.031$, $p < .01$, however, there was also a significant Group x Trial Type interaction, $F(1, 14) = 6.58$, $MSE = 0.002$, $p < .05$.

Of primary interest is the comparison between the mean proportions of non-PR choices for each trial type with the proportion expected by chance (.50). For Group AP, the mean proportion of non-PR choices was significantly higher that chance on the AP vs. PR trials, $t(19) = 5.15$, $p < .01$, but did not differ from chance on the NAP vs. PR trials, $t(18) = 0.35$, *ns*. For Group C, the mean proportion of non-PR choices was marginally lower than chance on the AP vs. PR trials, $t(19) = 2.04$, $p < .10$, but did not differ from chance on the NAP vs. PR trials, $t(18) = 0.98$, *ns*. Thus, participants in Group AP most often generalized a novel property to a Creature more similar with respect to AP properties, but performed at chance level when AP-related similarity did not distinguish the two options. Participants in Group C did

not show this pattern, and evidenced a slight tendency toward the PR option in the AP vs. PR trials.

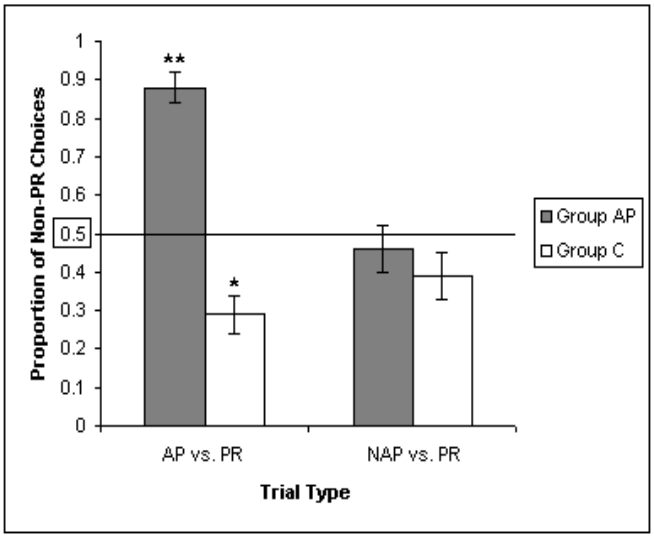


Figure 5: Mean proportion of Non-PR choices for each trial type in the Inference Task. *Note.* Asterisks indicate that the mean differs from chance (.50), ** $p < .01$ (two-tailed), * $.05 < p < .10$ (two-tailed).

Summary and Discussion

The main hypothesis of this experiment was that the utilization of a certain subset of Creature features in a goal-directed task would influence categorization and reasoning. It was further predicted that an ideal goal-related category structure would account well for the performances of these participants.

The results of the category construction task are consistent with these predictions. The APi structure correlated highly with Group AP's final categorizations. Even though Group C's categorizations also correlated highly with APi, this correlation showed no increase from this group's initial performance. The APi structure also accounts well for Group AP's performance on the inference task. When the APi structure distinguished between the two options, the participants often chose the AP-related Creature, when the APi structure did not distinguish the two options, participants performed at chance level. Altogether these findings suggest that participants in Group AP developed categories organized around a particular subset of goal-related Creature properties (AP-related properties). Participants in Group C did not evidence such change. By controlling the frequency of participants' exposure to the domain as well as its structure, the differences between groups can be attributed to differences in the goal-directed tasks that each group performed.

This research is still in its early stages and would surely benefit from increasing the sample size and adding further tests of participants' new domain knowledge. Nevertheless, the findings of this experiment serve to highlight the role that goal-directed activity could play in categorization. Given the extensive and varied nature of such activity within and across cultures, this role may deserve more attention than it has previously been paid.

Acknowledgements

This research was completed as part of a Master's degree requirement at the University of Illinois at Chicago, and was supported by a fellowship from the National Sciences and Engineering Research Council of Canada. Special thanks to Jeremy Eagles and Melinda Jensen for their assistance in running participants and entering data, and to Len Newman, Jim Pellegrino, and Rob Youmans for helpful discussions.

References

Barsalou, L. W. (1991). Deriving categories to achieve goals. In: G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 27). New York: Academic.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In: K. A. Ericsson, & J. Smith (Eds.), *Toward a General Theory of Expertise*. New York: Cambridge University Press.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129(4),* 592-613.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49-96.

Medin, D. L., Ross, N., Atran, S., Burnett, R. C., & Blok, S. V. (2002). Categorization and reasoning in relation to culture and expertise. In: B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 41). New York: Academic.

Murphy, G. L. (2002). *The Big Book of Concepts.* Cambridge, MA: MIT Press.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185-200.

Proffitt, J. B., Medin, D. L., & Coley, J. D. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(4),* 811-828.

Ross, B. H. (1996). Category learning as problem solving. In: D. L. Medin (Ed.), *The Psychology of Learning and Motivation* (Vol. 35). New York: Academic.

Ross, B. H. (2000). The effects of category use on learned categories. *Memory & Cognition, 28(1),* 51-63.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, Naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(4),* 641-649.

Solomon, K. O., Medin, D. L., & Lynch, E. L. (1999). Concepts do more than categorize. *Trends in Cognitive Sciences, 3(3),* 99-105.

The Acquisition of Intellectual Expertise: A Computational Model

Lisa C. Kaczmarczyk (lisak@cs.utexas.edu)
Department of Computer Sciences, The University of Texas at Austin
1 University Station C0500, Austin, Texas 78712 USA

Risto Miikkulainen (risto@cs.utexas.edu)
Department of Computer Sciences, The University of Texas at Austin
1 University Station C0500, Austin, Texas 78712 USA

Abstract

Intellectual expertise means knowledge and ability that a person has that allows them to solve complex problems. It is important to understand how people become experts so that we can improve educational strategies, and help learners achieve their full academic potential. Unfortunately, the process of acquiring intellectual expertise is not well understood. Artificial neural networks (ANNs) have already been successful in modeling other types of human learning. This paper shows that they can also be trained as a model of expert human learning, and address many of the difficulties found in trying to study expertise in humans. The results confirm three hypotheses: (1) An artificial neural network can be used as a model to investigate how people learn under different training scenarios; (2) Different methods for delivering the training material result in different final performance; (3) The best performance is achieved by incrementally increasing the complexity of the material. These results provide educators with computational evidence that structured, integrated delivery methods are better for learners than oversimplification and isolation of learning tasks.

Introduction

An intellectual expert has achieved a level of cognitive development in which she or he can rapidly grasp subtleties of complex problems, and produce very high quality solutions. A goal of formal education is to help students achieve an expert level of understanding in their chosen field. It is important to understand the nature of expertise so that we can improve educational strategies. As a result of many research studies about expertise, we know a lot about the characteristics of experts. However, there is a lot we do not understand about how to become an expert. It is not easy to create experts, whether human or computational. The learning process is complex and human studies are difficult. Understanding how to acquire intellectual expertise has proven elusive for educators, psychologists and students alike.

A primary goal of the study reported here is to increase understanding of the process by which humans become intellectual experts. In particular, how can people develop the ability to look at a problem statement and immediately select the best solution strategy? The second main goal is to understand this process in the context of formal instruction; specifically, how does the strategy by which material is delivered to the learner affect learning and conceptual development?

This paper presents results from a series of computational experiments examining how different delivery methods influence learning and conceptual development. These experiments use a real-world adult educational problem: the ability to identify correct solution strategies for calculus integration problems. The goal is to show that an artificial neural network can be used as a model to investigate how people learn under different training scenarios, and that different delivery methods result in different overall performance. The main results include: (1) errors are higher on final exams when different problem types are learned in isolation; (2) cramming just prior to taking final exams does not significantly improve performance. Different delivery strategies affect learning in different ways: (1) traditional sequential delivery methods inhibit learning and retention; (2) integrated delivery methods increase learning and retention; (3) the best performance comes from delivery methods that incrementally increase the complexity of material. These results can be applied to developing better training methods for people.

Prior Research on Intellectual Expertise

Studies of human expertise and understanding have revealed key information about experts. We know that experts and novices categorize problems differently, and that this categorization takes place before the subject attempts to solve the problem (Chi, Feltovich, and Glaser 1981). We also know that experts can categorize problems without solving them (Robinson and Hayes 1978). Finally, there is strong evidence that routine problems are solved not by intense calculating but rather by recognizing a type of problem (categorizing) and then using the stored knowledge about how to solve problems of that type (Reiman & Chi '89 referenced in (Ross and Spalding 1991).

Most studies of expertise have focused on what an expert knows, rather than the process by which she or he attained expertise. As a result, we know a lot less about this learning process than we do about expertise itself. Expert behavior does not simply follow a script: the greatest expertise is the result of long-term practice (Hayes 1989) that is consciously goal directed, self-monitoring, and self-adjusting within the setting of each particular task (Garner 1990). In addition, many studies have shown that meta-cognition (self-appraisal and self-management of cognition) is critical for successful academic learning (literature surveyed by Paris and Winograd (1990)). Since we know that experts categorize extremely well, it is possible that categorization ability and goal-directed meta-cognition enhance one another. When these abilities merge, intuition may be the result: there is strong evidence that experts rely upon their accurate intuition and a holistic recognition of appropriate actions (Dreyfus and Dreyfus 1986).

Cognitive scientists have often studied mathematics learning, due to the abstract nature of its concepts. Bruner has even suggested that learning mathematics may be viewed as a microcosm of all intellectual development (Bruner and Kenney 1965). A particularly interesting early connectionist model of mathematics learning was presented by Viscuso, Anderson, and Spoehr (1989). Their artificial neural network (ANN) simulated qualitative reasoning while doing multiplication. In summarizing their model, Viscuso et al correctly pointed out that the most important contribution of their model was that it mimicked the manner in which experts rely not so much on formal logic and rules but on their "sense" of what is correct. Another interesting ANN system learned to perform arbitrarily long addition problems (Cottrell and Tsung 1993). Their model learned the implicit underlying rule of addition. This system showed that ANNs can account for conceptual development: the network learned an important concept on which it had not been explicitly trained. In the decade since these studies were published, there has been quite a bit of work in related areas, such as the development of basic numerical abilities in infants and children (literature surveyed in (Ahmad, Casey, and Bale 2002)), and in childhood strategy development (Bray, Reilly, Villa, and Grupe 1997). However, we still do not understand how adult human experts learn to "sense" important concepts. It is important to understand this ability, so that we can better educate students.

The Calculus Domain

Calculus, at its most fundamental level, is based upon abstract cognitive concepts. As a result, understanding how people best learn calculus requires understanding the mind. The current educational debates over mathematics and science education partly result because we do not understand enough about how the brain produces cognition and conceptual understanding. In order to become calculus experts, students need to understand complex concepts and intuitively select the most efficient methods to solve problems. Educators need to understand what methods of delivering material will most help students achieve these abilities.

In the last decade math and science have been at the center of an increasingly wide-spread national concern with properly educating citizens for the new technological age. In college, students who want to major in science or engineering usually have to first perform well in calculus, which turns out to be a major obstacle for many of them.

In order to clearly identify what kinds of problems calculus students were having at at the University of Texas at Austin (UT), we conducted structured interviews with mathematics faculty and teaching assistants (TAs). The results fit well with the psychological literature on expert/novice behavior. Faculty and TAs reported that novice learners (in this case UT students) are often unable to select the correct solution strategy. This problem arises before they even have a chance to exhibit computational difficulties and prevents many from reaching timely, correct solutions. Conversely, the experts claimed an ability to "just see" the correct strategy, yet were unable to articulate how they knew. Probing revealed that although there are "rules of thumb" to assist in this domain, they are not comprehensive and do not cover many common scenarios. Experts instead pointed to general patterns and categorization that they have learned to recognize via extensive practice.

Successful problem solvers categorize math problems based upon underlying structural similarities and fundamental principles (Schoenfeld and Herrmann 1982). These categories are often grouped based upon solution strategies, that the experts then use to calculate an answer (Owen and Sweller 1989). How such strategies are formed is poorly understood. What regularities are most likely to be noticed, and how does the form in which the initial procedure is learned affect what is noticed? From the point of view of education, are there ways of managing how learners practice, to enhance the likelihood that they will notice these regularities, and incorporate this information into their problem-solving strategies?

One of the first instructional decisions is what order to present the material in, and how to move from one concept to the next. There are many possible orderings of material, and a computational model can be used to explore them. The model presented in this paper, described in the next section, contributes to achieving this research goal.

The ANN Model

The particular calculus problem chosen for the study is to decide whether a given integration problem should be solved with Simple Integration (Simple), Integration by U-Substitution (Usub), or Integration by Parts (Parts). This section describes the architecture of the artificial neural network as well as the training and test data, its encoding, and the experimental methods used in all the experiments described in this paper.

Architecture and Data

The model is an artificial neural network utilizing the backpropagation algorithm (for details of the algorithm see Bishop 1995) created using the LENS network simulator (Rohde 1999). The network is fully connected, and has 55 input nodes and 20 hidden nodes. The 55 input nodes make up a vector large enough to represent the features of one calculus integration problem containing up to four terms. The 20 hidden units were determined to be appropriate by experimentation; the results were not effected by small changes in size.

The input data consists of 957 calculus integration problems based upon examples found in college level calculus textbooks. Feature coding is a logical choice for representing them, given that both novices and experts use the features of a problem to determine which approach to use (Chi et al. 1981). The 55 unit input vector contains a series of 0s and 1s that map operators/operands to their location in the calculus integration problem. Short problems are padded with blanks. The vector consists of

- Four 2-unit terms representing constants and variables.
- Four 8-unit Unary Operators, representing $\sin, \cos, \tan, \cot, \sec, \csc, \ln$, exponentiation $e(x)$.
- Three 5-unit Binary Operators, representing multiplication, division, exponentiation $\hat{}$, addition, subtraction.

For example, the problem

$$3 + cos(x) - sin(y) + ln(x)$$

is coded in postfix form as: 01 00000000 10 01000000 00 10000000 10 00000010 00010 00001 00010, where the components are

01 : No Variable; Constant (i.e. 3)
00000000 : NONE (i.e. no unary operator for the constant)
10 : Variable (i.e. x); No Constant
01000000 : cos (of the variable x)
10 : Variable (i.e. y); No Constant
10000000 : sin (of the variable y)
10 : Variable (i.e. x); No Constant
00000010 : ln (of the variable x)
00010 : +
00001 : -
00010 : +

The network has three output nodes, each of which represents one of the possible integration strategies, Simple, Usub, Parts. Because the network is trained with one active target at a time, it learns to represent how confident it is in each choice (Bourlard and Wellekens 1990). For example, if the network reports activation values at 12%, 85%, 3%, then it is quite confident in the second category, considers the first category possible but unlikely, and the third category extremely unlikely (but not absolutely impossible). This percentage represents the *confidence level* that the network has in each answer.

Experimental Design

The calculus integration problems were divided into 10-fold cross-validation training and test sets (splits, or learning experiments). In each experiment the training set was input to the network, one problem at a time, in random order, and the test set was used to measure performance. Validation sets were not used because each learning experiment represented training one subject and the training time had to be constant, to compare how well the subjects learned. Three different types of learning experiments were run. Each experiment was run ten times, randomly resetting the initial network weights each time. Thus the whole study consisted of 300 learning experiments. This way it was possible to model the behavior of many different subjects and watch for both emergent patterns and individual variation.

During the test phase, there was always only one correct answer to a problem. This answer, called the "Best", was the answer suggested in a textbook, or by a calculus expert (faculty, TA). For each test problem the network reported how confident it was that the solution strategy was either Simple, Usub or Parts. If the confidence level for all solutions was below 80%, the problem was considered having "stumped" the network.

Results

Two sets of experiments (Drill and Test, Fully Integrated Learning) validated the ANN as a model of human learning. These experiments showed that the model accurately matches results from past educational research. In addition, these experiments provide insight into how the learning process occurs. The third set of experiments provided a computational prediction that a different type of learning (Incremental) produces the best performance.

Validating the Model: Drill and Test Learning

The first set of experiments, called "Drill and Test", mimicked a classic form of delivery that results in poor long-term retention in humans (Resnick and Ford 1981). In this method, concepts are introduced to the learner one at a time, with no overlap between topics. At the end of each topic, the learner is given a midterm exam (of previously unseen examples) on that concept.

After it has been trained with all concepts, the learner is given an opportunity to "cram", i.e. train on all concepts for a short period of time. At the end of all material, there is a comprehensive exam consisting of the entire test set.

In order to monitor the progress of learning quantitatively, and to compare to other approaches, each network was also tested during each epoch in two ways: (1) with the current midterm exam, illustrating the performance that the teacher would see in the classroom (Figure 1a), and (2) with the comprehensive exam, monitoring progress in learning the entire task, but broken into separate numbers for the different concepts (Figure 1b).

The main result was that the model, like humans, only remembers the most recently introduced concept well. More specifically, in 100 experiments run using Drill and Test, most networks (83%) rapidly learned to identify each of the concepts in turn (Simple, Usub, Parts). On midterm exams, the network often recognized 100% of the problems belonging to the concept that had just been studied. However, in spite of the opportunity to cram first, when the comprehensive final exam was given, these learners performed poorly, averaging 41.65% (standard deviation 6.35). The highest score was 54.55%. The remaining 17% of network learners were unable to make the switch from Simple to Usub problems, and then to Parts problems: their Usub and Parts midterms usually scored 0%. When these learners crammed and then took their comprehensive exams, they scored on average 17.29% (standard deviation 4.95), with a high score of 26.92% All learners in these experiments were extremely confident in their answers, even when they were wrong.

Validating the Model: Fully Integrated Learning

A second set of experiments mimicked human learning using an approach called "Fully Integrated Learning". This approach is inspired by the immersion experiences popular in foreign language learning (Spolsky 1989): the learner is placed in an environment where she or he is completely surrounded by the stimuli to be learned. The cognitive mechanisms that enable a foreign language student to sort out important grammatical features might not be that different from those cognitive mechanisms that sort out features of mathematical structures. In the Fully Integrated Learning experiments, there was only one training period, during which the networks were trained on all of the problem types simultaneously. During each epoch, the Simple, Usub and Parts training problems were input to the network in random order. Exams using the entire test set were given after every training epoch.

Fully Integrated Learning produced significantly better results than the Drill and Test delivery experiments (Figure

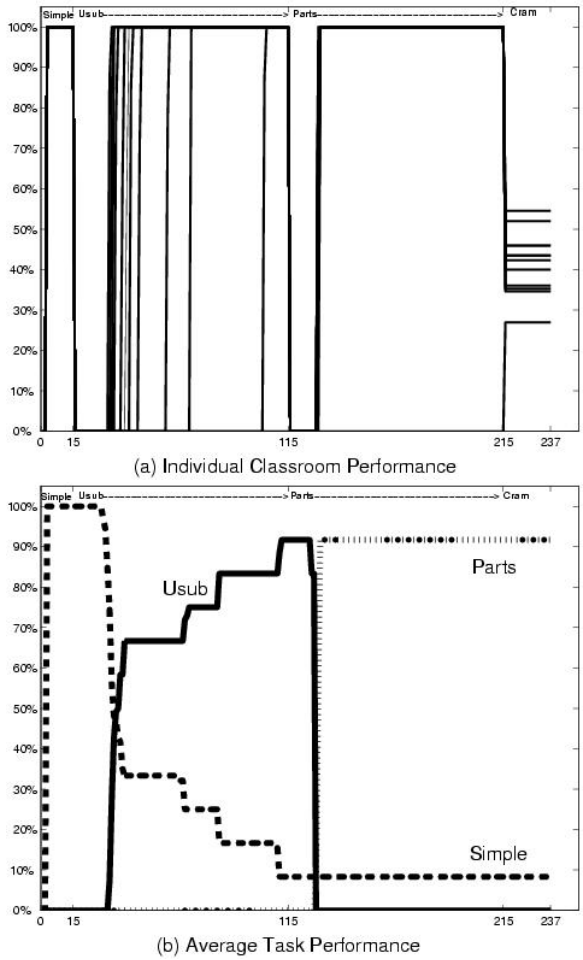


Figure 1: **Drill and Test Learning.** (a) The classroom performance of 12 representative learners, i.e. their accuracy on the current midterm (Simple, Usub, Parts, Cram periods) and comprehensive exam. Exam scores are on the y-axis, and the training epoch is shown along the x-axis. Scores on the comprehensive exam were poor - even with the aid of a cram session the highest score was 54.55%. (b) The average performance of all learners on the comprehensive exams, broken down by concept. Each problem type is forgotten when a new topic is learned.

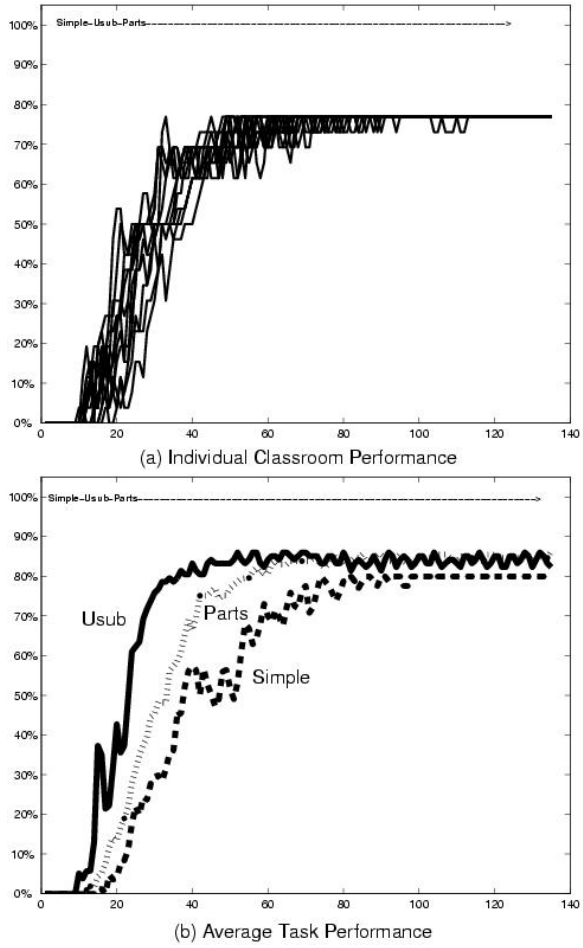


Figure 2: **Fully Integrated Learning.** (a) The classroom performance of 12 representative learners on the comprehensive test set over the course of learning. The learners initially failed the exams, but their scores rapidly increased, and finally plateaued. Improvement was not smooth, reflecting the trial and error process of learning. The best exam score was 80.76%. (b) Average performance of all learners broken down by concept. Usub problems were learned fastest, Simple problems slowest. Final results for Simple, Usub and Parts were similar.

2). The average score on the final comprehensive exam was 76.99% (standard deviation 7.94). The highest score was 80.76%. In contrast to Drill and Test, confidence in Fully Integrated learning closely reflected exam scores. The errors that were made on the exams followed a pattern of slow, gradual learning, spread across all problem types. The Fully Integrated Learning results as a whole replicated human data showing that immersion results in better longer-term retention than does Drill and Test.

Extending the Model: Incremental Learning

The third set of experiments was designed to test the hypothesis that the best learning of material is obtained by a delivery approach called "Incremental Learning". This approach is inspired by the result in the machine learning community that it is often most effective to tackle large computational tasks by starting with small problems and gradually increasing their complexity (Elman 1991; Gomez and Miikkulainen 1997). When there are a large number of co-dependent variables, it is hard to discover the role that each one plays in the problem and its solution. Therefore, an Incremental Learning delivery introduces new, increasingly complex concepts along with reinforcement of old concepts.

As with the Drill and Test experiments, there were three training periods. The network was first trained to identify Simple problems. During the second training period, Usub problems were added to the Simple problems, and for the third training period, Parts problems were added. The classroom performance was measured with Simple tests during the the first period, Simple and Usub test problems during the second, and the entire test set during the third (Figure 3a). The progress in learning the entire task was monitored with the entire test set, broken down by concept (Figure 3b). As in the Drill and Test experiments, Simple-only midterms very rapidly reached scores of 100%. When Usub prob-

lems were introduced, test scores began to fluctuate severely. Scores would drop to, or near, zero, rebound, and then drop again, as the network struggled to distinguish the new concept (Usub) from the old concept (Simple). Over time, although fluctuation continued, overall test scores increased. In a few cases, SU midterm scores reached 100%, however the majority of cases peaked at 70-75%. When Parts problems were introduced, the pattern of fluctuating scores was accentuated. Midterm scores immediately plummeted, although it is interesting to note that even the downward drop was often not smooth, but marked by brief plateaus and recoveries. Performance continued to deteriorate for longer than in the SU training segment, with scores fluctuating lower and lower. In contrast to the SU midterm scores, SUP midterm scores appeared to tighten in closer and closer to complete failure (for a while nearly all midterms fluctuated well under 20%). This behavior is predictable, because it is harder to distinguish three concepts from one another than two concepts. Eventually, performance began to improve, with prominent individual differences, as each network learner identified subtle patterns to accurately identify each concept. Eventually, virtually all midterm scores surpassed 70%. The average score on the final comprehensive exam was 81.9% (standard deviation 8.23). It is important to note that the final test results for Incremental Learning were better than either Drill and Test or Fully Integrated Learning, in spite of interim results that sometimes appeared poorer than either other type of experiment. The maximum exam SUP score was 95.6%, higher than any score reached in a Fully Integrated learning experiment. As evaluated with a t-test, the Incremental Learning final exam scores were higher than those of the Fully Integrated learning ($t = 1.9574, df = 11.869, p = 0.07423$).

The types of errors that the network made followed a pattern. As each new training period began, the network appeared to "flail", choosing first one answer then another on successive exam questions. However, this "flailing" gradually lessened and the network learned to correctly select each problem type simultaneously. As with the Fully Integrated Learning experiments, the learners' confidence levels closely reflected exam scores. The Incremental learning experiments showed that the best performance is achieved by introducing increasingly complex concepts gradually, allowing learners to build on their existing knowledge, and gradually pay more attention to finer distinctions.

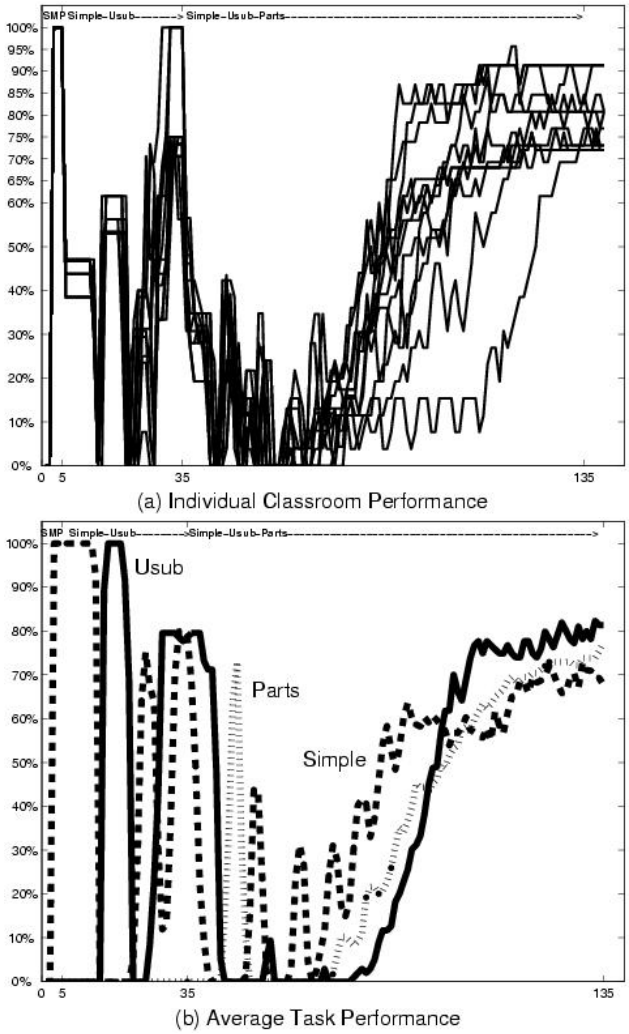


Figure 3: **Incremental Learning.** (a) The classroom performance of 12 representative learners on the current midterm (Simple, Simple-Usub, Simple-Usub-Parts). The maximum comprehensive exam score was 95.6%, higher than any score reached in a Fully Integrated learning experiment. (b) Average performance of all learners broken down by concept. Each problem type followed the same pattern of fluctuation between learning and apparent forgetting. Over time fluctuation lessened and performance improved for all problem types. Simple problems fluctuated the most and the longest.

Discussion and Future Work

Calculus integration problems that are often given to novice learners were used to study the process of learning to accurately categorize them by solution strategy. These strategies - Simple Integration, Integration by U-substitution, Integration by Parts - represent complex concepts that students need to intuitively master in order to become calculus experts. Drill and Test experiments and Fully Integrated experiments validated the model by showing that it can mimic known data about human learning. Drill and Test experiments supported the hypothesis that delivery methods that rigidly separate concepts during learning result in poor long-term retention of material. Also supported was the hypothesis that when concepts are reinforced inconsistently, only the most recently introduced concept is remembered, and that cramming does not improve learning. The nearly perfect midterm exam scores seen in Drill and Test experiments were misleading. They implied a level of interim learning and understanding which was not supported when the final exam require the learner to distinguish complex concepts.

Fully Integrated learning experiments supported the hypothesis that if problems that belong to one concept are introduced along with problems that belong to other concepts, error rates are smaller than when the same concepts are introduced separately. Over time, Fully Integrated learners performed quite well on their exams and although they are not perfect, can be claimed to have learned the task.

The results for Incremental Learning were very different from either Drill and Test or Fully Integrated learning. By introducing new problem types in a structured manner, the

network learner is allowed to focus on a smaller set of characteristics at the beginning of learning. Just as the first concept (Simple integration problems) is acquired, additional problems (Usub) are mixed in. The resulting confusion is apparent in the fluctuating midterm scores. Over time, as the learner grapples with the two contrasting problem types, confusion diminishes and midterm scores rise. When Parts problems are introduced, it becomes again more difficult to discriminate between the concepts. However, it is far more difficult to compare three related problem types than two. The confusion lasts longer and is more difficult to resolve, and individual learner differences become more apparent. Fortunately, the "priming" effect of the previous training segments allows most Incremental Learning learners to eventually do well, and in most cases better than the Fully Integrated learners.

An interesting direction of future research is to analyze the conceptual development that took place in the model during the different types of delivery methods. Using techniques such as Independent Component Analysis (ICA) of hidden layer representations it may be possible to discover how the network learners represent the problems as the learning progresses. In addition, the predictions on Incremental Learning can be tested in a study with human subjects. If confirmed, these results strongly suggest that a structured incremental approach should be used in teaching for expertise.

Conclusion

The experiments reported in this paper support the following three hypotheses: 1) An artificial neural network can be used as a model to investigate how people learn under different training scenarios 2) Different delivery methods result in different overall performance 3) Incremental Learning results in better performance than either Drill and Test or Fully Incremental learning. These results provide new insight into how humans learn complex cognitive tasks. As a result, educators have computational evidence that structured, integrated delivery methods lead to better performance for learners than oversimplification and isolation of learning tasks. They also have evidence that introducing many complex concepts at the same time does not produce the best learning either. The work encourages educators to focus on finding the optimal balance between introducing complexity and providing structured guidance. Finally, educators are reminded that interim results that reflect struggle with complex concepts will result in longer term performance gains than near perfect results in the short term.

References

Ahmad, K., Casey, M., and Bale, T. (2002). Connectionist simulation of quantification skills. *Connection Science*, 14(3):165–201.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Univ. Press.

Bourlard, H., and Wellekens, C. J. (1990). Links between Markov models and multilayer perceptrons. *IEEEPAMI*, PAMI-12:1167–1178.

Bray, N., Reilly, K., Villa, M., and Grupe, L. (1997). Neural network models and mechanisms of strategy development. *Developmental Review*, 17(2):525–566.

Bruner, J., and Kenney, H. (1965). Representation and mathematics learning. In Morrisett, and Vinsonhaler, editors, *Monographs of the Soc. for Research in Child Development ser. 99*, vol. 30-1. Univ. Chicago Press.

Chi, M., Feltovich, P., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5:121–152.

Cottrell, G., and Tsung, F. (1993). Learning simple arithmetic procedures. *Connection Science*, 5(1):37–58.

Dreyfus, H., and Dreyfus, S. (1986). *Mind Over Machine: The Power of Human Intuition and Expertise*. Free Press, Macmillan, Inc.

Elman, J. (1991). Incremental learning, or the importance of starting small. *Cognitive Science*, 443–448.

Garner, R. (1990). When children and adults do not use learning strategies. *Rev. of Educational Research*, 60(4):517–529.

Gomez, F., and Miikkulainen, R. (1997). Incremental evolution of complex general behavior. *Adaptive Behavior*, 5:317–342.

Hayes, J. (1989). *The Complete Problem Solver*. LEA.

Owen, E., and Sweller, J. (1989). Should problem solving be used as a learning device in mathematics? *JRME*, 20(3):322–328.

Paris, S., and Winograd, P. (1990). How metacognition can promote academic learning and instruction. In *Dimensions of thinking and cognitive instruction*. LEA.

Resnick, L., and Ford, W. (1981). *The Psychology of Mathematics for Instruction.* LEA.

Robinson, C., and Hayes, J. (1978). Making inferences about relevance in understanding problems. In Revlin, R., and Mayer, R., editors, *Human Reasoning*. V.H. Winston and Sons.

Rohde, D. (1999). Lens: The light, efficient network simulator. *Technical Report CMU-CS-99-164, Department of Computer Science*.

Ross, B., and Spalding, T. (1991). Some influences of instance comparisons on concept formation. In Fisher, D., Pazzani, M., and Langley, P., editors, *Concept Formation*. Morgan Kauffman.

Schoenfeld, A., and Herrmann, D. (1982). Problem perception and knowledge strucuture in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 8:484–494.

Spolsky, B. (1989). *Conditions for Second Language Learning*. Oxford Univ. Press.

Viscuso, S., Anderson, J., and Spoehr, K. (1989). Representing simple arithmetic in neural networks. In Tiberghien, G., editor, *Advances in Cognitive Science*, vol. 2. Ellis Horwood and John Wiley and Sons.

Transfer of learning between isomorphic artificial domains: Advantage for the abstract

Jennifer A. Kaminski (kaminski.16@osu.edu)
Center for Cognitive Science
Ohio State University
210F Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Andrew Heckler (heckler@mps.ohio-state.edu)
College of Mathematical and Physical Sciences
Ohio State University
425 Stillman Hall, 1947 College Road
Columbus, OH 43210, USA

Abstract

Transfer between isomorphic domains was investigated. Thirty college undergraduate students learned two isomorphic artificial systems. One system was concrete in the sense that it was perceptually rich and dynamic, while the other was abstract, involving written symbols. The results show significant positive transfer from the abstract domain to the concrete domain and no significant transfer from the concrete to the abstract.

Introduction

One of the goals of successful learning is transfer, or the ability to apply acquired knowledge outside the learned situation. Although a desired outcome of learning, spontaneous transfer is notoriously difficult to achieve. In the past few decades, numerous studies document poor or non-existent spontaneous transfer across isomorphic situations (Ben-Zeev & Star, 2001; Gholson et al., 1997; Holyoak, Junn, & Billman, 1984; Holyoak & Koh, 1987; Schoenfeld & Herrmann, 1982). Poor performance has been attributed to surface features distracting from underlying structure.

Which aspects of the learning situation facilitate transfer? A widely held belief in the education community has been that learning and transfer of mathematical and scientific knowledge is facilitated by the use of concrete representations of more abstract mathematical and scientific principles. In the past several decades, the use of concrete representations has been a growing part of the mathematics curriculum. Concrete representations include both physical manipulatives as well as specific instantiations of abstract concepts. They are often perceptually rich and meaningful. Mathematical concepts are traditionally represented in an abstract symbolic form, while applications of the mathematics to scientific and real-world scenarios can be thought of as concrete instances of the abstract concept.

The National Council of Teachers of Mathematics (NCTM) reform movement launched in 1989 promoted the role of such representations in the curriculum. For example, Dienes blocks (Dienes, 1960) are used in elementary mathematics education to teach arithmetic and place value. Dienes blocks are concrete proportional representations of the base-ten number system. The belief of educators who use the blocks is that through their use, young children will not only be able to represent and execute arithmetic problems, but will also be able to gain insight into the structure of the base-ten number system (Fuson & Briars, 1990).

Much support for the use of either concrete manipulatives or concretely situated applications in the learning of mathematics comes from constructivist educators. Cobb, Yackel, and Wood (1992) propose that students actively construct mathematical knowledge in social contexts. Furthermore, they suggest that topics which are applications of mathematics such as those from real-world or scientific situations provide good initial instructional activities. That is, instruction of mathematical concepts should be initiated through applications of the mathematics as opposed to initiated in symbolic form.

These approaches to learning and transfer seem to echo the Piagetian theory, according to which education should parallel the process of cognitive development, and the ability for abstraction is not achieved before the formal operational stage (Inhelder & Piaget, 1958). At the same time, children's reasoning during the preceding stage (i.e., the concrete operational stage) was said to be limited to objects and physically possible situations. If learning parallels the process of development, then transfer from more concrete to more abstract representations should be more efficient than the reverse.

However, there are strong reasons to doubt this view. First, it has been demonstrated that concrete, perceptually rich objects are more likely to be considered objects than symbols denoting other entities (DeLoache, 1987; 2000). In a series of studies by DeLoache and colleagues, very young children were shown the location of a toy in either a photograph or a physical model of a scale room. They were then asked to retrieve the toy from the actual room. Almost all (88%) of the children shown the photograph were able to make an errorless retrieval of the toy, while only 16% of the children shown the physical model were able to do so. When the model was placed behind a screen, children's retrieval rate improved. Furthermore, slightly older children are very successful in this task. However, when older children were encouraged to play with the model, performance dropped significantly. These studies demonstrate that children have difficulty treating perceptually rich objects as symbols. Decreasing the salience of the object increased the ease of its symbolic use.

Second, there is a large body of literature on analogy (analogy is variant of transfer of knowledge from one domain to another) indicating that properties that are not a part of to-be-learned knowledge (i.e., surface features) may hinder rather than facilitate learning (e.g., Ross, 1987; 1989).

Third, there is recent evidence that there might be a competition between abstract and concrete representations of the same situation, and salient concrete representations may distract learners from more abstract regularities (Goldstone & Sakamoto, 2003).

Finally, there is evidence that transfer from abstract instantiations of knowledge may be in fact easier than transfer from concrete to abstract instantiations (Bassok & Holyoak, 1989). Bassok and Holyoak examined transfer between more abstract algebraic knowledge and more concrete physics knowledge, namely between arithmetic-progression problems and isomorphic constant-acceleration problems. High school and college students (who were unfamiliar with both of these domains) learned one of these topics and then were posed word problems involving the other topic. The measure of transfer was whether the learned method had been applied to the structurally isomorphic problems in the unstudied domain. Students who had learned arithmetic-progression first easily and spontaneously applied the learned method to correctly solve constant-acceleration problems. However, the students who learned the physics topic showed essentially no transfer of method to the arithmetic-progression problems. The results of this study suggest that transfer is more likely to occur from a more abstract instantiation to a concrete isomorph.

While the Bassok and Holyoak study (1989) certainly implies that more transfer occurs from abstract to concrete domains, confounds in the study limit such a broad conclusion. The chosen topics in mathematics and physics, as any mathematical and physical topics, do not exist in isolation. Any individual has many associations with each, including related prior learning as well as attitudes and beliefs. Specifically, the amount of mathematics learned through elementary, middle, and high school is significantly more than the amount of physics learned. This disparity of learning most likely exists between mathematics and any of its isomorphic applications. Furthermore, through the course of education, students develop an expectation that mathematical concepts can effectively and appropriately be applied to other domains such as physics, chemistry, economics, to name just a few. It is doubtable that student have as strong expectations that scientific strategies can be used to solve purely mathematical problems.

The purpose of this study was to investigate transfer across two isomorphic domains: one that used a set of abstract symbols, and another that used concrete perceptually-rich objects. To eliminate potential confounds stemming from prior knowledge, both domains were artificially constructed to be algebraic Abelian groups of order three. In other words, each is isomorphic to the integers under addition modulo three. Therefore, both domains included three classes of entities and a set of specific transformation rules described in Figure 1. The first, more abstract, domain (hereafter "Mathematics") was presented to the participants as a symbolic language in which three types of symbols, denoted as ♦, ❄, and ★, combine to yield a resulting symbol. The combination of symbols is expressed as written statements such as ★, ❄ → ♦. The second, more concrete, domain (hereafter "Science") involved interactions between three-dimensional objects from three classes. The objects dynamically interact to form a resulting object. The appearance of the objects

and interactions was designed to be dissimilar to any particular science.

The goal of the reported experiment was to investigate transfer across the two isomorphic artificial domains. Transfer was measured by comparing average test scores on a given domain as a function of prior learning of another domain.

Method

Participants

Participants in the experiment were undergraduate students from Ohio State University who received partial credit for an introductory psychology course. Thirty students participated in the experiment. Fifteen participants were in the math-then-science condition and fifteen were in the science-then-math condition. Information was presented to individual participants via computer.

Materials and Design

Materials included two sets of entities (i.e., abstract meaningless symbols and concrete, perceptually-rich objects), and a set of transformations rules (see Table 1).

Table 1. Example of stimuli and transformation rules across the two domains.

	Mathematics	Science
Elements	♦ ❄ ★	
Associativity	For any elements *x*, *y*, *z*,: ((*x* , *y*) , *z*) is equivalent to (*x* , (*y* , *z*))	
Commutativity	For any elements *x*, *y*: *x* , *y* is equivalent to *y* , *x*	
Identity	There is an element, I, such that for any element, *x*: *x* , I is equivalent to *x*	
Inverses	For any element, *x*, there exists another element, *y*, such that: *x*, *y* is equivalent to I	
Specific Rules:	♦ is the identity	 is the identity
	❄ , ❄ → ★ ★ , ★ → ❄	Operands / Result    

Information about each domain was given as a computer presentation. The training in both domains was essentially isomorphic. The rules of the domain, namely commutativity, associativity, and the rules governing specific elements, were explicitly stated.

The experiment included four phases presented over one hour: (1) Training in domain X, (2) Test in domain X, (3) Training in domain Y, (4) Test in domain Y, with participants randomly assign to a particular order of learning (i.e., math-then-science or science-then-math).

Training included introduction of transformation rules, followed by questions with feedback. Several detailed examples were given. Testing consisted of twenty multiple choice questions designed to measure recall of the given rules and deeper conceptual understanding of the system. For both domains, the test questions were completely isomorphic and were presented in the same order.

The presentation of the two domains differed by storyline. The artificial mathematics was presented as a symbolic language discovered on an archaeological search. Symbols of different categories combine to yield a resulting symbol. The artificial science was explained to be a phenomenon observed on a planet outside of our solar system. Objects from different classes of shapes interact to form a resulting shape. The presentation of the artificial science included movie clips demonstrating the interactions. Two or more objects move toward each other. When they come in contact, an interaction occurs and results in a predictable object.

Each subject was randomly assigned to one of two orders: math-then-science (M→Sc) or science-then-math (Sc→M). Participants in the first group received training and testing in the artificial mathematics immediately followed by training and testing in the artificial science. Subjects in the second group received training and testing first in the science and then in the mathematics. Following training and testing in both domains, a brief interview was conducted.

Students' scores on the mathematics and science tests were recorded. They were also asked to rate the similarity of the two domains on a scale from one to five. A rating of one indicated that the domains are completely different and a rating of five indicated that the domains are structurally identical with different representations of the objects.

Test scores for mathematics and science were compared across the two conditions, math-then-science and science-then-math. Transfer due to mathematics first was taken to be the difference in the average science score for M→Sc and the average science score for Sc→M. In other words, transfer is the improvement in science score due to having

previously learned the mathematics. Similarly, transfer due to science was taken to be the difference in average mathematics score for Sc→M and the average mathematics score for M→Sc.

Procedure

All training and testing was presented on a computer screen. Participants were tested in a quiet room in a lab by a female experimenter. They proceeded through training and testing at their own pace, and their responses were recorded by the researcher. After training and testing in both domains, a brief interview was conducted.

Results and Discussion

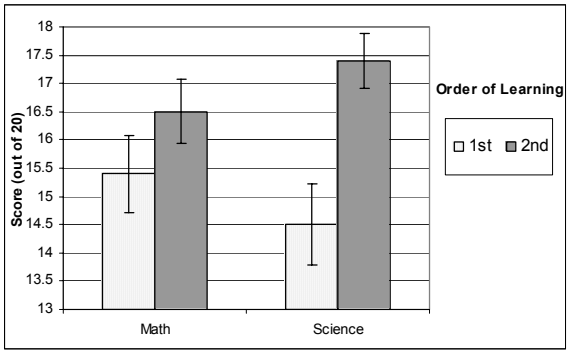
Students were able to learn the artificial mathematics and the artificial science. With the exception of one student, all test scores were significantly above chance (i.e., 7/20) in both math-then-science and science-then-math conditions. The one student who had a science score of 9/20 was removed from the data analysis. This score is not significantly different than chance and is also greater than two standard deviations from the mean (mean = 16.8, standard deviation = 2.8).

All students indicated that they noticed similarities between the two topics. Under both conditions the students rated the domains as highly similar. On a similarity scale from one to five, the mean rating given by math-then-science participants was 4.4 (*SD* = .65). The mean rating given by science-then-math students was 4.6 (*SD* = .63).

The data on transfer across the domains are presented in Figure 1. These data were subjected to a 2 (Domain: Math vs. Science) by 2 (Order: Learned First vs. Learned Second) mixed ANOVA with Domain as a repeated measure. The analysis revealed a significant Domain by Order interaction, $F(1, 27) = 24.15$, $p < .0001$. At the same time, none of the main effects was significant, both ps > .28.

Planned comparisons indicated that there was a significant difference in performance as a function of learning order. Students in the math-then-science condition performed significantly better on the science test than students in the science-then-math condition, independent-samples $t(24) = 3.26$, $p<.01$. However, there was no significant difference in mathematics scores across conditions, $p > .229$.

Figure 1. Mean test scores for mathematics and science shown as first and second domain studied.



The higher average science score for the M→ Sc group suggests that their prior knowledge of the mathematics improved their learning of the science. Therefore, significant transfer was found from the abstract symbolic domain to the concrete, but the reverse was not found.

In order to understand why the symbolic representation promoted transfer, while the concrete representation did not, it is necessary to take a closer look at the process of transfer. Not only does transfer require recognition and mapping of analogous relational structure from a source domain to a target domain, the elements of the source domain need to act as symbols. In other words, the objects or aspects of the source domain need to act as placeholders that can refer to something else, namely the objects of the target domain. Concrete representations are perceptually rich and consequently engage the perceptual system. Perceptually rich representations can easily convey associated properties and overall similarity (Goldstone & Barsalou, 1998). However, the specific characteristics of objects or elements are often irrelevant to concepts. Maintaining dissociation between the relational structure and the characteristics of the given elements is often crucial to accurate analogical reasoning. The salience of surface attributes often misleads students in the course of problem solving by distracting them from the underlying structure (Ben-Zeev & Star, 2001; Gholson, Smither, Buhrman, Duncan, & Pierce, 1997; Holyoak, Junn, & Billman, 1984; Holyoak & Koh, 1987; Schoenfeld & Herrmann, 1982). Perceptual objects convey affordances that may be helpful, but may also be irrelevant to the underlying concepts.

However, concrete representations may be beneficial under some conditions. For example Goldstone and colleagues (Goldstone, Son, & Patton, under review) have argued that maximum transfer occurs through "concreteness fading" where concrete representations progressively become idealized.

The conclusions of their study were that transfer is promoted through multiple representations. Furthermore, the authors conclude that while idealized displays promote internal representation not deeply embedded in a single domain, concrete displays have the advantage of a strong intuitive link between the real world and the modeled world. In other words, concepts can get a partial free ride from familiar concrete instances. However, in the course of learning mathematics and science, there are many concepts for which obvious concrete models may not exist. In the absence of a familiar concrete model on which concepts can freeload, does an artificially constructed concrete representation have benefits over a symbolic representation?

This notion of concepts getting a free ride from concrete representations as suggested by Goldstone and his colleagues is certainly appealing and is intuitively very reasonable. From a pedagogical perspective, there seems to be definite merit in concreteness fading provided that instructors do not allow learning to become deeply embedded in the concrete example. For students, the concept may become the concrete model and not the abstraction necessary for true understanding and transfer. Furthermore, many concepts may not have obvious and familiar representation in the real world. Mathematical concepts, by their very nature, are not bound to concrete contexts. Their transfer depends on attending to their relational structure and not salient surface features of a particular instance.

The goal of this experiment was to take a closer look at the merits of abstraction and concreteness for transfer. The familiarity or intuitive link between the concrete model and the concept were intentionally removed. No significant transfer from the concrete to the abstract was found, while significant transfer from the abstract to the concrete was exhibited.

Concrete representations may be difficult to treat as symbols in novel, complex concepts. Perceptually rich representations convey more information than leaner representations. As the degree of richness increases, it likely becomes more difficult to recognize the representation as an object itself as well as a reference to its intended referent. Successful transfer requires the elements of the source domain to be treated as symbols. Concrete representations engage the perceptual system. Rich percepts convey much information, a large portion of which is unrelated to a task in question. When that information correlates with the conceptual structure, learning may be facilitated. However, when the attributes are irrelevant to the concept, learners may not see the relevant analogy. Even when the analogy is perceived, it is difficult for rich percepts to be used as symbols, as demonstrated in this experiment. Participants in both the math-then-science and the science-then-math conditions recognized similarities between the domains. However, only the students who learned the symbolic mathematics prior to the concrete science were able to transfer information

Acknowledgments

This research is supported by a grant from the National Science Foundation (REC # 0208103) to Vladimir M. Sloutsky

References

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 153-166.

Ben-Zeev, T., & Star, J. R. (2001). Spurious correlations in mathematical thinking. *Cognition and Instruction, 19*, 253-275.

Cobb, P., Yackel, E., & Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. *Journal of Research in Mathematics Education, 21*, 2-33.

DeLoache, J. S. (1987). Rapid change in symbolic functioning of very young children: Understanding of pictures and models. *Child Development, 62*, 736-752.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child Development, 71*, 329-338.

Dienes, Z. P. (1960). *Building up mathematics.* London: Hutchinson Educational Ltd.

Fuson, K. C. & Briars, D. J. (1990). Using base-ten blocks learning/teaching approach for first and second grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *21*, 180-206.

Gholson, B., Smither, D., Buhrman, A., Duncan, M. K., & Pierce, K. A. (1997). Children's development of analogical problem-solving skill. In L. D. English (Ed.) *Mathematical reasoning: Analogies, metaphors, and images* (pp. 149-189). Mahwah, NJ: Erlbaum.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12,* 306-355.

Goldstone, R. L. & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65*, 231-262.

Goldstone, R. L. & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology, 46(4)*, 414-466.

Goldstone, R. L., Son, J., & Patton, Z. (2003). The transfer of scientific principles using concrete and idealized simulations. Unpublished manuscript.

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development, 55,* 2042-2055.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, *15*, 332-340.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *13*, 629-639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*, 456-468.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 484-494.

Representational Shifts in a Multiple-Cue Judgment Task with Continuous Cues

Linnea Karlsson (linnea.karlsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87 Umeå, Sweden

Peter Juslin (peter.juslin@psyk.uu.se)
Department of Psychology, Uppsala University
SE-751 42 Uppsala, Sweden

Henrik Olsson (henrik.olsson@psyk.uu.se)
Department of Psychology, Uppsala University
SE-751 42 Uppsala, Sweden

Abstract

Research on multiple cue judgment with continuous cues and a continuous criterion has been dominated by statistical modeling of the cue utilization with linear multiple regression. In this study we apply two cognitive process models to investigate the relative contributions of explicit abstraction of the cue-criterion relations and memory for concrete exemplars in a multiple-cue judgment task. The task was an extension of a previous task with binary cues (P. Juslin, H. Olsson., A-C. Olsson, 2003) and involved multiple continuous cues that either combined by addition or multiplication. As predicted by the process model $\sum$ (P. Juslin, L. Karlsson, & H. Olsson, manuscript) explicit abstraction of cue-criterion relations were induced in the additive task, while exemplar memory was induced in the multiplicative task.

Introduction

Multiple-cue judgment research has traditionally been concerned with statistical modeling of judgment data. Rather exquisite regression models have been developed that describe multiple-cue judgment as *a*) well fitted by a linear additive model; *b*) only taking a few cues into account; *c*) hard to report on subjectively; *d*) characterized by cue weightings that differ greatly between individuals; and *e*) plagued by considerable inconsistency in the weighting of the cues (see Brehmer, 1994; Cooksey, 1996; Hammond & Stewart, 2001).

In the light of the cognitive revolution it might seem puzzling that this field of research has not benefited from the growth of cognitive modeling as a means to track the underlying cognitive representation and process of judgment, a growth seen in related fields like categorization learning (but see for example Bott & Heit, 2004; Busemeyer, Byun, DeLosh, & McDaniel, 1997; or DeLosh, Busemeyer & McDaniel, 1997, single-cue learning). Categorization – which is in many ways similar to multiple-cue judgment (see Juslin, Olsson, & Olsson, 2003) – has invited extensive investigation of the cognitive representations and processes that underlie behavior. A plethora of models, ranging from an emphasis on how abstract rules or prototypes guide category decisions to a domination of memory for category exemplars are thus available in cognitive science today. In this study we apply the methods of cognitive modeling to a typical multiple-cue judgment task. By connecting research on cognitive science to judgment and decision making research, we can gain an understanding of what cognitive representations and processes guide the judgments, and how this is manifested in the results of the traditional statistical modeling (e.g., Cooksey, 1996).

Arguably, it is not mere coincidence that linear, additive models fit multiple cue judgment data well and that categorization is often well captured by exemplar models that entail a linear additive combination of retrieved exemplars (Juslin, Karlsson, & Olsson, manuscript). Imagine how you sequentially consider and weigh the pros and cons of different aspects of a car before you purchase it (its looks, reliability, etc). You may weigh them differently but positive qualities *add* to and negative qualities *subtract* from your overall opinion. Likewise, you may sequentially consider exemplars of similar cars that you are aware of: similar cars (e.g., same model) that have worked properly *add* to the appeal of the car and cars that that have been frustrating *subtract* from it.

We have proposed a general process model, $\sum$, that captures the essentials of multiple-cue judgment, both when it is driven by consideration of cue-criterion relations and exemplar retrieval (Juslin et al., manuscript). The assumptions in $\sum$ are that our controlled and explicit thought processes have an architectural constraint enhancing sequential, real-time consideration of multiple pieces of evidence (cues or exemplars). The process involves successive adjustment of an estimate, a process structurally compatible with linear, additive cue integration (Einhorn, Kleinmuntz, & Kleinmuntz, 1979) and exemplar models (Nosofsky & Johansen, 2000).

The key assumption is that, in effect, all integration of information involves addition (or subtraction). This hypothesis suggests that explicit and controlled thought processes are apt at performing cue-integration only in

tasks where the cue-criterion relations in the task indeed combine by addition. By contrast, a task that involves non-linear or multiplicative cue combination requires capitalization on exemplar memory (Medin & Schaffer, 1978; Nosofsky & Johansen, 2000). Exemplar memory involves no strong computational commitments to particular task structures. With a division of labor between distinct representations we are better equipped to adapt to different task structures. We propose that the judgment process adapts to specific task environments and predict that in a multiple-cue judgment task with continuous cues we can induce a shift between qualitatively distinct processes by manipulating the structural properties of the environment: additive cue combination should promote cue abstraction and multiplicative cue combination should promote exemplar memory.

Judgment Task and Cognitive Models

The judgment task involves judgment of a continuous criterion based on continuous cues. The task concerns judgments of the effectiveness of different species of herbs as medical treatments to a lethal virus. The effectiveness is measured as the *maximal amount of a chemical substance* (mg) that can be extracted from the species. The species have four continuous dimensions (C_1, C_2, C_3, C_4), and each cue dimension can take a value between 0 and 10. The judgment of effectiveness requires inference from these dimensions, which are presented as verbal statements (e.g., weeks of bloom per year, geographic place of growth).

The tasks involve two manipulations. First, there is one condition in which all cues are related to the criteria positively and linearly, and one condition in which two cues are positively and two cues are negatively linearly related to the criterion. This manipulation makes the task a matter of function learning. Second, there is a manipulation of whether the effects of the four cues on the criterion combine by *addition* or *multiplication*.

In the additive condition the criterion is a linear, additive function of the continuous cues:

$$c = 500 + 4 \cdot C_1 + 3 \cdot C_2 + 2 \cdot C_3 + 1 \cdot C_4 + \varepsilon . \qquad (1)$$

C_1 is the most important cue with *coefficient* 4 (i.e., a relative weight .4), C_2 is the second to most important with coefficient 3, and so forth. The cues are uncorrelated. ε is a normally and independently distributed random error with a standard deviation that produces a multiple correlation R between cues and criterion of .9 (i.e., defining the ecological validity of the cues).

In the multiplicative condition the criterion c is a multiplicative function of the four cues:

$$c = 509.05 + 0.54545 \cdot e^{(C_1 \cdot 4 + C_2 \cdot 3 + C_3 \cdot 2 + C_4 \cdot 1)/18} + \varepsilon , \qquad (2)$$

with the same coefficients as in the additive task (Eq. 1). The effectiveness varies between 500 and 600 mg of chemical substance in the additive task and 509 to 650 mg in the multiplicative task. However, the training ranges are hold equal for the two conditions. The range of cue values observed in the two tasks is therefore the same. Moreover, the criterion in the multiplicative condition is an exponential function of the criterion presented in the additive condition.

We use two structural models to derive predictions, a cue-abstraction (CAM) and an exemplar model (EBM). $\sum$ implies that in the additive task CAM should be the correct structural description of the process, whereas in the multiplicative condition EBM should be the appropriate description[1]. The CAM assumes that participants abstract explicit cue-criterion relations in training that are mentally integrated at the time of judgment. When presented with a probe the participants retrieve rules connecting cues to criterion (e.g., "More weeks in bloom gives more effectiveness"). The rules specify the sign and importance of each cue with a cue weight. For example, after training the rule for C_1 may specify that high C_1 goes with an increase in the criterion.

The CAM implies that participants compute an estimate of the criterion c based on sequential consideration of cues. For each cue, the estimate of c is adjusted according to the cue weight ω_{iA} (i=1…4). The final estimate $\hat{c}_{CA}$ is a linear additive function of the cues C_i,

$$\hat{c}_{CA} = k + \sum_{i=1}^{4} \omega_{iA} \cdot C_i , \qquad (3)$$

where $k = 500 + .5 \cdot (100 - 10 \cdot \sum \omega_{iA})$. If ω_{1A}=4, ω_{2A}=3, ω_{3A}=2, and ω_{4A}=1, Eq's 1 and 3 are identical and the model produce perfect judgments. The intercept k constrains the function relating judgments to criteria to be regressive around the midpoint (550) of the interval [500, 600] (Juslin et al., manuscript).

Although ruled out by the predictions of $\sum$, we also consider the possibility that participants have correctly abstracted the multiplicative cue-criterion relations by fitting a multiplicative cue-abstraction model to the data:

$$c = 509.05 + 0.54545 \cdot e^{(\sum_{i=1}^{4} \omega_{iM} \cdot C_i)/18} \qquad (4)$$

where ω_{1M} are the best fitting subjective cue weights in the multiplicative cue abstraction model.

EBM is commonly applied to classification, but here we apply it to a continuous criterion. EBM implies that participants make judgments by retrieving similar exemplars (herb species) from memory. When the exemplar model is applied to judgments of a continuous criterion variable, the estimate $\hat{c}_E$ of the criterion c is a weighted average of the criteria c_j stored for the J exemplars, where the similarities $S(p,x_j)$ are the weights:

$$\hat{c}_E = \frac{\sum_{j=1}^{J} S(p, x_j) \cdot c_j}{\sum_{j=1}^{J} S(p, x_j)} . \qquad (5)$$

[1] $\sum$ is a model of the real-time process of judgment that becomes structurally identical with a CAM when the representations fed to the process are abstracted cues and structurally identical to an EBM when the process is fed by exemplars. The structural description refers to the relationships between stimulus features and the response (Juslin et al., manuscript).

p is the probe to be judged, x_j is exemplar j (j= 1…J), and $S(p,x_j)$ is the similiarity between probe p and exemplar x_j. Eq. 5 is the *generalized context model* (GCM: Nosofsky, 1984; 1986), which generalizes the original version of the context model (Medin & Schaffer, 1978). The similarity $S(p,x_j)$ between exemplars is found by transforming the distance between them. The distance between a probe p and an exemplar j is,

$$d_{pj} = h\left[\sum_{M=1}^{M} w_m \left| x_{pm} - x_{jm} \right|\right], \qquad (6)$$

where x_{pm} and x_{jm}, respectively, are the values of the probe and an exemplar on cue dimension m, the parameters w_m are the attention weights associated with cue dimension m, and h is a sensitivity parameter that reflects overall discriminability in the psychological space (the sensitivity parameter is usually denoted c, but we changed that to avoid confusion with the criterion c). Attentional weights vary between 0 and 1 and are constrained to sum to 1. The similarity $S(p,x_j)$ between a probe p and an exemplar j is assumed to be a nonlinearly decreasing function of their distance (d_{pj}),

$$S\left(p, x_j\right) = e^{-d_{pj}}. \qquad (7)$$

In the experiment, herb species with a criterion above 590 and below 510 are not included in the training phase. This makes it possible to distinguish between the models as they provide different predictions (Figure 1). In the training phase, all exemplars have effectiveness between 510 and 590. If participants have estimated the correct cue weight for each cue they should compute the most extreme judgments for the extreme exemplars that are left out in the training phase. More specifically, whenever participants have correctly identified the sign of each cue (i.e., whether it increases or decreases the criterion) they should make more extreme judgments for the exemplars with all cues at their maximum and the exemplars with all cues at their minimum, as illustrated on the left-side of Figure 1. By contrast, the exemplar model computes a weighted average of the criteria between 510 and 590 stored with the exemplars and this can never produce a value outside of this observed range (Erickson & Kruschke, 1998; but see DeLosh et al., 1997). Moreover, because of the nonlinear similarity function of the GCM the most extreme judgments tend to be made for the second to most extreme exemplars. For these exemplars the judgment is dominated by retrieval of the identical stored exemplars and these identical exemplars are the most extreme that were encountered in the training phase. These predictions are illustrated on the right side of Figure 1.

For new exemplars in the mid range of the criterion cue abstraction suggests no systematic difference between new exemplars and old exemplars matched on the criterion: the cognitive process is the same regardless of whether a specific exemplar has been encountered before or not. The exemplar model predicts more precise judgments for the old exemplars because for these exemplars the participants can benefit from previous identical exemplars with the correct criterion c.

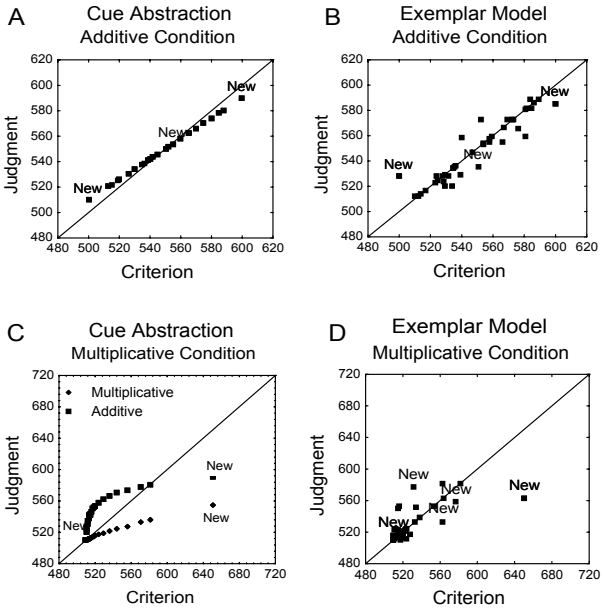


Figure 1: Predictions by cue-abstraction model (CAM) and exemplar model (EBM) in additive and multiplicative task environments. Panel A: CAM in additive task environment (with slightly regressive weights, 3.2, 2.4, 1.6, & .8). Panel B: EBM in additive task environment (s = .25 and h = 10). Panel C: Additive [CAM(A)] and multiplicative [CAM(M)] cue-abstraction models in multiplicative task environment (with weights, 3.2, 2.4, 1.6, & .8). Panel D: EBM in multiplicative task environment (s = .25 and h = 10). The choice of values for the parameters is arbitrary and only used for illustrative purposes.

The Experiment

In the experiment we manipulated whether participants were confronted with a task that involved additive or multiplicative cue combination. For the reasons outlined in the introduction, we predicted that the additive task (Eq.1) should promote explicit cue abstraction with additive cue integration (Eq. 3). A multiplicative task (Eq. 2) should cause a shift to a qualitatively different process, that is, to exemplar memory (Eq. 5).

The sign of the linear relations between cues and criterion was also manipulated. For half of the participants all four cues were positively related to the criterion and for half of the participants two cues were positively and two cues were negatively related to the criterion. In line with the assumptions of ∑, we predict that in an additive task, whether cue directions are negative or positive should not affect the ability to perform cue abstraction. In a multiplicative task, both with homogeneous

and heterogeneous cue directions, exemplar memory is predicted to prevail over cue abstraction.

Method

Participants

Thirty two undergraduate students volunteered, receiving a payment of 60-99 SKr, depending on their performance. Twenty participants were male and 12 were female, all in the age between 20 and 32.

Materials and Procedure

The experiment consisted of a training phase and a test phase. In the training phase, the participant learned to judge the effectiveness of each species of the herb by means of outcome feedback. The effectiveness was measured as the amount (mg) of the fictitious chemical substance *Ranulin*. In the training phase, the effectiveness varied between 510 and 590 mg. The species were shown as four written propositions on a computer screen. At each trial in the training phase, the participant was to answer the question "How many milligrams of Ranulin does this specie contain?". After giving a response they received the correct answer: "This specie contain 540 milligrams of Ranulin". The four dimensions were: number of weeks in bloom, the optimal amount of iron in the ground, the degrees of latitude where it does well, and the amount of water it emits per leaf area. Each dimension varied "pseudo-continuously" in 11 equidistant steps that ranged between 0 and 10, yielding a total of 11^4 different exemplars. In the training phase, a random sample of 300 exemplars was drawn from this distribution and shown to the participant. A pause of two minutes was given to the participant after the first 150 trials.

In the test phase, participants were to judge the effectiveness of the species of the herbs but now *without* outcome feed-back. In the test phase, new exemplars were included. The test phase consisted of 44 judgments of *a*) 20 randomly chosen old exemplars shown in training, *b*) 20 randomly chosen new exemplars, drawn from the training distribution and *c)* 4 extreme exemplars, with criterion values outside the training range (eg. the exemplars with cue values [0,0,0,0] and [10,10,10,10])

In the condition with heterogeneous cue directions, for half of the participants, negative sign was assigned to the cues with objective weight 4 and 2, and for half of the participants to the cues with weights 3 and 1.

Dependent Measures

The measure of performance is *Root Mean Square Error* (*RMSE*) of judgment (i.e., between judgment and criterion). Measures of model fit are *the coefficient of determination* (r^2) and *Root Mean Square Deviation* (*RMSD*) between predictions and data from the test phase.

Results

A two-way ANOVA with environment (additive vs. multiplicative) and cue directions (homogeneous vs. heterogeneous) as between-subject factors shows two main effects on RMSE (Table 1), but no interaction. In the additive condition performance is significantly better ($F(1.30) = 20.36$; $MSE = 32.36$; $p = 0.000$). Also, when the cue directions are homogeneous RMSE is lower compared to when the cue directions are heterogeneous ($F(1.30) = 20.36$; $MSE = 6.37$; $p = 0.018$)

Table 1: Judgment performance in the experiment as measured by the Root Mean Square Error (RMSE) between judgment and criterion.

		Condition		
Cue directions	Index	Add.	Mult.	Mean
Homogeneous	RMSE	11.69	21.23	*16.46*
Heterogeneous	RMSE	17.21	25.90	*21.56*
Mean	RMSE	*14.45*	*23.56*	

Mean judgments are shown in Figure 2. In the additive homogeneous condition the judgments are a linear function of the criterion and no extra- or interpolation effects are visible. The best fitting regression lines for old and new judgments coincides.

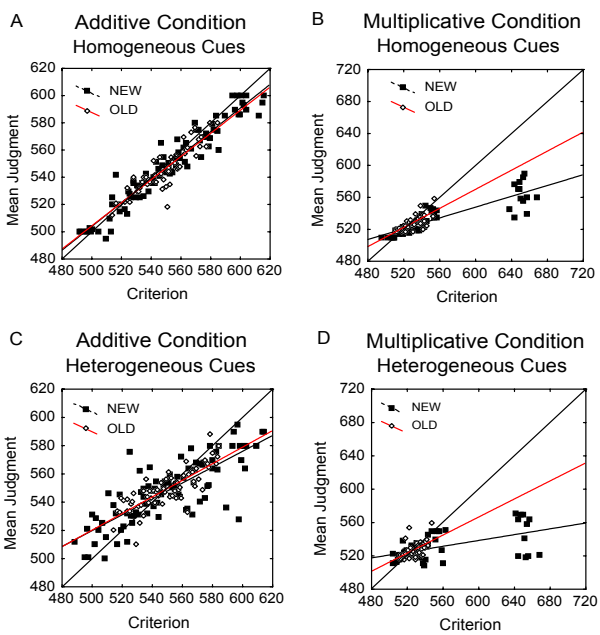


Figure 2: Mean judgments for the different conditions. Panel A: additive, homogeneous. Panel B: multiplicative, homogeneous. Panel C: additive, heterogeneous. Panel D: multiplicative, heterogeneous. Best-fitting regression lines are based on a) the old exemplars seen in training or b) the new exemplars introduced at test.

In the multiplicative homogeneous condition the judgments clearly deviate both from the identity line and the best fitting regression line based on old exemplars. Although the judgments are a positive function of the criteria in the training range there is evidence for an inability to extrapolate. The judgments are not extrapolated beyond the range of training.

In the additive heterogeneous condition (Figure 2C) the judgments are still close to the optimal judgment line, although there are signs of extra- and interpolation effects. In the multiplicative heterogeneous condition (Panel D) the mean judgments are a positive function of the criterion, but the inability to extrapolate is obvious.

Quantitative model predictions were obtained by fitting the models in the introduction (Eq. 3, 4 & 5) to the latter half of the *training phase* with Mean Square Error between predictions and data as the error function (Juslin et al., 2003; manuscript).

Table 2: Model fit: Root Mean Square Deviations (RMSD) and r^2 for the additive cue-abstraction model (CAM(A)) the exemplar model (EBM) and the multiplicative cue-abstraction model (CAM(M)) in the four conditions.

	CAM(A)		EBM		CAM(M)	
Cond.	r^2	RMSD	r^2	RMSD	r^2	RMSD
Add:						
Homogen.	.77	10.83	.75	13.73	-	-
Heterogen.	.53	12.67	.51	13.40	-	-
Mult:						
Homogen.	.21	28.26	.74	8.50	.70	12.20
Heterogen.	.18	48.62	.34	19.65	.32	18.97

The models were thus fitted to data from the training phase and applied with these parameters to the wider range of herb species in the *test phase*. This implies cross-validation for exemplars presented in training and genuine predictions for new exemplars. To capture individual differences, the models were applied to individual data. Table 2 shows the mean fit for the three models. In the additive condition cue abstraction is the overall dominant model, regardless of the cue directions. In the multiplicative condition exemplar memory describes the data best with regard to r^2 and the multiplicative cue-abstraction model yields a smaller mean *RMSD*. The rather low fit of all three models in the heterogeneous conditions may be explained by larger noise in these data since this task is presumably more difficult than the homogeneous task. Figure 3 shows the proportion of participants best accounted for by each model in terms of *RMSD*. In the additive homogeneous condition most of the participants are accounted for by the cue-abstraction model. In the multiplicative homogeneous condition the reverse is true, namely that the exemplar model produces the best explanation. In the additive heterogeneous condition the proportion of participants explained by the cue abstraction model decreases. In the multiplicative heterogeneous condition, as hypothesized the exemplar-based model continues to provide the best explanation of data for most of the participants. The multiplicative cue-abstraction model describes some of the participants in both the homogeneous and the heterogeneous multiplicative tasks.

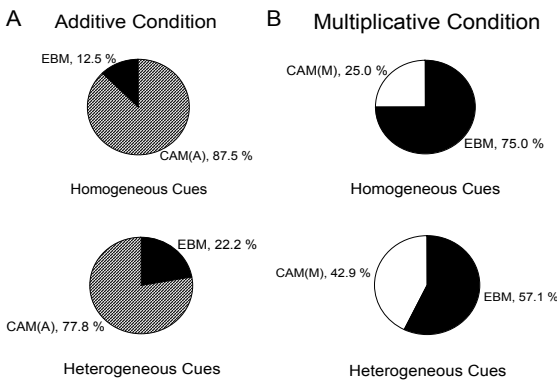


Figure 3: The proportion of participants accounted for by any of the three models in the additive and the multiplicative conditions in terms of *RMSD*. Panel A: additive condition. Panel B: multiplicative condition.

Discussion

The results reported in this paper support the assumptions made by ∑ that multiple-cue judgment processes conceal an effective division of labor between qualitatively distinct cognitive processes (Juslin et al., manuscript). Cognitive modeling supports the hypothesis that in a multiple-cue judgment task where the cues combine by addition, ∑ is fed with representations in form of abstracted knowledge of the relations between cues and criterion. On the other hand, in an environment where the cues relate to the criteria by a multiplicative function we seem to be equipped with no means to explicitly abstract the underlying structure. In such tasks, people seem to resort to the back-up process of exemplar-memory.

The fact that exemplar-memory plays part also in additive tasks is not a coincidence, since both processes allows accurate performance in training. That the multiplicative cue-abstraction model provide an explanation for some of the participants in the multiplicative task is more surprising. This is probably an effect both of large noise in data and of its high correlations to the exem-

plar model. Figure 2 B & C yields no evidence for successful extrapolation beyond the range of training.

The bad performance in the multiplicative heterogeneous condition, together with the low fit of the models makes it unfair to draw conclusions regarding what cognitive process that has dominated the judgments in this condition. What makes this task difficult to learn? A tentative hypothesis would be that in training there is a bias towards the abstraction of specific rules (eg. *rule bias;* see for example Ashby et al., 1998; Juslin et al., 2003; manuscript). Presumably, a period of extensive hypothesis testing is therefore taking place at beginning of training. However, the multiplicative heterogeneous task may be inductive of more extensive hypothesis-testing procedures. The back-up of exemplar memory is thus postponed, and thereby learning may be impaired.

An interesting approach to the interpretation of the data would be to consider how an exemplar-model augmented with linear extrapolation would account for the results (see *EXAM*; DeLosh et al., 1997; Busemeyer et al., 1997, for results on single-cue learning). *EXAM* suggests that, although learning has been in the form of exemplar-memory, abstraction of cue-criterion relations is possible at test. When encountered with a new exemplar at test, familiar exemplars and their stored criterion are retrieved from memory. An extrapolated judgment for the new exemplar is then made possible through linear regression based on the old exemplars. How this model explains the data reported in this paper remains to be tested, although a first qualitative evaluation of the data in Figure 2 can be made. Given the data in the additive condition, *EXAM* is likely to produce the same fit as the cue-abstraction model. In the multiplicative condition *EXAM* would predict no difference between the regression made on old exemplars and the regression made on new exemplars. This difference is however apparent in the data in Figure 2 (Panel B & C) and thus suggests the refutation of *EXAM* in favor of EBM.

The main interpretation to be drawn from the results reported in this paper is that the human judge, under the constraints imposed by $\sum$, adapt to different task structures by means of representational shifts. This highlights how the task is a powerful predictor of cognitive process in human multiple-cue judgment.

Acknowledgments

Bank of Sweden Tercentenary Foundation supported this research.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105,* 442-481.

Bott, L. & Heit, E. (2004). Non-monotonic Extrapolation in Function Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30 (1)*, 38-50.

Brehmer, B. (1994). *The psychology of linear judgment models. Acta Psychologica, 87.* 137-154.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning Functional Relations Based on Experience With Input-Output Pairs by Human and Artificial Neural Networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge, Concepts and Categories*. London: UCL Press.

Cooksey, R. W. (1996). *Judgment analysis. Theory, methods and applications.* San Diego. Academic Press, Inc.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 968-986.

Einhorn, J. H., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Regression models and process tracing analysis. *Psychological Review, 86*, 465-485.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107-140.

Hammond, K. R., & Stewart, T. R. (Eds.) (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford University Press: Oxford.

Juslin, P., Karlsson, L. & Olsson, H. (manuscript) *Exemplar Memory in Multiple-Cue Judgment: A Division of Labor Hypothesis.*

Juslin, P., Olsson, H., & Olsson, A-C. (2003) Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General132 (1),* 133-156.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning, *Psychological Review, 85*, 207-238.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 7,* 375-402.

Smith, J. D, & Minda, J. P (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26, 3-27.*

Activation of Non-Target Language Phonology During Bilingual Visual Word Recognition: Evidence from Eye-Tracking

Margarita Kaushanskaya (m-kaushanskaya@northwestern.edu)
Northwestern University
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, Il 60208

Viorica Marian (v-marian@northwestern.edu)
Northwestern University
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, Il 60208

Abstract

Russian-English bilingual and English monolingual participants were tested on the Picture-Word Interference task modified for use with an eye-tracker. Distractor words were 1) non-words in English, but viable phonological words in Russian, 2) control bigram matched non-word stimuli, and 3) English translations of the Russian words. Russian-English bilinguals looked at the phonological Russian words more than monolingual participants, and took longer to name pictures accompanied by these stimuli than did monolingual participants. Proportion of eye-movements and reaction times to the other two types of distractor stimuli did not differ for the two groups. These results suggest that phonology of the non-target language is activated automatically during visual word recognition in the target language, even for written stimuli that do not carry orthographic information for the non-target language.

Introduction

The task of reading is hard enough when the reader reads in one language alone. In the case of a bilingual reader, the picture is even more complex: Not only does a bilingual reader need to process the written information in one language, but he or she may have to contend with information from his or her other language that also becomes activated. The activation of the non-target language during reading in the target language is implemented in the Bilingual Interactive Activation (BIA+) model of visual word recognition (Dijkstra & Van Heuven, 1998; 2002). The BIA+ model is a localist connectionist model with elements from both the dual-route models of reading (e.g., Coltheart et al., 1993; Coltheart et al., 2001; Ziegler et al., 2000) and the connectionist models of reading. (e.g., Plaut et al., 1996; Seidenberg and McClelland, 1989; Van Orden and Goldinger, 1994). The BIA+ model proposes that lexical access of a visually presented word in a bilingual is non-selective, i.e., when a word is presented, information for that word, both orthographic and phonological, is activated for both of the bilingual's languages.

Support for such non-selective processing of written information in bilinguals (i.e., activation of the non-target language orthography and/or phonology when involved in a reading task requiring use of only the target language) has accumulated over the past three decades (e.g., De Groot, Delmaar, & Lupker, 2000; Nas, 1983; Van Heuven, 2000; Van Heuven, Dijkstra, & Grainger, 1998). Activation of non-target language phonological information has been reported for bilingual readers of languages with shared alphabets, such as English and French (Jared & Szucs, 2002), Dutch and English (Dijkstra, Grainger, & van Heuven, 1999), Spanish and Catalan (Costa, Miozzo, & Caramazza, 1999), and Dutch and French (Brysbaert, Van Dyck, & Van de Poel, 1999). Activation of phonological information for the non-target language has also been reported in the case of bilinguals who speak languages with entirely different alphabets, like Hebrew and English (Tzelgov et al., 1996). However, it is still largely unknown whether phonological information for the non-target language is activated when the letter string in the target language carries little resemblance to the non-target language orthography (but see Feldman & Turvey, 1983; Lukatela et al, 1978).

In this experiment, we used a modified Picture-Word Interference (PWI) task to test whether phonological information for Russian is automatically activated during processing of non-words in English. The stimuli were constructed to contain English-specific letters, but constitute viable phonological Russian words. Using head-mounted eye-tracking methodology, we examined Russian-English bilinguals' ability to control their eye-movements to distractor words in the PWI task when these words contained phonological, but not orthographic information for Russian. The proportion of eye-movements to the distractor word signified the degree to which Russian phonological information drew the participants' eye-movements, while the reaction times to naming the target picture stimulus signified the degree to which the Russian phonological word interfered with picture naming in English.

Eye-tracking technology has been used to explore parallel activation of the two languages known to a bilingual during auditory perception tasks (e.g., Marian & Spivey, 2003 a, b). Eye-tracking has also been widely used in research of reading (e.g., Reichle, Rayner, & Pollatsek, 1999; Starr & Rayner, 2001). Unlike automatic eye-movements observed

in response to spoken instructions, eye-movements in reading are thought to be under partial cognitive control. The E-Z Reader model of eye-movement control in reading (Reichle, 1998; Reichle et al., 1999) posits that prior to programming an eye-movement to a particular word, a familiarity check takes place, which indicates whether a word is likely to be recognized by a reader. During this familiarity check, a reader gains information on low-level properties of the word (Engbert, Longtin, & Kliegl, 2002), which causes partial activation of the lexicon (e.g., Deutsch et al., 2002; Starr & Rayner, 2001). Experimental evidence shows that readers obtain both orthographic information (Liu et al., 2002), and phonological information (Wong & Chen, 1999) from the word before it is fixated. Models of eye movement control during reading account only for monolingual reading and it is still largely unknown whether control of eye-movements during bilingual reading is accomplished in the same way.

Recent research with bilinguals suggests that phonological information for the non-target language is automatically activated when reading in the target language. However, phonological activation for the non-target language when reading in the target language has not yet been explored for bilingual speakers whose two languages have partially overlapping alphabets. For these speakers, letter strings with alphabet-specific symbols often contain phonologically meaningful information for the non-target language. Russian-English bilinguals are faced with exactly this type of alphabetic overlap (Figure 1). Given the properties of the Russian Cyrillic alphabet and the English Roman alphabet, it is possible to test whether phonology of the non-target language is activated when its orthography is only partially present in the target language letter string. Consider the word COBA, which is the Russian word for "owl." COBA can be transcribed using the Roman alphabet, SAVA, which includes letters specific to the English alphabet. Letter strings like SAVA constitute phonological, but not orthographic, representations of Russian words. When they are presented to the Russian-English bilinguals, only phonology, but not orthography, associated with corresponding Russian words, should be activated.

The activation of the non-target Russian language during picture naming in English was measured using a modified PWI task. The objective of the PWI task is to name pictures, while ignoring the words also present on the screen. The interference from the distractor word is thought to arise due to the automatic reading of the word, which then interferes with selection of the appropriate name for the picture at the level of the lexico-semantic system. Thus, reading during this task is largely automatic; in fact, it is counter-productive to the successful and fast completion of the task. Unlike a classic PWI task, where a written stimulus is presented inside a picture, we separated the written stimulus and picture presentation, such that a picture was in one quadrant of the computer screen, while a written stimulus was in another quadrant of the computer screen. The instructions to the participants were the same as in the regular PWI task: To ignore the word, and name the picture. In the regular PWI task, these instructions do not prevent reading of the words because the words are presented inside

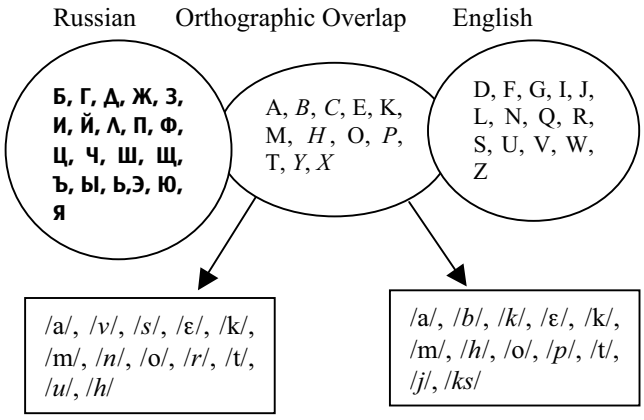


Figure 1: Overlapping symbols in orthographies of Russian and English, and the associated phonemes in each language.

the picture, and therefore, the participants necessarily look at them. In the modified PWI task, the word and the picture are in different locations on the screen so that the participant sees that there is a word, but does not need to look at it in order to recognize the picture. We tracked the participants' eye-movements while they were completing the task in order to determine whether particular words drew more eye-movements than others.

Based on the models of cognitive control of eye movements in reading (Reichle, 1998; Reichle et al., 1999), the proportion of eye-movements to distractor stimuli during the modified PWI task should be indicative of the degree to which participants were able to control their eye-movements to the written stimuli. The stimuli presented as distractors on the PWI task words were 1) phonological Russian words, 2) English translations of the Russian stimuli, and 3) control stimuli (non-words in both Russian and English that were controlled for English bigram frequencies to equal phonological Russian stimuli). We hypothesized that if phonology of Russian is automatically accessed even when the orthographic shape of the word is not Russian (i.e., contains English-specific letter symbols), Russian-English bilingual readers will make more eye-movements to the phonological Russian words than the monolingual English speakers. Consequently, we hypothesized that if the phonological information for Russian activates the lexico-semantic information for Russian, Russian-English bilinguals will have longer reaction times to the pictures accompanied by the phonological Russian words than monolingual English speakers.

Methods

Participants

Fifteen Russian-English bilingual speakers (Mean Age=24.27 years, SD=4.80) and 15 English monolingual speakers (Mean Age=21.00 years, SD=4.68) participated in the experiment. All Russian-English bilingual participants were born in the former Soviet Union, and immigrated to the United States at a mean age 14.56 years (SD=5.35). All bilingual participants filled out a Language Experience and

Bilingual Status Questionnaire (LEABS-Q) at the end of the experiment. LEABS-Q is a comprehensive questionnaire containing questions about language proficiency, modes and ages of language acquisition, current language usage, etc. (Marian, Blumenfeld, and Kaushanskaya, 2003). Self-reported proficiency measures were later determined based on the participants' answers.

Reading fluency and reading comprehension were assessed by administering a passage reading task in English to the monolingual participants, and in English and Russian to the bilingual participants. When tested in English, bilingual participants were found to read orally with similar speed, $t(28)=0.94$, $p=0.36$, have as many errors while reading, $t(28)=0.75$, $p=0.46$, and comprehend as much of the content, $t(28)=0.96$, $p=0.35$, as the monolingual participants. For bilingual speakers, fluency of reading, $t(14)=1.89$, $p=0.08$, and comprehension of content, $t(14)=0.38$, $p=0.71$ were comparable for Russian and English. However, bilingual participants were significantly faster when reading in English ($M=2.71$ words/sec, $SE=0.12$) than in Russian ($M=2.12$ words/sec, $SE=0.11$), $t(14)=4.44$, $p<0.01$.

Design

Two dependent variables were considered: reaction time, and proportion of eye-movements to the word. The experiment followed a 4x2 Mixed Design with two independent variables – condition (no word, phonological Russian word, non-word control stimulus, and English word) as a within-subjects variable, and group (bilingual vs. monolingual) as a between subjects variable. For the proportion of eye-movements data, condition variable included only three levels (phonological Russian word, non-word control stimulus, and English word).

Materials

Twenty-three target pictures of common concrete objects were selected from the IMSI MasterClips picture database; all pictures were transformed into black-and-white drawings of equal size using PhotoShop.

Twenty-three words that were semantically related to picture names, i.e. belonged to the same superordinate category, were selected. The 23 words were then translated into Russian to create stimuli that were phonological representations of Russian words, spelled using the English alphabet. For instance, for the picture of a duck, the distractor word selected was *chicken*, which translates to YTKA in Russian, and yields UTKA when spelled using the English alphabet.

Control stimuli for the phonological Russian stimuli were constructed by creating non-words comparable to Russian phonological words in length and bigram frequencies. Paired-samples t-tests confirmed that Russian phonological stimuli ($M=4355.02$, $SE=2244.50$) and phonological controls ($M=4263.35$, $SE=2362.53$) were similar in their bigram frequencies ($t(22)=0.36$, $p>0.05$).

For each picture, there were 4 conditions: (1) picture – no word, (2) picture – English semantic distractor, (3) picture – phonological Russian semantic distractor, and (4) picture – non-word control stimulus. The picture-no word condition was used as a baseline, to establish that eye-movements to words, if occurred, were due to the presence of the word in a particular location, and not to the location itself. For each condition, a panel divided into four quadrants was constructed – a picture was placed into the middle of one quadrant, and the word was placed into the middle of another quadrant. For each condition, a picture and all the words in the three conditions were placed in the same quadrants; the positions of pictures and words were counterbalanced across the four possible quadrants. To increase the time between target picture presentations, 16 filler picture stimuli were added to the experiment.

Apparatus

All stimuli were presented on a G5 Macintosh Monitor using SuperLab experimental software. Naming times were measured from the presentation of the picture to the onset of triggering the microphone response by the participant's voice. A headband-mounted ISCAN eyetracker was used to record participants' eye-movements during the PWI task. A scene camera, joined to the view of the tracked eye, provided an image of the participant's field of view. A second camera, which provided an image of the participant's left eye, allowed the ISCAN software to track the center of the pupil and the corneal reflection; gaze position was indicated by white crosshairs superimposed over the image generated by the scene camera. The output was recorded onto a digital mini-tape via a Cannon Digital Camera; it was later loaded into the FinalCut Editing software for frame-by-frame playback analysis.

Procedure

All participants were tested individually. Training for the Picture-Word Interference task was presented first. Each picture used in the PWI task was presented in the middle of the screen; the participant was instructed to name it into the microphone.

The Picture-Word Interference task was presented next. Prior to initializing the task, the calibration of the eye-tracking equipment was completed. To increase the sensitivity of equipment, calibration was done on 9 fixation points. The fixation values were then mapped onto the corresponding monitor locations; the fixation location was indicated by a white cross-hair that moved synchronically with the eyes. After successful calibration, the PWI task was initiated. Each participant was instructed to fixate on the cross that appeared prior to each picture stimulus; he/she was also instructed to name pictures into the microphone as fast and as accurately as possible, and ignore the text on the screen.

At the end of the experimental session, two proficiency measures were administered to each participant. The first was the reading ability measure: Each participant read a short passage in English into the microphone, and answered 8 multiple-choice questions about it afterwards. Each bilingual subject also read a short passage in Russian, and answered 8 multiple-choice questions about it. Lastly, the LEABS-Q was administered to each participant.

Coding

Reaction times were recorded using SuperLab software by measuring the time between the presentation of the picture and the initiation of the vocal response into the microphone. Accuracy was assessed by reviewing the participant's recorded performance. The eye-tracking data, consisting of super-imposed cross-hairs onto the field of view, were coded for proportion of eye-movements to the distractor words. Eye-movements to the distractor word were considered to have occurred when the crosshairs have crossed into the quadrant containing the word. Ten percent of the data were coded by a second, independent coder, who did not speak Russian. Point-to-point reliability for coding of proportions of eye-movements was 96%.

Results

Trials on which participants made errors accounted for 4.40% of the data. Picture naming errors were analyzed separately, while errors like false starts were omitted from the analyses.

Reaction times

A 4x2 Anova with condition (no word, Russian words, Phonological controls, and English translations) as a within-subjects variable, and group (monolingual and bilingual) as a between-subjects variable yielded a main effect of condition, $F(1, 28)=8.39$, $p<0.01$, and a significant two-way interaction between condition and group, $F(1, 28)=5.35$, $p<0.05$ (Figure 2).

Post-hoc pair-wise comparisons for condition adjusted for multiple comparisons using the Bonferroni method revealed that both groups had shorter reaction times to pictures without distractor words (M=809.70, SE=14.93) than to pictures accompanied by Phonological Russian words (M=861.77, SE=16.73), $p<0.05$, Phonological controls (M=852.80, SE=16.75), $p<0.05$, or English translations (M=855.36, SE=16.47).

Post-hoc pair-wise comparisons between groups for each condition revealed that bilingual participants had longer reaction times to Phonological Russian words (M=912.80, SE=23.66) than monolingual participants (M=810.74, SE=23.66), $F(1, 28)=9.30$, $p<0.01$. There was no significant difference between bilinguals' and monolinguals' reaction times to Phonological controls $F(1, 28)=3.28$, $p>0.05$, to pictures alone, $F(1, 28)=1.48$, $p>0.05$, or to English translations, $F(1, 28)=0.17$, $p>0.05$.

Proportion of Eye Movements to the Word

A 3x2 Mixed Ancova, with condition (Phonological Russian words, Phonological controls, and English translations) as a within-subjects variable, and group (monolingual and bilingual) as a between-subjects variable, with speed of reading as a covariate, was used to analyze the proportion of eye movements to the three types of distractor words.

Results revealed a main effect of group, $F(1, 27)=4.44$, $p<0.05$, with bilinguals looking more at the written stimuli (M=0.46, SE=0.04) than monolinguals (M=0.34, SE=0.04) and a marginally significant interaction between condition and group, $F(1, 27)=3.74$, $p<0.06$ (Figure 3).

Post-hoc Univariate Analyses of Covariance revealed that bilinguals looked more at the Phonological Russian words (M=0.50, SE=0.05) than monolinguals (M=0.31, SE=0.05), $F(1,27)=7.66$, $p<0.05$, but the two groups did not differ in their proportion of looks to the Phonological controls, $F(1, 27)=2.11$, $p>0.05$, or the English translations, $F(1,27)=2.09$, $p>0.05$.

Error Analysis

On the PWI task, monolingual participants committed 11 mis-naming errors, where a picture was named using the distractor word (for example, naming a picture of the chicken "duck"). Bilingual participants committed 8 mis-naming errors. Of these, 5 were committed with the English distractors being the word stimuli. The Mann Whitney test for independent samples revealed that monolingual speakers of English misnamed more pictures using the English

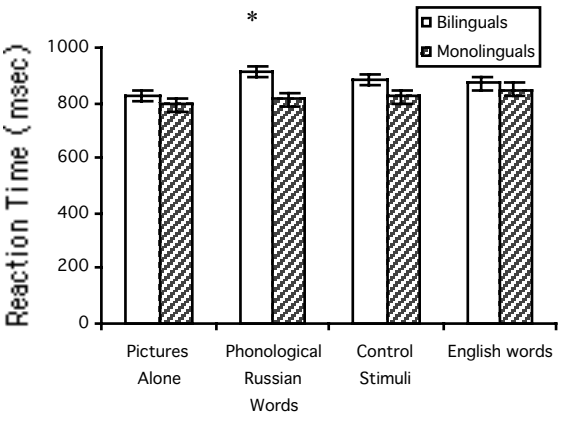


Figure 2: Reaction times of bilingual and monolingual participants when naming pictures alone, pictures, accompanied by phonological Russian words, by bigram-matched non-word control stimuli, and by English words.

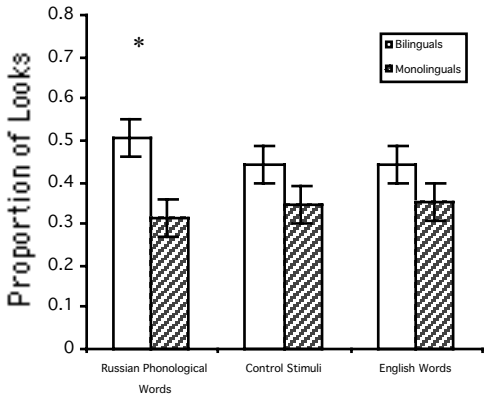


Figure 3: Mean proportion of looks to distractor stimuli made by bilingual and monolingual participants when distractors were phonological Russian words, bigram-matched non-word control stimuli, and English words.

distractor words than bilingual participants, Mann Whitney $U(15)=73.50$, $p<0.06$.

Three mis-naming errors were committed by bilingual participants with distractors being phonological Russian stimuli, such as naming a picture of a *collar* "sleeve," when the distractor word on the screen was RUKAV, a phonological word for "sleeve" in Russian. This number was not significantly different from the number of errors committed by monolingual participants, Mann Whitney $U(15)=90.00$, $p>0.05$.

It is interesting to note that while interference of Russian written stimuli did occur during naming of pictures in English, none of the bilingual participants had switched into Russian when naming pictures. Instead, a spontaneous translation of the Russian distractor into English had occurred, and this, in turn, interfered with naming of the picture. This pattern of errors is consistent with the observation made by Costa, Miosso, and Carammazza (1999), who suggested that items from two languages, while activated in parallel, do not compete for selection during production.

Discussion

Russian-English bilinguals looked at English non-words that composed meaningful phonological Russian words more than monolingual English speakers, while the proportions of eye-movements made to control non-word stimuli and to English translation equivalents were comparable for the two groups. This finding suggests that phonological information in the non-target language drew bilinguals' eye-movements. Therefore, phonological information for the non-target language was automatically activated for these stimuli. According to models of eye movement control in reading, the stimuli that carried phonological information for Russian drew the bilinguals' eye movements because these stimuli carried meaningful information for them, but not for monolingual speakers of English.

Literature on eye movements during reading suggests that before fixating a word, a reader obtains useful information from its parafoveal preview; this information is used by the reader to decide whether to fixate on the word or not (Reichle et al., 1999; Starr & Rayner, 2001). While the E-Z Reader model of eye-movement control in reading posits cognitive control of eye-movements during reading of sentences (Reichle, 1998), it also seems to explain the behavior of the participants in this experiment, where they recognized single words. The decision to look at the word during the modified Picture Word Interference task appears to be dictated by the amount and quality of information a reader gleans from its parafoveal preview. Russian-English bilinguals were better able to control eye-movements to English non-word stimuli that did not carry any phonological information for Russian, than to English non-word stimuli that did carry phonological information for the non-target language.

Testing bilinguals who speak languages with partially overlapping alphabets allows for separating the contributions of orthographic codes and phonological information they carry to the parallel activation of the two languages when processing print. Our findings are in line with connectionist models of visual word recognition (e.g., Plaut et al., 1996; Seidenberg & McClelland, 1989; Van Orden & Goldinger, 1994), which propose that phonological information for a word is automatically activated. The only way to obtain interference effects from phonological Russian words is by activation of the Russian language via its phonology because the orthographic information for Russian is not present in these stimuli. The dual-route models of visual word recognition, while allowing for an indirect phonological route to the lexicon, postulate that this route is specialized for reading non-words (e.g., Coltheart et al, 2001; Ziegler et al., 2000).

Russian-English bilinguals had longer reaction times when naming pictures accompanied by phonological Russian stimuli that constituted words semantically related to the picture names than monolingual speakers; reaction times to non-word control stimuli and English translations were comparable for the two groups. Because slower naming times on the PWI task result from the interference of the written stimulus with the picture name at the lexical-semantic level, this finding suggests that not only was phonology for the Russian language activated, as indicated by greater proportion of eye movements made to these by the bilingual than the monolingual speakers, but that this information was meaningful enough to get processed to the lexico-semantic level. This observation is further supported by the analysis of the error data: Although bilingual participants committed only 3 misnaming errors when the distractor word was a phonological Russian word, the mere fact that these errors exist support the idea that the phonological information for Russian was meaningful enough to activate the relevant lexico-semantic information.

In conclusion, we have successfully shown that phonology of the non-target language is activated for the target-language stimuli that bear little resemblance to the non-target language orthography. Furthermore, we have shown that activation of non-target language phonology is enough to produce interference effects during picture-naming in the target language. These findings extend the idea of parallel activation of languages in bilinguals to those languages in which shared orthographic symbols map onto distinct phonological representations, and inform models of bilingual reading on the role of phonology in the lexical access of written words. Finally, the idea that eye-movements during reading are under at least partial cognitive control offers an intriguing possibility that the bilinguals in this experiment exhibited a measure of cognitive control over interference from the non-target language during the modified PWI task. Future experiments might be able to explore the idea of cognitive control over language interference in bilinguals further by manipulating the amount of meaningful information present for the non-target language, and by testing different groups of bilingual speakers.

Acknowledgments

We thank Karla McGregor, Doris Johnson, Henrike Blumenfeld, Caitlin Fausey, and members of the Bilingualism and Psycholinguistics Research Lab for their helpful suggestions during the course of this research.

References

Brysbaert, M., Van Dyck, G., & Van de Poel, M. (1999). Visual word recognition in bilinguals: Evidence from masked phonological priming. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 137-148.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589-608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegel, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204-256.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection. *Journal of Memory and Language*, *41*, 365-397.

De Groot, A.M.B., Delmaar, P., & Lupker, S.J. (2000). The processing of interlexical homographs in translation recognition and lexical decision: support for non-selective access to bilingual memory. *The Quarterly Journal of Experimental Psychology*, *53A(2),* 397-428.

Deutsch, A., Frost, R., Pollatsek, A., & Rayner, K. (2002). Early morphological effects in word recognition in Hebrew: Evidence from parafoveal preview benefit. *Language and Cognitive Processes*, *15*, 487-506.

Dijkstra, T., Grainger, J., & van Heuven, W.J.B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41*, 496-518.

Dijkstra, T., & Van Heuven, W.J.B. (2002). The architecture of the bilingual visual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5(3),* 175-197.

Dijkstra, T., & Van Heuven, W.J.B. (1998). The BIA model and bilingual visual word recognition. In J. Grainger and A.M. Jacobs (eds.), *Localist Connectionist Approaches to Human Cognition*, pp.189-225. Mahwah, NJ: Earlbaum Associates.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamic model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42(5),* 621-636.

Feldman, L.B. and Turvey, M.T. (1983). Word recognition in Serbo-Croatian is phonologically analytic. *Journal of Experimental Psychology: Human Perception and Performance, 9(2),* 288-298.

Jared, D and Szucs, C. (2002). Phonological activation in bilinguals: evidence from interlingual homograph naming. *Bilingualism: Language and Cognition, 5,* 3225-239.

Liu, W., Inhoff, A.W., Ye, Y., & Wu, C. (2002). Use of parafoveally visible characters during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance, 28*, *1213-1227.*

Lukatela, G., Savic, M., Gligorijevic, B., Ognjenovic, P., & Turvey, M.T. (1978). Bi-alphabetical lexical decision. *Language and Speech*, *21*, 142-165.

Marian, V., Blumenfeld, H., & Kaushanskaya, M. (2003). Language proficiency and bilingual status questionnaire (LEABS-Q). Poster presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL. Manuscript in preparation.

Marian, V., & Spivey, M. (2003 a). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics*, *24*, 173-193.

Marian, V., & Spivey, M. (2003 b). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6*, 97 - 115.

Nas, G. (1983). Visual word recognition in bilinguals: evidence for a cooperation between visual and sound based codes during access to a common lexical store. *Journal of Verbal Learning and Verbal Behavior, 22*, 526-534.

Plaut, D.C., McClelland, J., Seidenberg, M.S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103,* 56-115.

Reichle, E. D. (1998). Towards a model of eye-movement control in reading. *Psychological Review*, *105*, 125-157.

Reichle, E.D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research*, *39*, 4403-4411.

Seidenberg, M.S., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523-568.

Starr, M.S., & Rainer, K. (2001). Eye movements during reading: Some current controversies. *TRENDS in Cognitive Sciences*, *5*, 156-163.

Tzeglov, J., Henik, A., Sneg, R., & Baruch, O. (1996). Unintentional word reading via the phonological route: The Stroop effect with cross-script homophones. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 336-349.

Van Heuven, W.J.B. (2000). *Visual word recognition in monolingual and bilingual readers: Experiments and computational modeling.* Ph.D. Thesis, University of Nijmegen.

Van Heuven, W.J.B., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Langauge*, *39*, 458-483.

Van Orden, G.C., & Goldinger, S.D. (1994). The interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1269-1291.

Wong, K.F.E., & Chen, H-C. (1999). Orthographic and phonological processing in reading Chinese text: Evidence from eye fixations. *Language and Cognitive Processes*, *14*, 461-480.

Ziegler, J. C., Ferrand, L., Jacobs, A. M., Rey, A., & Grainger, J. (2000). Visual and phonological codes in letter and word recognition: Evidence from incremental priming. *Quarterly Journal of Experimental Psychology, 53A*, 671-692.

Should Politicians Stop Using Analogies? Whether Analogical Arguments Are Better Than Their Factual Equivalents

Mark T. Keane (mark.keane@ucd.ie)
Amy Bohan (amy.bohan@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland

Abstract

In political argumentation, analogies are often used to convince an audience of one's views. For example, in political debates leading up to the Iraq War, one such analogical argument was that Saddam Hussein was like Hitler and therefore Saddam should be forcibly ousted. But are all analogical arguments really convincing? In this paper we investigate whether analogical arguments are actually more convincing than factual arguments. In Experiment 1 we asked people to rate analogical and factual arguments for various propositions and found that people considered factual arguments more convincing. In Experiment 2, we asked people to think more explicitly about the analogical mappings but still found that people considered the analogical arguments less convincing than the factual ones. These findings suggest that people are *not* more easily convinced by an analogical argument then a straight factual one, suggesting that perhaps politicians should re-consider their rhetorical tactics after all.

Introduction

Is the aftermath of the Iraq War like Germany post-WWII or Northern Ireland or, indeed, is it another Vietnam? In the furious political debate following the Iraq War, politicians on both sides have used different analogies to bolster their arguments. In science, analogies are often used to discover something new about natural phenomena, but in politics they are used to convince an audience of one's views. In this paper, we consider whether such analogical arguments are more convincing than their equivalent, factual arguments.

Though classical rhetoric has long advocated the use of analogy in argumentation (Plato, *Phaedo*, trans. 1871, 71c-d being a prime exponent of the craft) and political science regularly analyses the analogies used in political debate (Blanchette & Dunbar, 2001), we know of no studies that have systematically determined whether people actually find analogical arguments more cognitively convincing than their factual equivalents. This gap in the literature is all the more surprising when one considers the amount of research on the separate topics of argumentation and analogy. The nature of argumentation has been elaborated in a rich literature in philosophy, logic and psychology (e.g., Rips 2002; Voss & Van Dyke, 2001). Similarly, the nature of analogy has been empirically explored in many studies, supported by clearly articulated theory that has been modeled computationally (see Gentner, 1983; Holyoak & Thagard, 1995; Keane, 1997; Keane, Ledgeway & Duff, 1994; Hummel & Holyoak, 1997). Yet, the two areas have not been combined in a systematic study of their cognitive underpinnings. In the present paper, we attempt such a combination.

The Present Experiments

We propose a novel paradigm for assessing people's evaluation of arguments that pits analogical arguments and their factual equivalents against one another. In our experimental setup, people are presented with a proposition and a two-fact argument supporting this proposition (see Figure 1). They are then asked to rate how good they found this argument as a warrant or support for the proposition. For a given proposition, the argument was either two facts or two equivalent analogical facts.

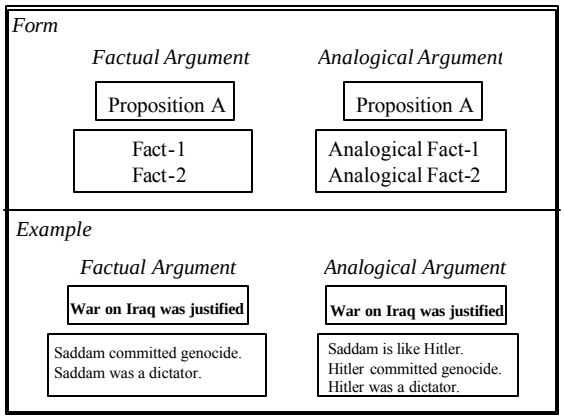


Figure 1: Abstract form and a gloss of a sample argument used in the experiments.

For example, the Iraq War argument suggests that going to war with Iraq was justified because Saddam was a dictator and had committed genocide in his country (see gloss in Figure 1). The analogical equivalent suggests that war on Iraq was justified because Saddam is like Hitler, and Hitler was a dictator and had committed genocide in his country. In this way, the analogical argument presents the same facts about Saddam but through the lens of a WWII analogy. This is the typical way in which politicians use analogies, suggesting a parallel in an analogous domain that supports their argument in a current domain.

From a cognitive perspective, there are several reasons why analogical arguments might be more convincing than factual ones. Essentially, an argument is defined as "a course of reasoning aimed at demonstrating the truth or falsehood of something" (Kuhn, 1991, pg.12). That is, in any argument the aim is to convince an audience of the truth or falsehood of certain facts, that these facts in some way support or warrant your proposition and that, therefore, your proposition is justified as right or correct (Toulmin, 1958; Kuhn, 1991). Therefore if there is agreement that Saddam's dictatorial powers and genocidal activities are bad and that these facts warrant the action of war as a response; then making war is, in some way, a necessary response to the facts given. One of the key steps in this process is getting the audience to accept the warrant as a necessary link between the facts and the proposition. Cognitive models of analogical thinking, show us that people use analogies to make high-level causal inferences about analogous domains (c.f., Keane, 1988). In this case, the comparison to Hitler provides a match to WWII where these dictator and genocide facts were viewed as essential reasons for military intervention. Thus, the analogy provides a previous case where the facts caused or strongly warranted military intervention, inviting the inference that war is therefore appropriate in Iraq too.

We report two experiments on the role on analogy in argumentation. These experiments used a wide variety of topical arguments from different domains covering alcohol abuse, military service, university entrance exams and traffic congestion policy. The analogies used also varied in the distance of the domains from one another; some involved close domains (e.g., Iraq War and WWII), others involved distant domains (e.g., Art and Foreign Languages). In the experiments, no single individual saw both the factual and analogical versions of a given argument. We also gathered people's ratings of their *a priori* belief in the proposition (i.e., their agreement/disagreement with it) to check for any belief bias in their assessment of the argument. In Experiment 1, we made a direct comparison of people's goodness ratings for the factual and analogical arguments to various propositions. In Experiment 2, we replicated this test with an intervention that encouraged people to reflect more on the analogy. To presage our findings, the evidence suggests that people are *not* more easily convinced by an analogical argument over a straight factual one, suggesting that politicians might indeed want to re-consider their rhetorical tactics.

Experiment 1

This experiment examined whether analogical arguments were deemed to be better (i.e., more convincing) than their factual equivalents for a variety of topical propositions. People were shown 10 different propositions (5 with factual arguments, 5 with analogical arguments) and asked to carry out two tasks on each: a belief task and an evaluation task. In the belief task, they were shown the proposition on its own and asked to rate their agreement/disagreement with it on a 7-point scale. In the evaluation task, they were shown the proposition and the argument (factual or analogical) and asked to rate its goodness as an argument for the proposition on a 7-point scale. The order of these tasks was counterbalanced in two different conditions. If our politicians are right then the analogical arguments should be considered to be better than their factual equivalents.

Method

Materials. Ten propositions were created based on either currently debated topics (e.g., the Iraq War, the school examination system, societal effects of drugs, utility of GM foods) or long-standing debated topics (e.g., the introduction of the death penalty, military service, public funding of the arts). For each of these propositions, a two-fact argument was created based on the typical reasons used to support these propositions. Analogies were then developed that had clear one-to-one correspondences to the conceptual objects and relations used in the original facts. Eight different materials sets were made up from random selections of particular materials and arguments, such that each set contained 10 unique propositions, 5 of which had corresponding analogical arguments with the other 5 having corresponding factual arguments. This material-group variable is not reported in the results as an analysis of people's ratings shows that it had no reliable effect on results found.

For every material set, two booklets were collated for the two tasks. The belief-rating booklet had a cover sheet explaining that people should rate how strongly they agreed/disagreed with the proposition on a 7-point scale, followed by 10 pages with a single proposition and rating scale shown on each page. The evaluation booklet had a cover sheet explaining that people should rate how good/bad they thought the argument was for the proposition on a 7-point scale regardless of their beliefs, followed by 10 pages with a proposition plus its corresponding (factual /analogical) argument and a scale shown on each page. The items in every booklet were randomly ordered for each participant.

Participants & Design. Thirty-two native English-speaking undergraduates at University College Dublin took part in the experiment. The order of the tasks was counterbalanced so that half the participants received the belief task before the evaluation task (belief-then-evaluation conditions) while the other half received the tasks in the opposite order (evaluation-then-belief conditions). So, the design was a 2 argument-type (factual or analogical) x 2 task-order (belief-then-evaluation or evaluation-then-belief) one with argument-type being within-participants and task-order being between-participants.

Procedure. In the evaluation task, participants read instructions that explained the 1-7 argument goodness scale (1 being "very bad", 7 being "very good" and 4 being "neither good nor bad"), and a sample proposition was

shown with a factual argument and another shown with an analogical argument. The participants were asked to take their time over each decision and to make "an *objective* assessment of the arguments. That is, to make a judgement regardless of your agreement or disagreement with the proposition". Each proposition-argument pair was presented on a separate page with a marked space for participants to note their 1-7 goodness rating. In the belief task, the instructions and materials were presented in the same way, except that the proposition alone was presented and the instructions explained that people were to rate how strongly they disagreed/agreed with the proposition on the 1-7 agreement scale (1 being "strongly disagree", 7 being "strongly agree" and 4 being "no opinion").

Table 1: Percentage of good arguments and mean goodness ratings for both experiments

	Analogical		*Factual*	
Measure / *Experiment*	%Good	Mean Rating	%Good	Mean Rating
Expt. 1				
belief-then-evaluation	21.3%	2.59	58.8%	4.3
evaluation-then-belief	30%	3.11	42.5%	3.76
Mean	25.6%	2.85	50.6%	4.03
Expt. 2				
belief-then-evaluation	33.8%	3.43	68.4%	4.59
evaluation-then-belief	43.8%	3.80	46.2%	3.90
Mean	38.8%	3.6	57.2%	4.25

Results

Table 1 summarises the main results of the Experiment showing that the factual arguments were considered to be better than the analogical ones on several different measures.

Percentages of Good and Bad Arguments. A rough feel for people's responses to the arguments can be gleaned by re-classifying their ratings into ordinal groups of good (> 4), bad (< 4) or indifferent (=4) according to how they rated the argument on the goodness scale. Overall, 320 arguments were evaluated in the experiment, 160 factual and 160 analogical. Of the factual arguments, 38.1% (61) were rated as bad and 50.6% (81) as good (the remainder being indifferent). Of the analogical arguments, 67.5% (108) were rated as bad and 25.6% (41) as good (the remainder being indifferent). Collapsing across the order conditions, this result was found to be reliably different using a Chi-squared analysis, $\chi^2(1) = 26.032$, $p < 0.0001$, N=291. However, an inspection of the percentages clearly shows that task-order has an impact too, in that more arguments were considered to be good in the belief-then-evaluation conditions (40%) than in the evaluation-then-belief conditions (36%). Indeed, on the face of it, there appears to be an interaction between task-order and argument-type that is more easily revealed using the ratings measure.

Ratings of Arguments. A 2x2 ANOVA was carried out on the ratings data for the between-participant variable of task-order and within-participant variable of argument-type. All analyses of variance by participants and by items were performed by respectively treating participants (F_1) and sentences (F_2) as a random factor. These analyses revealed a main effect of argument-type with the factual arguments (M=4.03) being rated as being better than the analogical arguments (M=2.85), $F_1 (1, 286) = 40.02$, $p < 0.0005$, MSe = 111.628; $F_2 (1, 307) = 40.30$, $p < 0.0005$, MSe = 111.628. There was also a reliable interaction between task-order and argument-type $F_1 (1, 286) = 8.10$, $p < 0.005$, MSe = 22.578; $F_2 (1, 307) = 7.20$, $p < 0.008$, MSe = 19.938. Planned pair-wise comparisons revealed that the factual/belief-then-evaluation condition was reliably different to all the other conditions using Bonferroni adjustments (ps<0.0005). None of the other comparisons were reliably different to one another.

The Impact of Belief on Evaluation. One of the key questions was whether people's prior beliefs in the proposition would have any impact on their rating of the goodness of the argument, even though we asked people to be as objective as possible. If people were rating the arguments in line with their beliefs then we should, for example, find that people gave high goodness ratings to arguments in which they strongly agreed with the proposition and low goodness ratings to arguments with which they strongly disagreed. However, there is little evidence of such a relationship. The correlation between participants' belief ratings and their goodness ratings for the items is low and not reliable, using Pearsons product-moment correlation $r(319) = 0.36$, $p<0.0005$.

Discussion

This experiment reveals three main findings: (i) analogical arguments are *not* considered to be better than their factual equivalents, (ii) people's *a priori* agreement/disagreement with the proposition does not affect their subsequent evaluation of the goodness of an argument for that proposition, (iii) people find factual arguments much better if they are first asked to rate their belief in the proposition.

The first of these findings should be a surprise for most politicians, as it shows that they might as well be using straight-forward factual arguments to present their views. In the next experiment, we explore whether this result may have occurred because people did not process the analogy sufficiently to draw out all its implications.

The second finding suggests that people can separate their belief in the proposition from their assessment of its goodness, when they are instructed to do so. In other words

that people can maintain a level of objectivity in evaluating arguments.

The third finding of a task-order effect was as unexpected as it is interesting. It shows that if someone rates their belief in a proposition and subsequently sees a factual argument for that proposition they consider it to be better than the same argument presented before they give their belief rating (this effect does not occur for analogical arguments). Why should this occur? One possibility is that when people are first asked to rate their agreement with the proposition, they must think of their own arguments for the proposition. Then, when they are subsequently shown some arguments for the proposition many participants may find them more convincing because they are similar to their own arguments. In contrast, when participants are first asked to evaluate the argument and proposition (before being asked for their belief) there is less opportunity to think of their own arguments, less opportunity to recognize similarities and, hence, less of a boost to the goodness rating of the argument. No parallel benefits are found for the analogical arguments because people do not readily think of their own analogical arguments when rating their belief in the proposition (see Gick & Holyoak, 1980; Keane, 1985, 1988, on people's tendency *not* to explore analogical possibilities without instructions to do so). In the next experiment, we attempt to replicate this task-order effect to determine whether it is robust.

Experiment 2

In Experiment 1, we found that people failed to be convinced by analogical arguments relative to their factual equivalents. This result could be due to the amount of cognitive processing people have to carry out on analogical arguments as opposed to factual arguments. In the analogical case, they must understand the analogical arguments, map the corresponding objects and relations between the two domains, then apply the mappings to the proposition's domain and, finally, evaluate it. In the factual case, they merely have to understand the argument, relate it to the proposition and evaluate it. Maybe participants in Experiment 1 did not bother to draw the analogy and, hence, marked these arguments down. We should note that this explanation is somewhat implausible as we know from the literature that people readily appreciate and understand analogies (see Keane, 1988; Holyoak & Thagard, 1995). So, in this experiment, we explicitly asked people to report their mapping of key objects between the two domains to ensure that the analogy was being properly processed. We also ran the task-order manipulation again to see if it could be replicated.

Method

Participants, Materials & Design.. Thirty-two native English-speaking student volunteers at University College Dublin took part in the experiment. The materials were the same as those used in Experiment 1, as were the grouping of material sets and organization of task booklets. As before, the design was a 2 argument-type (factual or analogical) x 2 task-order (belief-then-evaluation or evaluation-then-belief) one, with argument-type being within-participants and task-order being between-participants.

Procedure. The procedure was as in Experiment 1, except for one change to the analogical argument materials. In each case where an analogical argument was presented, people were shown two boxes listing three key objects from each domain of the analogical argument (as shown in Figure 2). The participants were asked to draw lines between the corresponding objects in the analogy. For example, in the Saddam Hitler analogy, Hitler corresponds to Saddam, Germany corresponds to Iraq etc. They were asked to perform this mapping before they rated the analogical argument in the evaluation task.

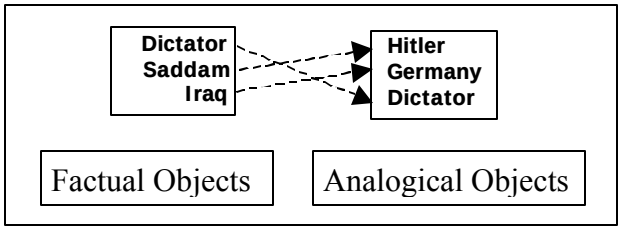


Figure 2: Example of the object-mapping task used in Experiment 2

Results and Discussion

Table 1 summarises the main results of the experiment showing that the factual arguments were considered to be better than the analogical ones on several different measures. The pattern of findings replicates those found in Experiment 1, with a strengthening of the effects being found.

Percentages of Good and Bad Arguments. Re-classifying people's responses into the ordinal groups of good (> 4), bad (< 4) or indifferent (=4) we found that (i) of the 159 factual arguments evaluated 35.2% (56) were rated as bad and 57.2% (91) as good (the remainder being indifferent), (ii) of the 160 analogical argument evaluated 52.5% (84) were rated as bad and 38.8% (62) as good (the remainder being indifferent). Collapsing across the task-order conditions, this result was found to be reliably different using a Chi-squared analysis, $\chi^2(1) = 11.093$, $p < 0.0009$, N=293.

Ratings of Arguments. A 2x2 ANOVA was carried out on the ratings data for the between-participant variable of task-order and within-participant variable of argument-type. All analyses of variance by participants and by items were performed by respectively treating participants (F_1) and sentences (F_2) as a random factor. These analyses revealed a main effect of argument-type with the factual arguments (M=4.25) being rated as being better than the analogical arguments (M=3.6), F_1 (1, 255) = 14.17, $p < 0.001$, *MSe* =

37.154; F_2 (1, 279) = 7.06, $p < 0.016$, MSe = 38.465. There was also a reliable interaction between task-order and argument-type F_1 (1, 255) = 8.63, $p < 0.006$, MSe = 22.623; F_2 (1, 279) = 10.02, $p < 0.005$, MSe = 23.646. Planned pair-wise comparisons revealed that the factual/belief-then-evaluation condition was reliably different to all the other conditions using Bonferroni adjustments (*p*s<0.005). None of the other comparisons were reliably different to one another. So, again, we replicate the task-order x argument type interaction found in the previous experiment. The main effect of task order was not reliable.

The Impact of Belief on Evaluation. Again, we found as in Experiment 1, that there is little evidence to suggest that people's prior beliefs in the proposition affected their assessment of the argument. The correlation between participants' belief ratings and their goodness ratings for the items is low and not reliable, using Pearsons product-moment correlation r(318) = 0.243, p<0.0005.

Conclusions from Experiments 1 and 2. So, again we find that the analogical arguments were considered to be less convincing than the factual ones, even when we ensure that people have mapped the analogy appropriately. However, it is noteworthy that, relative to Experiment 1, their analogical arguments seem to be rated as slightly better (e.g., 38.75% are considered good arguments in Experiment 2, relative to 25.63% in Experiment 1).

General Discussion

The results of these experiments suggest that politicians should stop using analogies, as they do not seem to provide much more than a sugar coating on the convincingness of a straight, factual argument. Overall, we have shown several novel findings about the use of analogy in argumentation. First, we have seen that analogical arguments are generally not considered to be as good as factual arguments. Second, we have seen that it is very hard for analogical arguments to challenge the goodness of factual arguments (in other experiments we have found that even when the full factual argument is given along with the analogical argument, the evaluations do not go higher than the plain factual arguments). Third, we have found that factual arguments' ratings can be boosted if people are asked to reflect on the proposition in advance of rating them. Finally, we have seen that people can separate their beliefs in a proposition from their evaluation of an argument to that proposition, showing a noteworthy objectivity in their evaluations. To traditional rhetoricians this evidence may seem unwelcome and unconvincing. In the remainder of this section, we consider three main objections that might be raised to our findings.

The Arguments Were Not Very Good. One argument against the evidence would be to maintain that the arguments used were not very good; that if you had better arguments then different results would be found. Unfortunately, we do not have data on how many people in a population find a given argument to be good or bad relative to some proposition, so it is hard to judge whether our arguments are in some way unrepresentatively poor. What we do know is that people only found 35-38% of our factual arguments to be bad (50%-60% of these arguments being considered good). On the face of it, using the "you can fool some of the people all of the time…" adage these figures appear to be reasonable levels of goodness. As such, we would argue that there is no obvious deficiency in the arguments used. Furthermore, we should also note that many of the arguments used were ones that people have used to support these propositions in everyday life.

Maybe Our Analogies Are Not Very Good. If one admits that the arguments are adequate, then a further objection could be that the analogies were, in some way, inadequate. Again this is a hard objection to assess given that we have little idea of the space of possible analogies used in argumentation. What we can say is that all of the analogies used conform to what is deemed to constitute an analogy in the literature; they involve one-to-one mappings, they involve matching relational structure and they suggest inferences by analogy connecting the arguments and the proposition (c.f., Gentner, 1983; Hummel & Holyoak, 1997; Keane et al., 1994). But, what if some are, in some way, better than others.

To explore this possibility, we presented a separate group of 16 participants with a mixture of 10 analogies and non-analogies asking them to rate the goodness of the analogies on a 7-point scale. Of the 10 materials used in the experiment only one received a bad goodness rating (i.e., < 4), all of the reminder being rated as being good (with mean ratings from 4.25 to 5.25). Overall, people reliably distinguished the analogies (M= 4.6) from the non-analogies (M = 2.5), using a dependent t-test, t(157) = 8.10, $p < 0.0005$. So, the failure of the analogical arguments cannot be attributed to the poorness of the analogies.

Are There Other Ways in to Improve Analogies? A final objection is that we have not appropriately intervened to boost the analogy. We saw that asking people to plot the object mapping improves their goodness ratings for the analogy arguments. Perhaps there is some other intervention that might boost them further. It is unclear to us what this intervention might be. However, this objection in a sense misses the point. If we did find some intervention that promotes analogical arguments is it quite likely to be quite artificial. In the cut and thrust of political debate the facts of the matter are generally known (though may not be stated explicitly) and the analogy is provided to be understood on the spot (without, for example, asking people to specify the object mappings).

Acknowledgments

This work was funded in part by grants from University College Dublin and Science Foundation Ireland under Grant No.03/IN.3/I361 to the first author.

References

Blanchette, I. & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. Memory & Cognition 29(5), 730-735.

Gentner, D (1983). Structure-mapping: A theoretical framework for analogy, Cognitive Science, 23, 155—170.

Gick, M.L., & Holyoak, K.J. (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.

Holyoak, K. J. & Thagard, P. (1995). Mental Leaps: Analogy in Creative Thought. The MIT Press, Cambridge, MA, 1995

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representation of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.

Keane, M.T. (1997). What makes an analogy difficult?: The effects of order and causal structure on analogical mapping. Journal of Experimental Psychology: Learning, Memory & Cognition, 23, 946-967.

Keane, M.T. (1998). Analogical Problem Solving. Ellis Horwood Ltd, Chichester.

Keane, M.T., Ledgeway, T.& Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. Cognitive Science, 18, 287-334.Rips, L.J. (2002). Circular Reasoning. Cognitive Science, 26, 767-795

Kuhn, D. (1991). The skills of argument, Cambridge University Press.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49*, 67-96.

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory and Cognition, 26*(5), 965-978.

Toulmin, S. (1958). The uses of argument. Cambridge University Press.

Voss, J.F., & Van Dyke, J.A. (2001). Argumentation in Psychology: Background Comments. Discourse Processes, 32(2&3) 89-111.

Zwaan, R.A., Magliano, J.P., & Graesser, A.C. (1995). Dimensions of situation-model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 386-397.

Information Visualizations for Supporting Knowledge Acquisition - The Impact of Dimensionality and Color Coding -

Tanja Keller (t.keller@iwm-kmrc.de)
Virtual Ph.D. Program "Knowledge Acquisition and Knowledge Exchange with New Media"
University of Tuebingen, Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Peter Gerjets (p.gerjets@iwm-kmrc.de)
Multimedia and Hypermedia Research Unit, Knowledge Media Research Center
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Department of Applied Cognitive Psychology and Media Psychology, University of Tuebingen
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Bärbel Garsoffky (b.garsoffky@iwm-kmrc.de)
Multimedia and Hypermedia Research Unit, Knowledge Media Research Center
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Abstract

So far, information visualizations, i.e., graphical representations of huge amounts of abstract data which do not have a natural visual representation, have mainly been used to support *information-retrieval tasks*. In this paper we investigate whether information visualizations are also suitable to foster tasks that focus on *knowledge acquisition or learning*. In addition, we address the issue of how information visualizations have to be designed to be efficient learning tools. We conducted an experimental study which provided evidence that information visualizations can foster knowledge acquisition and that 2D-information visualizations are better suited for knowledge acquisition than 3D-ones. In addition, we found slight performance improvements due to using color to code information.

Technological innovation allows storing fast growing quantities of information. Accordingly, it has become increasingly important to develop efficient methods to structure large and complex information sets. Recently, there have been several attempts to tackle this challenge by using information visualizations, i.e., graphical representations of large amounts of abstract data which do not have a natural visual representation (Wiss, Carr, & Jonsson, 1998). For instance, information visualizations have been used to display abstract data like document collections or text-based information contents in the WWW. So far, information visualizations have mainly been investigated with regard to technical issues and in the context of information-retrieval tasks – where they proved to be very useful to improve users' ability to use information. However, it is not clear whether information visualizations can also foster knowledge acquisition or learning. Additionally, little is known about the cognitive processes involved in, and maybe supported by the use of information visualizations as learning tools. Therefore, the aim of our empirical study was to investigate to what extent multidimensional information visualizations are superior compared to a non-spatial representation when the task is to memorize a data set and to acquire an understanding of the relationships embedded within this set. Moreover, we were interested in the design of information visualizations for learning. Particularly, we investigated experimentally whether information visualizations should be two-dimensional or whether a third spatial dimension may be helpful for knowledge acquisition. Finally, the question is addressed whether knowledge acquisition with spatial information visualizations can be further enhanced by using color coding to represent attributes of data.

Figure 1 shows a simplified sketch of the type of spatial information visualization used in the empirical study presented in this paper. In this sketch, three attributes of four information units A, B, C, and D are represented by means of three spatial dimensions.

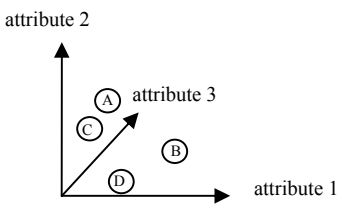


Figure 1: Simplified sketch of a 3D-information visualization.

Information units pool those parts of data sets that belong together. The units can be described by their values on numerous different attributes. Typically, only a subset of these attributes can be represented spatially. Thus, other attributes of the information units may be represented textually or by other codes (e.g., color coding).

What is the Pedagogical Potential of Information Visualizations?

There are different cognitive theories arguing that information visualizations may be efficient tools to enhance the acquisition of knowledge on large and abstract data sets,

whereby knowledge acquisition refers to understanding and memorizing abstract data and their interrelations.

- Firstly, *theories of computational effectiveness* pay specific attention to the inferences learners have to make in order to understand a task or a domain. The argument here is that some representational codes facilitate some inferential (learning) processes better than others. In their seminal work, Larkin and Simon (1987) found for example that search processes in physics are performed much easier with diagrammatic representations than with textual ones. This idea that different representations with the same "content" can still offer different processing opportunities is called "computational effectiveness" (Larkin & Simon, 1987). Following this idea, spatial information visualizations may allow learners to draw inferences very easily on how different information units are related to each other with regard to those attributes that are represented spatially. In this respect, information visualizations are rather similar to concept maps because both of them allow arranging information units spatially in a specific way. Concept maps are 2D-diagrams that illustrate relationships between concepts in a domain by representing these concepts as nodes. These nodes are connected by labeled lines in order to represent their interrelations. It could already be shown that concept maps foster processes of knowledge acquisition (Tergan, 2003), as these representations provide learners with a better understanding of the structures underlying a domain without imposing high cognitive demands on them to extract this information. Due to the aforementioned similarities between information visualizations and concept maps, these processing advantages should also hold for information visualizations.

- Secondly, the *cognitive theory of multimedia learning* (Mayer, 2001) is based on a dual-channel assumption which proposes that textual information is processed and encoded in a verbal system, whereas pictures or graphics are predominantly processed in a pictorial system. The theory assumes that a well-designed combination of text and graphics leads to better memory retention than the use of only one representation. The reason for this is that using the capacity of both memory systems should lead to more information being processed than using only one of the systems. In addition, dual coding might contribute to the construction of a stronger mental model, if the information of both processing systems has to be integrated actively. In addition, research on spatial cognition differentiates between a what-system and a where-system for visual cognition (Landau & Jackendoff, 1993). The where-system is used to process the location of objects, whereas the what-system is dedicated to identify features of an object itself. Memory studies revealed that the where-system operates more effectively with respect to speed and accuracy than the what-system (e.g., Amorim, Trumbore, & Chogyen, 2000). Representing attributes of information units by means of two or three spatial dimensions (instead of a purely textual representation or color coding) might accordingly improve the processing of these attributes by deploying a more efficient processing system.

- Thirdly, following the *cognitive load theory* (Sweller, van Merriënboer, & Paas, 1998), instructional procedures should be designed to prevent cognitive overload. More specifically, the amount of cognitive processing not directly relevant to learning - and thus causing extraneous cognitive load - should be reduced. The necessity of avoiding high levels of extraneous load is especially relevant when the contents to be learned are complex in relation to learners' level of prior knowledge. In this case, the representation of the learning contents imposes a considerable amount of intrinsic cognitive load so that substantial extraneous load can lead to overload in that no more capacity for processes of understanding is left. Cognitive processes directly relevant to understanding and learning are causing germane cognitive load. There are thus two reasons why information visualizations might be particularly appropriate to facilitate learners' acquisition of complex data structures that consist of highly interrelated information units. Firstly, distributing different attributes of information units across different memory and processing systems might provide additional processing resources that can be used to increase germane cognitive load. Secondly, providing learners with a spatial representation of some attributes of information units might reduce extraneous cognitive load by reducing search processes as well as making it easier to draw inferences on how different information units are related to each other with regard to these attributes.

According to these theoretical considerations it can be hypothesized that information visualizations might have a substantial pedagogical potential because they allow to deploy cognitive resources available for learning in a way that is more appropriate than it is with conventional representations of large sets of information units (e.g., spreadsheets).

How to Design Information Visualizations for Knowledge Acquisition?

Beyond the general claim that information visualizations are tools that might foster the acquisition of knowledge on large and abstract data sets, we are also interested in the issue of designing profitable visualizations. Particularly, the study reported in this paper addresses how dimensionality of information visualizations and color coding of attributes might affect learning.

2D- versus 3D-information visualizations?

Although, there are a few empirical studies investigating the dimensionality of information representation in general, nearly none of these studies is related to information visualization or even learning with information visualization. Furthermore, these findings seem to be rather inconsistent and depending heavily on the concrete tasks accomplished with the information representation. For instance, Park and Woldstad (2000) found that 2D-displays are superior to 3D-displays for performing telerobotic tasks. Contrarily, the study of Risden, Czerwinski, Munzer, and Cook (2000) compared 2D- and 3D-browsers with regard to the ease of information retrieval and concluded that 3D-visualizations are preferable. However, only a small number of studies demonstrated the superiority of 3D-representations. In sum, the existing evidence is by no means

sufficient to decide whether information visualizations should be 2D or 3D in the context of learning tasks.

From a more theoretical point of view, one might assume that representing three attributes of information units spatially should be superior to representing only two attributes in a spatial format because of the abovementioned advantages of spatial representations in general (i.e., distribution of information across processing systems, superiority of the where-system, computational effectiveness). However, 3D-information representations might at the same time impose additional extraneous cognitive load onto learners due to the fact that they are usually associated with an increased interactivity and with additional orienttation demands. For instance, 3D-visualizations usually have to be equipped with the option to look at information units from different viewpoints (e.g., by rotating the visualization) to counteract the problem that information units might be concealed by other units. As a result, this interactivity may impose additional cognitive processing demands because learners must control the interaction with the environment and maintain orientation.

We studied the role of the dimensionality of information visualizations empirically to decide whether the advantages or the disadvantages of introducing a third spatial dimension prevail in knowledge acquisition.

Should information visualizations for knowledge acquisition be color-coded or not?

The issue whether it might be helpful to enhance information visualizations by color coding of particular attributes of information units seems to be less ambiguous than the role of dimensionality. As color is a basic element of visual perception (Treisman, 1987), color coding can be expected to make information more salient. Therefore, color coding should provide learners with a better understanding of the structures underlying a domain. It has been shown that coloring objects increases learners' ability to retrieve object information from memory. As the color of objects is stored in long term memory together with other object information (e.g., Hanna & Remington, 1996), color information provides an additional cue for memory retrieval. It can thus be hypothesized that color-coded information visualizations should be superior to those without color coding. However, it remains an open question whether color coding and dimensionality will interact when they are combined with each other. On the one hand, combining spatial representation and color coding results in multiple memory traces which should enhance learning; on the other hand, encoding the same attribute of an information unit by means of two different representational codes might make it necessary to map two representational systems onto each other. This might involve the processing of redundant information which in turn can result in additional extraneous cognitive load and learning impairments. Therefore, it is unclear and subject to experimental investigation whether introducing a double coding of particular attributes of information units will support or hinder knowledge acquisition.

Experiment

In this experiment we first investigated whether information visualizations are more suited to foster knowledge acquisition than text-based information representations. Secondly, we analyzed whether dimensionality and color coding of information visualizations influence learning.

Method

Participants Subjects were 100 students (56 female, 44 male) of the University of Tuebingen, Germany. Average age was 24 years.

Materials and procedure This work is associated with the European project "Mummy" of the Computer Graphics Center in Darmstadt (Germany), which focuses on mobile knowledge management using multimedia-rich portals for context-aware information processing, e.g., at construction sites. Therefore, our experimental environment was designed to provide architects with an overview on the details of their construction projects. Each project is described by values on six different project attributes, namely "rate of return", "construction costs per sqm", "number of problems", "construction progress", "size of construction site", and "construction volume".

With regard to the procedure, first the participants received a booklet for measuring different control variables like retentiveness in a paper-pencil test. Afterwards, they received an introduction to the experimental environment and its usage. To ensure that all participants saw the same information, the exploration of the environment during the subsequent practice phase was standardized. In the learning phase subjects were given 50 minutes to accomplish five tasks. In the context of these tasks they had to find 14 of the 42 information units and had to learn the data contained in these information units. Consecutively, subjects received another booklet containing 35 test tasks. In this test phase the learning materials were no longer available. There were no time limits during testing. Finally, participants had to fill out a questionnaire asking for difficulties regarding the use of the learning materials, the strategies used as well as assessing the cognitive load experienced during learning.

Design and dependent measures As an experimental baseline, the information on the construction projects was represented by means of a spreadsheet which listed 42 construction projects (i.e., information units) alphabetically (Figure 2). The first column in Figure 2 represented the name of the construction projects, whereas the other columns contained the values of these projects with regard to the six aforementioned attributes. The last column listed further project information beyond these attributes.

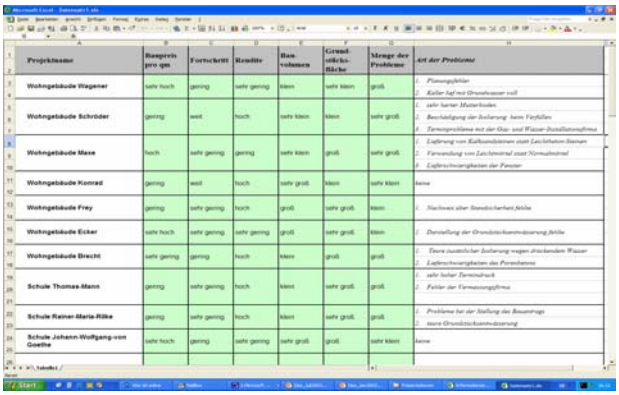


Figure 2: Excel spreadsheet representation (baseline).

To reduce complexity, the range of possible attribute values was restricted to four (i.e., very small, small, big, and very big). Due to the spreadsheet size it was impossible to see the data of all projects without scrolling. A pilot study showed that using this spreadsheet to memorize the abstract data set and to recognize relations between information units was a very difficult task for the subjects.

In order to implement our experimental manipulation, we represented the same data set by means of information visualizations that were either 2D or 3D and that were either monochrome or used color to represent one of the attributes. All of the manipulations (i.e., dimensionality and color coding) referred to the same specific attribute ("construction progress") and the way it was represented. In the 2D-information visualizations both "size of construction site" and "construction volume" were visualized spatially, i.e., they were represented by the axis of the 2D-information space (Figure 3). Information units were arranged in this information space according to their values on these two attributes. In Figure 3 the information units are represented by squares (labeled by their project name). The value of the attribute "construction progress" was represented by a digit attached to the project label. This digit was visible in all four information visualizations.

Figure 3: Two-dimensional color-coded information visualization with opened pop-up window.

The remaining project attributes ("rate of return", "construction costs per sqm", and "number of problems") as well as the further project information could be accessed through pop-up windows by clicking on the information units. In Figure 3, one pop-up window is opened. The pop-up windows could be moved with the mouse by learners in case the window concealed information of interest. To facilitate orientation, the project label of the viewed information unit changed its color from white to red and position lines from the information unit to the axes appeared while contacting the unit with the mouse pointer (position lines, see Figure 4).

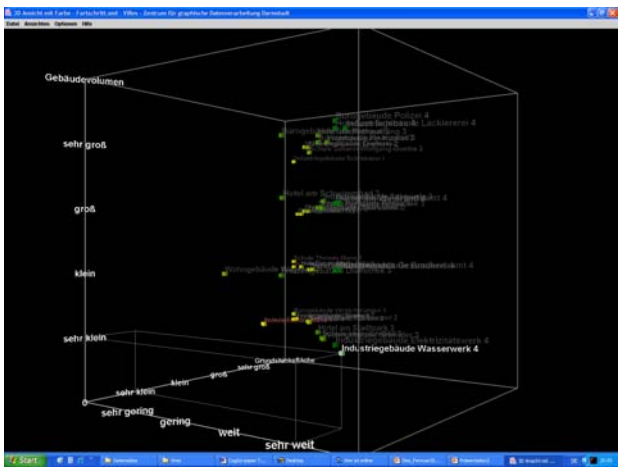


Figure 4: Three-dimensional color-coded information visualization with position lines.

In the 3D-information visualizations a third axis was included to visualize the attribute "construction progress" spatially (Figure 4). To ensure that all information units would be visible in the 3D-information visualizations, the users were allowed to rotate the vertical axis by moving the visualization with the mouse button pressed. To avoid "lost in navigation phenomena", users could push a home-button to attain the start perspective again at anytime.

The colored conditions differ from the monochrome information visualizations depending whether "construction progress" was additionally represented by means of color coding. In the monochrome conditions the information units were always presented in blue against a black background. However, in the colored conditions the information units were displayed in colors ranging from light yellow to dark green – indicating the values of "construction progress".

To sum up, the information visualization conditions differed in the representation format for the attribute "construction progress". In all information visualization conditions the values on this attribute were represented symbolically as a digit. In addition, in 3D-information visualizations the values on "construction progress" were visualized on the third axis. In 2D-conditions there was no spatial representation of this attribute. Furthermore, in polychrome information visualizations the values on the attribute "construction progress" were represented by means of the color of the information units. In monochrome conditions no color was used to visualize this attribute.

Dimensionality and color coding were both varied between subjects resulting in a 2x2-design (plus the baseline spreadsheet condition). Subjects were randomly assigned to the spreadsheet or to one of the four information visualization conditions.

With regard to the dependent measures, as a first dependent variable we measured *performance* with regard to the different knowledge tasks. Overall performance was calculated as the sum of both correct answers and partial correct answers. For 10 of the 35 tasks, partial credits were assigned to score subjects' answers. In the remaining tasks one point was assigned for each correct answer. For each task a maximum of one point was possible resulting in a maximum overall score of 35 points. *Relational performance* referred to tasks that asked for comparative judgments with regard to attribute values, whereas *item-specific performance* focused on specific attribute values. Both measures consisted of 15 tasks each. Five further tasks assessed *structural performance* which was concerned with the recognition of correlational structures within the data set. Furthermore, in each case four tasks were used to assess where-performance, what-performance, and varied-performance. *Where-performance* assessed knowledge on the attributes that were visualized spatially in all information visualizations, whereas tasks on *what-performance* registered knowledge on information always presented as text. Finally, *varied-performance* was concerned with knowledge on "construction progress", i.e., on the attribute whose representation was varied across conditions.

As a second dependent variable we measured learners' confidence with regard to the correctness of their answers. Learners rated each answer to a task with regard to whether they felt low, middle, or high confidence that their answer had been correct. In the *overall confidence* measure these ratings were summed across all tasks, whereby higher rating indicated higher confidence. This overall measure was subdivided into *confidence for correct answers* displaying a participant's belief in that a correct answer was correct. *Confidence for wrong answers* indicated a participant's conviction that a false answer was correct. Because there were 35 items for which every subject had to rate his or her confidence and because ratings ranged from one to three a maximum of 105 points was possible for each of the confidence scores.

As a third dependent variable, we assessed learners' subjective cognitive load by asking them how much *effort* they had to invest into learning and how *difficult* it had been to remember the contents. The effort and the difficulty ratings were given on a five-point scale, ranging from very low to very high.

Results and Discussion

The analysis of the data is divided into two parts: First, we compared the baseline spreadsheet condition to the overall means of all information visualization conditions in order to answer the question whether information visualizations in general are helpful for acquiring knowledge on large data sets compared to a purely text-based representation. In the second analysis, we assessed the effects of dimensionality and color coding by comparing the four information visualization conditions in an ANCOVA (dimensionality x color coding with retentiveness as a covariate, see below).

Do information visualizations foster learning? In a first step, we tested whether subjects achieved higher performance with information visualizations than with a spreadsheet, i.e., here we did not further differentiate between the different kinds of information visualizations. A two-tailed t-test for independent samples showed in fact a higher overall performance for information visualizations (M=20.80) compared to the spreadsheet (M=17.88; $t(98)=2.18$; $p<.05$). However, which kinds of information visualizations produced this effect? To answer this question, each of the four different kinds of information visualizations was compared to the spreadsheet separately (Table 1). Whereas the 2D-conditions were both superior to the baseline (without color coding: $t(38)=2.20$; $p<.05$; with color coding: $t(38)=3.53$; $p<.001$), there were no differences between the 3D-conditions and the spreadsheet condition (without color coding: $t(38)=0.34$; $p=.74$; with color coding: $t(38)=1.28$; $p=.21$).

Table 1: Means for performance, confidence, and cognitive load ratings for the information visualization conditions.

		Information visualizations			
		two-dimensional		three-dimensional-	
		monochrome	with color	monochrome	with color
Performance	**overall performance (35 tasks)**	**21.80**	**23.38**	**18.43**	**19.60**
	relational performance (15 tasks)	9.55	10.15	8.10	8.75
	item-specific performance (15 tasks)	10.35	10.98	8.88	9.10
	structural performance (5 tasks)	2.40	2.85	1.95	2.05
	where- performance (4 tasks)	3.05	3.10	2.70	2.40
	what- performance (4 tasks)	1.60	2.25	1.85	1.80
	varied-performance (4 tasks)	2.40	2.50	1.60	1.75
Confidence	**overall confidence (max. 105 points)**	**72.46**	**75.63**	**67.25**	**64.60**
	confidence correct answers (max. 105 points)	47.48	51.66	35.80	35.65
	confidence wrong answers (max. 105 points)	23.08	21.97	29.45	27.05
Cognitive load	**effort** (max. 5 points)	**3.65**	**3.60**	**4.15**	**3.95**
	difficulty (max. 5 points)	**3.40**	**3.35**	**3.75**	**3.75**

Which representation format of information visualizations is the most suitable for knowledge acquisition? In all analyses of variances reported here, we used retentiveness as a covariate because it was strongly associated with the dependent variables. In a first step we analyzed subjects' overall performance by an univariate ANCOVA (dimensionality x color coding).

Subjects who were presented with a 2D-information visualization outperformed subjects in the 3D-conditions ($F(1,75) = 15.16$; $p<.001$). Additionally, we obtained a marginally significant main effect for color coding in favor of "with color coding" ($F(1,75) = 2.87$; $p<.10$). There was no significant interaction between the two factors. The superiority of the 2D-information visualizations was not only confirmed for overall performance but also for the detailed performance measures - with one exception. There was no significant difference for the what-performance, but this was not astonishing because the information necessary to answer the respective tasks was represented the same way across all conditions. There were no main effects for color coding in the detailed performance measures.

Concerning the overall confidence learners felt regarding the correctness of their answers, we found that subjects learning with 2D-information visualizations were more certain that their answers were correct than subjects in the 3D-conditions ($F(1,75)=8.71$; $p<.01$). Further analysis revealed that learners in the 2D-conditions were not only more convinced that the correct answers they had given were correct ($F(1,75)=18.16$; $p<.001$). Moreover, they also felt more uncertain that their false answer might be correct ($F(1,76)=5.33$; $p<.05$). This pattern of results suggests that subjects in the 2D-conditions had a more accurate assessment of what they really knew. There were no main effects for color coding nor was there an interaction with respect to the overall confidence variable.

With regard to the cognitive load ratings registered after the test phase we found that subjects using 3D-information visualizations indicated that they had to invest more effort into learning than did those in the 2D-conditions ($F(1,76)=4.51$; $p<.05$). In addition, they also evaluated learning as being more difficult than subjects in the 2D-conditions ($F(1,76)=3.30$; $p<.10$). There were no main effects for color coding nor were there interaction effects.

Summary and Conclusions

In our experiment we provided evidence for the suitability of information visualizations for knowledge acquisition. Moreover, we demonstrated that in general 2D-information visualizations are more suitable to foster knowledge acquisition than 3D-ones. This could be due to the fact that learners had to invest more effort and experienced more difficulties during learning in the latter conditions. The question of whether these demands resulted from the necessity to rotate the 3D- information visualization will be addressed in further studies. With regard to the influence of color coding, there were only slight performance increases when information was displayed in color.

Acknowledgements

This work was supported by a scholarship of the DFG. We thank Matthias Grimm of the ZGDV, Darmstadt (Germany) for programming the information visualizations.

References

Amorim, M.-A., Trumbore, B., & Chogyen, P. L. (2000). Cognitive repositioning inside a „Desktop" VE: The constraints introduced by first- vs. third-person imagery and mental representation richness. *Presence*, *9*, 165-186.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*, 293-332.

Hanna, A., & Remington, R. (1996). The representation of color and form in long-term memory. *Memory and Cognition*, *24*, 322-330.

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. London: Erlbaum.

Landau, B., & Jackendoff, B. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brian Sciences*, *16*, 217-265.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65-99.

Mayer, R. E. (2001). *Mulimedia learning*. Cambridge: Cambridge University Press.

Park, S. H., & Woldstad, J. C. (2000). Multiple 2D-displays as an alternative to 3D-displays in telerobotic tasks. *Human Factors*, *42*, 592-603.

Risden, K., Czerwinski, M. P., Munzer, T., & Cook, D. B. (2000). An initial examination of ease of use for 2D and 3D information visualizations of web content. *International Journal of Human-Computer Studies*, *53*, 695-714.

Sweller, J., van Merriënboer, J.J.G., & Paas, F. W.C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296.

Tergan, S.-O. (2003). Managing knowledge with computer-based mapping tools. In D. Lassner & C. McNaught (Eds.), *Proceedings of the ED-Media 2003* (pp. 2514-2517). Honolulu, HI: University of Honolulu.

Treisman, A. (1987). Properties, parts, and objects. In K. R. Boff, L. Kaufman, & F. P. Thomas (Eds.), *Handbook of perception and human performance*. Oxford: Clarendon.

Wiss, U., Carr, D., & Jonsson, H. (1998). Evaluating three-dimensional information visualization designs: A case study of three designs. *Proceedings of the IEEE Conference on Information Visualization* (pp.137-144). London, England.

Learning Domain Structures

Charles Kemp, Amy Perfors & Joshua B. Tenenbaum
{ckemp, perfors, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

How do people acquire and use knowledge about domain structures, such as the tree-structured taxonomy of folk biology? These structures are typically seen either as consequences of innate domain-speci c knowledge or as epiphenomena of domain-general associative learning. We present an alternative: a framework for statistical inference that discovers the structural principles that best account for di erent domains of objects and their properties. Our approach infers that a tree structure is best for a biological dataset, and a linear structure ("left"–"right") is best for a dataset of people and their political views. We compare our proposal with unstructured associative learning and argue that our structured approach gives the better account of inductive generalization in the domain of folk biology.

Psychologists have argued that cognition in different domains draws on qualitatively different mental representations. Tree structures appear well-suited to representing relationships between animal species [1, 2, 10], while a one-dimensional structure (the liberal-conservative spectrum) seems better for representing people's political views. The possibility of different structures raises a fundamental question: how do people learn what kind of structure is appropriate in each domain?

The standard approach to this question is to reject one of its assumptions. Nativists deny that core structures are learned, at least for evolutionarily important domains like folkbiology. Instead, infants come equipped with innate knowledge about which structures are appropriate for which domains. Atran [1], for example, argues that folkbiology is a core domain of human knowledge, and that the tendency to group living kinds into hierarchies re ects an "innately determined cognitive structure." More generally, Keil [8] has argued that ontological knowledge obeys an innate "M-constraint", requiring the extensions of predicates to conform to rigidly tree-structured hierarchies of objects.

Alternatively, empiricists generally deny that structured representations are present at all. Domain-specific representations are merely emergent properties of unstructured, domain-general associative learning architectures. McClelland and Rogers [12], for example, have recently suggested that the acquisition of semantic knowledge in domains such as intuitive biology can be explained as learning in a generic connectionist network. Their architecture never explicitly represents any tree structure, although with repeated training, its hidden unit representations may implicitly come to approximate the taxonomic relations between biological species.

This paper proposes an alternative approach – structure learning – that combines important insights from both of these traditions. Our key contribution is to show how structured domain representations can be acquired within a domain-general framework for Bayesian inference. Like nativists, we suggest that different domains are represented with qualitatively different structures, and we show how these structured representations serve as critical constraints on inductive generalization. Like empiricists, though, we emphasize the importance of learning, and attempt to show how domain structures can be acquired through domain-general statistical inference. This is not only more parsimonious than the nativist position, but allows us to explain the origin of structured representations in novel domains, where the prior existence of domain-specific innate structure is highly implausible.

After describing our structure learning framework, we present two empirical tests of its performance. First, we show that it chooses the appropriate domain structure for both synthetic and real-world data sets. It correctly chooses a tree structure for a biological domain (animal feature judgments), and a linear structure for a political domain (US Supreme Court decisions). Second, we model two classic data sets of inductive judgments in biology [13] and show that our framework performs better than an unstructured connectionist approach.

Bayesian structure learning

Our proposal takes the form of a rational analysis. We aim to demonstrate the computational plausibility and explanatory value of Bayesian structure learning, but leave for future work the question of how these computations might be implemented or approximated by cognitive processes. Assume the learner's data consist of a binary-valued object-feature matrix D specifying the features of each object in a given domain. In biology, for instance, the rows of D might correspond to species, and the columns to anatomical and behavioral attributes. The entry in row i and column j would then specify the value of feature j for species i. Structure-learning includes computational problems at two levels. First, which *structure class* is most appropriate for the domain? Second, given a structure class, which structure

in that class provides the best account of the data?

For instance, suppose that a learner exposed to biological data ends up organizing animal species into a taxonomic tree. The first problem asks how she knew to use a tree rather than some other kind of structure. The second problem asks why she settled on one specific tree instead of the many other trees she might have chosen. Our focus here is on the first problem – the problem of inferring the right structure class for a domain. A solution to the second problem, however, falls out of our probabilistic approach.

We assume that learners come to a domain equipped with a hypothesis space of structure classes, either constructed from innate primitives or based on analogies with previously learned domains. For simplicity, this paper considers a hypothesis space of just three canonical classes: taxonomic trees, one-dimensional (linear) spaces, and independent feature models. People surely have access to other classes, including higher-dimensional spaces, at (non-hierarchical) clusterings, and causal networks. We leave it to future work to characterize the full range of structure classes accessible to human cognition. In particular, it is an open question whether this space is small enough to be explicitly enumerated as we do here, or is so large (perhaps infinite or uncountable) that it can be specified only implicitly through some generating mechanism. Future work should also consider the possibility that multiple structures may apply within a single domain.

Given a set of probabilistic models, Bayesian techniques can be used to evaluate which of the models is most likely to have generated some data [7]. Before these techniques can be applied to inferring domain structures, we need to associate each structure class in our hypothesis space with a probabilistic generative model for the features of objects. The next section defines these models, but here we show how Bayesian inference can be used to choose between them.

Let D be an object-feature matrix generated from one of several structure classes. The posterior probability of each class C_i is proportional to the product of the likelihood $p(D|C_i)$ and the prior probability $p(C_i)$. If we assign equal prior probabilities to each class (as we do throughout this paper), the best class is the class that makes the data most likely.

Computing the likelihood $p(D|C_i)$ requires integrating over all structures $\mathcal{S}$ belonging to structure class C_i:

$$p(D|C_i) = \int p(D|\mathcal{S}, C_i)p(\mathcal{S}|C_i)d\mathcal{S}, \qquad (1)$$

Intuitively, this means that a structure class C_i provides a good account of object-feature data D if the data are highly probable under a range of structures $\mathcal{S}$ in class C_i, and if these structures themselves have high prior probability within C_i. The following section explains how the fit of each structure to the data, $p(D|\mathcal{S}, C_i)$, is computed for several structure classes.

We estimate the integral in Equation 1 using stochastic approximations. First we run a Markov chain Monte Carlo simulation to draw a sample of m structures, $\{\mathcal{S}_j\}$, from the distribution $p(\mathcal{S}|D, C_i)$. We then approximate $p(D|C_i)$ by the harmonic mean estimator [7]:

$$p(D|C_i) = \left(\frac{1}{m} \sum_{j=1}^{m} \frac{1}{p(D|\mathcal{S}_j, C_i)} \right)^{-1}. \qquad (2)$$

This estimator does not satisfy a central limit theorem, and can be thrown off by a sample with very low likelihood. Despite its limitations, it is often sufficient to identify a model that is very much better than its competitors. In future work we plan to estimate these integrals more accurately using path sampling [4].

From structures to probabilistic models

We will work with three probabilistic models, each appropriate for a different structure class, and show how to compute the likelihoods $p(D|\mathcal{S}, C_i)$ for structures in each class. For simplicity we assume here that all features are binary, but our framework extends naturally to multi-valued or continuous features.

C_T: Taxonomic trees

Class C_T is the set of taxonomic trees — rooted trees with the objects in D as their leaves. This is a natural representation when the objects are the outcome of an evolutionary process. We restrict ourselves to ultrametric trees — trees where each leaf node is at the same distance from the root.

Assume that each feature is generated by a mutation process over the tree. We formalize the mutation process using a simple biological model [11]. Suppose that a feature F is defined at every point along every branch, not just at the leaf nodes where the data points lie. Imagine F spreading out over the tree from root to leaves — it starts out at the root with some value and could switch values at any point along any branch. Whenever a branch splits, both lower branches inherit the value of F at the point immediately before the split. Figure 1(a) shows one mutation history for a binary feature on a tree with four objects.

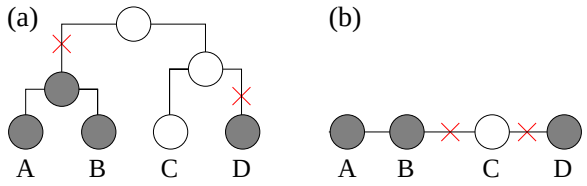


Figure 1: (a) A tree with four objects (A, B, C and D) and three internal nodes. A mutation history for a single feature is shown. The feature is off at the root, but switches on at two places in the tree. Shaded nodes have value 1, clear nodes have value 0, and crosses indicate mutations. (b) A line with four objects.

We formalize this model of mutation using a Poisson arrival process. Under this process, the probability that

F switches values between the beginning and end of any branch b is

$$p(\text{switch along branch } b) = \frac{1 \quad e^{-2\lambda|b|}}{2}, \qquad (3)$$

where $|b|$ denotes the length of b, and λ is the mutation rate. Note that the mutation process is symmetric: mutations from 0 to 1 are just as likely as mutations in the other direction. Asymmetric mutation processes may be more appropriate in some contexts.

Assume that the features are conditionally independent given the tree (i.e., their mutation histories are independent). We can then compute $p(D|\mathcal{T}, C_T)$, the probability of the data given tree $\mathcal{T}$ by multiplying probabilities for each feature vector taken individually. The necessary calculations can be organized efficiently using a Bayes net with the same topology as $\mathcal{T}$ [9].

Computing the total likelihood $p(D|C_T)$ requires integrating over the space of all trees (including variations in branch length and topology), as in Equation 1. We used the MrBayes [6] program for Bayesian phylogenetic inference to draw a sample of trees $\{T_i\}$ from the distribution $p(\mathcal{T}|D, C_T)$. We then estimated the likelihood $p(D|C_T)$ using the harmonic mean estimator (Equation 2).

C_L: One-dimensional (linear) spaces

Although trees seem appropriate for representing biological species and their properties, other domains will have other kinds of structures. Euclidean spaces figure prominently in mathematical models of similarity comparison, judgment, and choice, and probably should appear in any canonical list of structure classes. Let class C_L indicate the set of one-dimensional linear structures. Extensions to higher dimensions are easy in principle, if computationally more demanding.

A line $\mathcal{L} \in C_L$ is a one-dimensional structure where every node corresponds to an object in the domain. A line is a degenerate tree, but unlike the trees of the previous section, lines have no latent nodes. A four-object line is shown in Figure 1(b).

Features are generated over a line according to the mutation model of the previous section. Imagine that Feature F starts at the leftmost node with some value and spreads to the right with the possibility of switching value at any point. Again, the probability that adjacent nodes separated by a branch of length $|b|$ have different values of F is $\frac{1-e^{-2\lambda|b|}}{2}$.

As with C_T, we estimate the likelihood $p(D|C_L)$ with an approximate (MCMC) sum over all linear structures.

C_0: Independent Features

Class C_0 is similar to a null hypothesis. Unlike the previous models, it assumes no underlying relationships between objects in the domain. Each feature is distributed over objects independently of all other features. The pattern of overlap in feature extensions is thus completely unconstrained. More formally, C_0 assumes that feature vectors (columns of D) are generated by ipping weighted coins. Unlike the previous two cases, the likelihood $p(D|C_0)$ can be computed analytically. Suppose that θ_i is the weight of the coin for feature i, and our prior on θ_i is $\theta_i \sim$ Beta(,) (for each of our experiments we use = = 1). If column i of matrix D contains m_i ones and n_i zeros, it can be shown that $p(D|C_0) = \prod_i B(m_i + \ , n_i + \ )/B(\ , \ )$, where $B(\cdot,\cdot)$ is the beta function.

Model complexity and Occam's razor

The three models C_T, C_L, and C_0 vary significantly in their complexity. Both the tree model C_T and the linear model C_L include the independent feature model C_0 as a special case: when each object in C_T or C_L is a long way from its neighbors, feature values at adjacent object nodes are generated in effect by tosses of a fair coin. C_T is also more complex than C_L: in a domain with n objects, there are roughly 2^n more distinct tree structures than distinct linear structures, and the mutation process operating over each tree involves roughly twice as many potential mutation events.

A key feature of Bayesian model selection is that it automatically penalizes unnecessarily complex structures. Some form of Occam's razor is essential when comparing candidate domain structures of different complexities, where the more complex structure (e.g., trees) can more easily mimic the simpler structure (e.g., linear orders) than vice versa. A more naive approach to structure learning, such as choosing the structure that accounts for the most variance in the object-feature matrix D, would be biased against choosing the simpler model class, even when it really generated the observed data.

Empirical tests of structure learning

Synthetic Data

We created three synthetic datasets (unconstrained, tree-structured and linear) with 16 objects and 120 features each. The unconstrained set was constructed using model C_0. The tree-structured set was built by running the mutation process of C_T over a balanced tree with 16 leaf nodes. The linear set was built similarly by running the mutation process over a line with 16 nodes.

Table 1 shows log likelihoods computed for each dataset and structure class. The first row shows that the linear model C_L is better than the tree model C_T on the unconstrained data, but that both are worse than the independent features model C_0. Similarly, the linear model is preferred for the synthetic linear data. The results for the synthetic tree data are more interesting. Even though the data were generated over a tree, the structure class of choice is C_L.

To see why a linear order is a good hypothesis when a tree-structured domain is first encountered, imagine a picture of the true tree, then remove all the branches and internal nodes, leaving behind only the leaves in some linear order. Now join each leaf node to its immediate neighbors. This linear order is a better hypothesis than the true tree at first. The linear model C_L is simpler than the tree model C_T, and if the mutation rate is small, most concepts generated over the tree will be connected subsets of the linear order. Only as more features

Data	C_0	C_L	C_T
Synthetic Unconstrained	<u>59</u>	31	0
Synthetic Linear	0	<u>544</u>	300
Synthetic Tree	0	<u>210</u>	168
Biology	0	230	<u>339</u>
Political	0	<u>1312</u>	883

Table 1: Scaled log-likelihoods for three synthetic and two real-world datasets. Each row has been scaled additively so that its smallest entry is zero.

accumulate should a rational learner conclude that the extra complexity of a tree-structured model is necessary.

To confirm that the true domain structure will eventually win out, we generated a tree-structured set with 32 objects and 240 features and computed log likelihoods as more and more features were observed. Figure 2 shows that the linear structure is preferred while the number of observed features is small, but that the correct tree structure dominates in the end. This transition suggests that our Bayesian model may offer some insight into the dynamics of development. Piaget and others have argued that children move from simple to relatively complex conceptual structures as they mature. Our model shows an analogous shift in tree-structured domains.

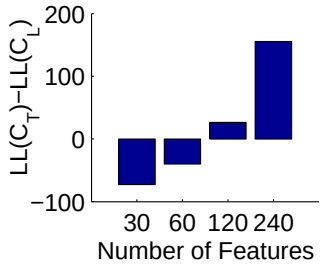


Figure 2: Differences between the log likelihoods of trees (C_T) and linear structures (C_L) on synthetic tree-structured data. Linear structures are preferred at first but the true structure becomes clear as more features are seen.

Biological and Political Data

We used our framework to infer the structure of a biological data set (expected to be tree-structured), and a political data set (expected to be linear). The biological set was constructed from human feature judgments collected by Osherson et al. (1991). Subjects were given 48 animals and 85 features (eg 'lives in water', 'has a tail') and asked to rate the "relative strength of association" between each animal and feature. Subjects gave ratings on a scale that started at zero and had no upper bound. Ratings were linearly transformed to values between 0 and 100, then averaged. We created a binary dataset by thresholding all values at the global mean.

The political dataset was taken from the Supreme Court database collected by Harold Spaeth (1998). We looked at the Burger court which served from 1981 to 1985. Spaeth records 8 possible types of voting behavior: we considered only the cases where every judge either joined the majority, dissented, or cast a regular concurrence (which we treated the same as a majority vote). This left a binary dataset containing votes for 9 judges on 637 cases.

Of the three classes in our hypothesis space, Table 1 confirms that trees provide the best account of the biological data and linear structures are best for the voting data. Note that a more naive approach to structure learning fails here. An additive tree model accounts for more of the variance of the Supreme Court data than a one-dimensional metric scaling solution. Choosing the model that accounts for the greatest proportion of the variance incorrectly favors trees, since it ignores the greater complexity of the tree model.

Once the structure class is known, we can identify the member of that class that makes the data most likely. For the animal data, we took our MCMC sample from the posterior over tree structures, and identified the most representative tree using the `consense` program in the PHYLIP package [3]. The resulting tree is shown in Figure 3(a). Similarly, the best linear structure for the Supreme Court data is shown in Figure 3(b).

The ultimate reason why trees are appropriate for biological data is that evolution is a branching process. It is harder to say a priori why the voting data should be one-dimensional, but the political spectrum ("left"–"right") is an extremely common notion, and others have analyzed Supreme Court data and found that the first dimension of a multidimensional linear model explains almost all of the variance [15]. Our results may explain in part why people represent these domains as they do, but the analysis is mute with respect to the precise mechanisms that give rise to these cognitive structures. Multiple learning mechanisms probably operate in both these domains. Likely mechanisms include inferences drawn from feature observations, as modeled explicitly by our Bayesian learning algorithm, as well as cultural transmission of knowledge, which surely occurs for structures like the "left"–"right" metaphor.

Structure learning versus empiricism

The conventional empiricist critique of structured domain representations has three lines of attack, well articulated recently by McClelland and Rogers [12]: (1) structured representations such as taxonomic trees are too rigid to deal naturally with exceptions or gradients of typicality; (2) it is not clear how structured representations can be induced from raw data; (3) unstructured associative learning architectures can match all of the supposed advantages that structured representations claim. Our work challenges all of these critiques. Previously [10], we showed that robustness to exceptions and sensitivity to typicality fall out naturally from defining a probabilistic generative model of object features in terms of a mutation process over a taxonomic tree (or other domain structure). Point (2) was addressed in the previous section, and now we turn to point (3). We show that learning explicitly structured domain representations provides a powerful source of inductive bias for reasoning about novel properties, and that this power is not easily matched by a generic connectionist architecture.

We compared our tree-structured model for the

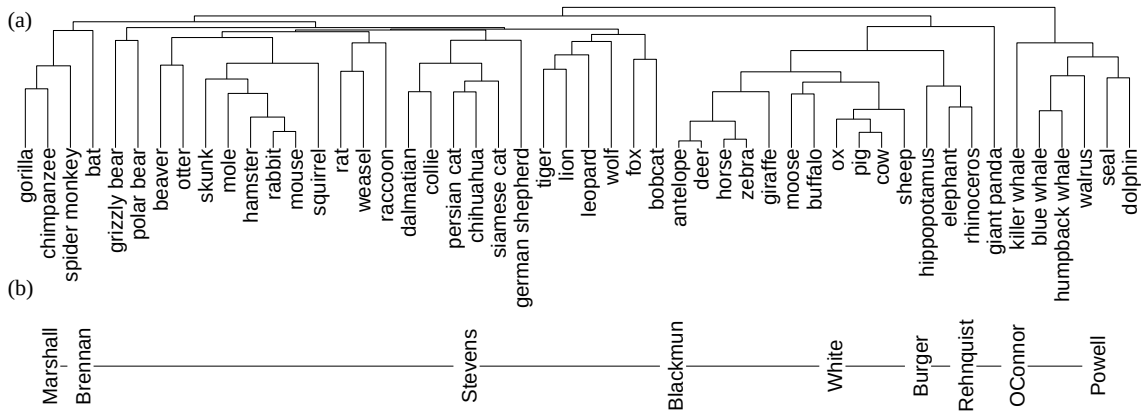


Figure 3: Structures (found via Bayesian structure learning) that best characterize two domains: (a) Mammal species and their properties, and (b) Supreme Court Judges and their decisions.

animal-feature data described above[1] with a connectionist model inspired by the work of McClelland and Rogers [12]. The network includes one input unit for each animal species and one output unit for each feature. We explored a wide range of network parameters in an attempt to achieve the best possible performance (see below). Following McClelland and Rogers, we trained each network on the full matrix D of object-feature associations, then tested how well the hidden-unit representations supported inductive projections for novel features.

In the inductive projection task, a new feature is introduced, and one or more examples of species with that feature are provided to the learner. The learner's task is to infer which other species have this novel property. Like Rogers and McClelland, we modeled this task by introducing a new output unit for the novel feature, freezing all weights except those connected to the new unit, and training the new unit's weights until it reliably produced the correct feature values for the given examples. We then tested the new unit's output when other species were presented as inputs.

We modeled this same induction task using our tree-based Bayesian framework, as described in [10]. Given a tree $\mathcal{T}$ inferred for the domain, the mutation process in model C_T induces a prior distribution over all possible labellings of the species (i.e., the leaves of $\mathcal{T}$). Given one or more examples of a novel property, this prior together with the machinery of Bayesian concept learning allows us to infer the most likely value of that property for all other species in the tree [10]. We used the tree shown in Figure 3(a), and set the mutation rate for the novel property to the value that best fit the 85 features in the biological data set. The resulting tree-based model has no free parameters.

The inductive projections of each model were compared with human argument ratings collected by Osherson et al. [13]. Osherson used a ten-animal domain: horse, cow, chimp, gorilla, mouse, squirrel, dolphin, seal and rhino. The specific set contains 36 two-example arguments, and the conclusion species is always "horse". The general set contains 45 three-example arguments, and the conclusion category is "all mammals." Unfamiliar (blank) predicates – e.g., "have biotinic acid in their blood" – were used for all these arguments. The tree-based Bayesian model rates the strength of general arguments by computing the probability that all ten animals in the domain have the property. The connectionist model rates general arguments by computing projections to each animal separately and adding these ten scores.

Table 2 shows correlations between model predictions and human judgments of argument strength. The first column summarizes the performance of two separate neural networks, re ecting the best performance we ever observed on each data set over a thorough two-stage exploration of the space of possible networks[2]. In the first stage, we tested many different network topologies and varied the learning rate, the number of training and testing epochs, and the presence or absence of momentum and bias. We then took the best-performing networks from the first stage and ran every possible combination of the two best architectures, three best learning rates, two best numbers of testing epochs, and three best numbers of training epochs. The best networks were trained for 20,000 epochs, tested after 250 epochs of training on each testing example, and had no momentum and a bias of -2. They had two hidden layers, typically with 10-30 units each, and a learning rate between 0.005 and 0.01. Even allowing different neural networks for the two datasets, we were unable to match the performance of the tree-based Bayesian model.

Our model differs from these connectionist models along at least two important dimensions, either or both of which could account for its superior performance. First, it uses explicit taxonomic structure and second, it uses Bayesian statistical inference. To isolate the ef-

[1] In order to model the behavioral judgments described below, we supplemented these data with feature ratings for two additional species, cow and dolphin, to give a total of 50 species.

[2] The majority of these tests were conducted with the original 48-animal feature ratings (substituting ox for cow and blue whale for dolphin), before we collected feature ratings for cow and dolphin. Qualitatively similar results were observed with the 50-animal dataset. The results reported in Table 2 re ect the best performance observed across either dataset.

	NN	NN (T)	Bayes (U)	Tree-Bayes	Sim Cov.
Specific	0.62	0.86	0.16	0.95	0.75
General	0.41	0.68	0.38	0.91	0.77

Table 2: Correlations between human judgments and five models for the specific (row 1) and general (row 2) inductive projection tasks described in the text.

fect of structure we implemented models that incorporate only one of these factors. NN(T) is a neural network that uses an explicitly taxonomic representation but not Bayesian inference. The network has 19 input units and a single output unit for the novel property. Input features are derived from the ten-animal tree — the subtree of Figure 3 that includes the ten animals used in this task. Each input node corresponds to a node in the tree, and a species is represented by switching on an input unit for each of its parent nodes in the tree (including a distinctive feature for itself). Species that appear nearby in the tree will share a relatively large number of ancestors and will therefore have similar representations. Bayes(U) is a model that uses Bayesian inference but without any explicit structural representation constraining hypotheses. The model is inspired by Heit's (1998) suggestion that priors for Bayesian induction could be derived from familiar features stored in memory [5]. Each of the 85 observed feature vectors is identified with a candidate hypothesis for generalization, e.g., the feature "nocturnal" gives rise to the hypothesis that the new property is true of all and only the nocturnal species. We assigned a prior probability of $\frac{1}{86}$ to each of these hypotheses and reserved a further $\frac{1}{86}$ for the hypothesis including all mammals.

Table 2 shows that NN(T) performed better than all of the networks explored previously. The tree-based Bayesian model performed better than Bayes(U) or a feature-based version of Osherson et al.'s (1990) similarity-coverage model (which also assumes no domain structure). These results suggest that generic approaches to biological induction may be improved by adding explicit representations of taxonomic structure. The tree-based Bayesian approach also performed better than the tree-based neural network, suggesting that both rational statistical inference and structured domain representations play important roles in guiding people's generalizations.

Conclusion

Our results are preliminary, with a focus on the domain of biology and just the taxonomic aspect of knowledge in that domain. No strong general claims can be made until we push this inquiry more deeply in the domain of biology, and more broadly into other domains. Even so, our work suggests a viable alternative to traditional nativist and empiricist accounts of domain knowledge. Contrary to a strong nativist view, the organizing structural principles of a domain may be learned. Contrary to a strong empiricist view, explicit representations of domain structure may be valuable for guiding inductive projections from sparse data. Structured domain representations and domain-general statistical learning thus need not exclude each other, and indeed are complementary. Statistical learning suggests how novel domain structures can be acquired, and these structures provide a powerful inductive bias for future statistical learning.

Acknowledgments We thank S. Sloman for providing the biological dataset (originally collected by Osherson and Wilkie), Doug Rohde for his neural net package, and Tom Griffiths and Sean Stromsten for helpful suggestions. Supported by NTT Communication Sciences Lab, the DARPA CALO program, and the Paul E. Newton Chair (JBT).

References

[1] S. Atran. Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21:547–609, 1998.

[2] A. Collins and M. R. Quillian. Retrieval time from semantic memory. *Jn of Verbal Learning and Verbal Behavior*, 8:240–247, 1969.

[3] J. Felsenstein. PHYLIP – Phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[4] A. Gelman and X. L. Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

[5] E. Heit. A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford and N. Chater, editors, *Rational models of cognition*, pages 248–274. Oxford University Press, New York NY, 1998.

[6] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[7] R. E. Kass and A. E. Raftery. Bayes factors. Technical Report 254, University of Washington, 1993. Revision 3: July 6, 1994.

[8] F. Keil. *Semantic and Conceptual Development*. Harvard University Press, 1979.

[9] C. Kemp, T. L. Griffiths, S. Stromsten, and J. B. Tenenbaum. Semi-supervised learning with trees. In *Advances in Neural Information Processing Systems*, 2003.

[10] C. Kemp and J. B. Tenenbaum. Theory-based induction. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, 2003.

[11] P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 19(6):913–925, 2001.

[12] J. McClelland and T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4:310–322, 2003.

[13] D. N. Osherson, E. E. Smith, O. Wilkie, A. Lopez, and E. Sha r. Category-based induction. *Psychological Review*, 97(2):185–200, 1990.

[14] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15:251–269, 1991.

[15] L. Sirovich. A pattern analysis of the second Rehnquist U.S. Supreme Court. *PNAS*, 100:7432–7437, 2003.

[16] H. J. Spaeth. United States Supreme Court judicial database, 1953-1996 terms. 1998. 8th ICPSR version.

Key Actions in Insight Problems: Further Evidence for the Importance of Non-Dot Turns in the Nine-Dot Problem

Trina C. Kershaw (tkersh1@uic.edu)
Department of Psychology
University of Illinois at Chicago
1007 W. Harrison St., Chicago, IL, 60607

Abstract

Key actions are single actions or behaviors that can be singled out as leading to the solution of a problem. In the nine-dot problem (Maier, 1930), Kershaw and Ohlsson (2004) proposed that non-dot turns are the key action necessary for solution. In two experiments, non-dot turns are further analyzed as the key action necessary for solving the nine-dot problem and its variants. Non-dot turns are found to be predictive of solution, while the classic conception of drawing lines outside the dots does not distinguish between solvers and non-solvers.

Key Actions in Problem Solving

Problem solving in everyday life, as well as the laboratory, can be quite difficult. Often a problem or activity can seem unduly difficult when one does not know the *key action* necessary for completing the problem. A key action can be defined as a single action or behavior that can be singled out as the key to the solution. Examples of key actions abound in everyday life. A proper roux cannot be made without engaging in continual stirring. Algebra problems become routine once one understands how to balance the equation and isolate the variables. Finally, as I have learned one too many times, data will invariably disappear if I have not completed the key action of backing it up!

In laboratory problem solving, many insight problems can be solved through the production of a key action. For example, using the pliers as the pendulum weight is the key action necessary for solving Maier's (1931) two-string problem, and moving objects in three-dimensional space is necessary for the six matches problem (Scheerer, 1963) and the eight-coin problem (Ormerod, MacGregor, & Chronicle, 2002). As a third example, the key action necessary for solving the prisoner and rope problem (Metcalfe & Wiebe, 1987) is to unravel the rope into two strands, then tie the ends of the two strands together to escape the tower.

Sometimes the key action can be realized without much struggle, depending on an individual's prior knowledge. A friend (and fellow insight researcher) worked on a farm growing up, where the splitting of rope to make it longer was a common occurrence. He instantly knew how to solve the prisoner and rope problem; the key action was easy to discover. In other insight problems, however, the key action is not easy to discover. The common use of pliers hinders their use as a pendulum weight in Maier's (1931) two-string problem (cf. Birch & Rabinowitz, 1951). Additionally, other insight problems may have multiple factors of difficulty preventing the discovery of the key action, such as in the nine-dot problem (Maier, 1930).

Finding the Key Action in the Nine-Dot Problem

The nine-dot problem (Maier, 1930; see Figure 1) is quite possibly the most difficult insight problem that has been studied, with a typical solution rate for unaided participants of 0% (MacGregor, Ormerod, & Chronicle, 2001). Problem solvers are required to connect all the dots in a 3 x 3 matrix by using four straight lines, without lifting their pens from the page or retracing any lines.

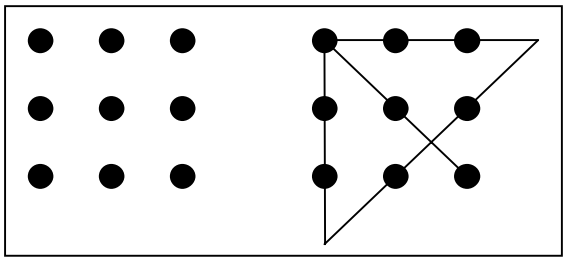


Figure 1: The Nine-Dot Problem and its Solution

The classic conception of the key action necessary for solving the nine-dot problem is that participants should draw lines that extend beyond the dots (Maier, 1930; Maier & Casselman, 1970; Scheerer, 1963). In a related conception, Lung and Dominowski (1985) claimed that the key action was drawing lines that did not begin or end on dots.

Kershaw and Ohlsson (2001) hypothesized that the key action necessary for solving the nine-dot problem was making non-dot turns, or turns that occur in the empty space between dots. The conception of non-dot turns as the key action came about through an inspection of two of MacGregor et al.'s (2001) nine-dot problem variants. The variant with no non-dot turn had a solution rate of 88% after four attempts, while the variant that required one non-dot turn only had a solution rate of 27% after four attempts. Kershaw and Ohlsson (2004) continued in this line of reasoning by explaining that the likelihood of producing a key action is dependent on the cognitive factors that underlie that action. In the nine-dot problem, multiple factors of difficulty are operating that each lower the probability of making a non-dot turn. Kershaw and Ohlsson (2004) distinguished three classes of difficulty: perceptual, knowledge, and process.

Perceptual factors include Gestalt properties of the nine-dot problem such as goodness of figure and figure-ground relationships. Making a non-dot turn requires that one both breaks the good figure of the square and views the white space beyond the dots as part of the problem. Knowledge factors refer to an individual's prior knowledge. Making a non-dot turn is hindered by people's prior experience with dot puzzles, such as connect-the-dot games played by children (cf. Weisberg & Alba, 1981). Process factors include the size of the search space, the specificity of the goal state, and the amount of mental lookahead necessary to find the solution. Making a non-dot turn is difficult because it is not obvious where to draw the first line or what the end state of the problem will be. In addition, people vary in the amount of mental lookahead they possess (cf. MacGregor et al., 2001), which affects the process of making non-dot turns. Kershaw and Ohlsson (2004) showed that perceptual, knowledge, and process factors interact to suppress the probability of producing the key action of non-dot turns in the nine-dot problem.

In the following experiments, non-dot turns are again examined as the key action necessary to solve the nine-dot problem. Experiment I follows up on Kershaw and Ohlsson (2004) but adds an additional possible facilitating factor: giving participants the first line of the solution, which should narrow the search space. Experiment II uses a think aloud methodology to explore what behaviors precede the production of non-dot turns.

Experiment I

Prior research by Kershaw and Ohlsson (2004; Kershaw, Ohlsson, & Coyne, 2003) has shown that increasing the number of non-dot turns leads to greater problem difficulty, such that the more non-dot turns a given problem requires, the harder that problem will be to solve.

Kershaw and Ohlsson (2004; Kershaw et al., 2003) increased solution rates through a training procedure that targeted the multiple factors of difficulty -- perceptual, knowledge, and process -- that hinder the production of non-dot turns. This training procedure was used in Experiment I.

A new facet of the procedure was to give participants the first line of the solution to each target problem. Weisberg and Alba (1981) raised the solution rate of the nine-dot problem to 62% by giving participants the first line in addition to instructing them to go outside of the box set up by the dots. The placement of this first line was chosen based on an analysis by MacGregor et al. (2001). For two of the target problems, the first line extended into the non-dot space.

The addition of the first line influenced the predictions for this experiment. One prediction was that first line would not affect the order nor the magnitude of the solution rates for the five target problems that were reported by Kershaw and Ohlsson (2004), with the 11-dot problem being the easiest and the three-turn problem being the most difficult. A second prediction was that the order of solution rate would remain the same, but that the magnitude would increase for all five problems. A third prediction was that the first line would differentially affect the solution rates for the problems, such that the displaced nine-dot and three-turn problems would show the greatest increase in solution rate due to their first lines cutting into the non-dot space.

Method

Participants and Design One hundred fifty undergraduates from UIC's participant pool participated in the experiment for course credit. No demographic data were collected about the participants.

Participants all received the same training, and one of five target problems.

Materials The first part of the training, the shape training, had a perceptual component in which participants learned to distinguish the shape of the nine-dot problem solution from other shapes (see Kershaw & Ohlsson, 2004; Kershaw et al., 2003). The second part of the training, the dot connecting training, featured problems made of black, filled dots presented on a grid of other unfilled dots as well as problems made of black dots that were alone on the page (see Kershaw & Ohlsson, 2001, 2004; Kershaw et al., 2003). In addition, the training contained a dialogue component in that participants were informed of the purposes of each training task.

The five target problems were taken from Kershaw and Ohlsson (2004; see Figure 2, the nine-dot problem was also used). The problems were modified by adding a diagonal line from the bottom right to the top left of the problem. The placement of the first line was chosen based on an analysis by MacGregor et al. (2001, Experiment 4). Participants were told to treat this line as the first line of the solution that they had to produce.

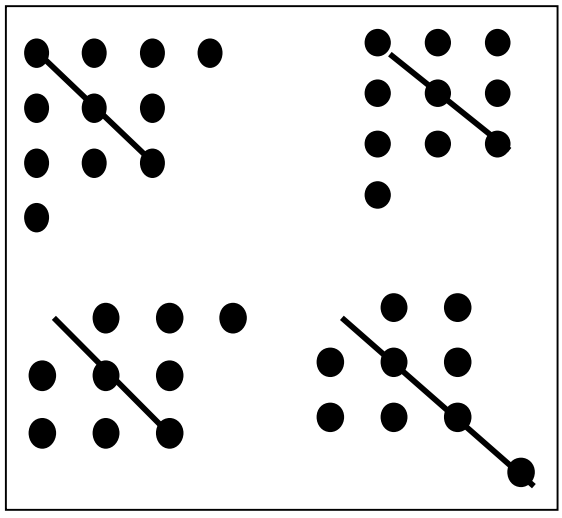


Figure 2: 11-Dot, 10-Dot, Displaced Nine-Dot, and Three-Turn Problems with First Line

Procedure Participants were seen in groups. Participants completed the shape training, and then the dot connecting

training. During the shape training, participants were told that the shape they learned was the shape that would be required to solve the target problem. During the dot connecting training, participants were told that it was necessary to draw lines outside the dots and turn in the empty space between dots. Participants were also shown the correct answer for judging a shape or connecting dots for each judgment or problem that was completed. In addition, they were continually reminded that what they were learning in the training would be applicable to the target problem.

After completing the training, participants attempted one of five target problems (the 11-dot, 10-dot, nine-dot, displaced nine-dot, or three-turn). Participants were given four minutes to connect all the dots using four straight lines. They were instructed to view the line in the problem as the first line, and to draw the remaining lines such that all lines could be drawn without lifting their pens from the page or retracing the lines.

Results

Kershaw and Ohlsson (2004) found the following solution rates for the five target problems (in the training condition): 11-dot, 97%; 10-dot, 80%; displaced nine-dot, 50%; traditional nine-dot, 40%; three-turn, 30%. In contrast, the respective solution rates for this experiment were 83%, 60%, 38%, 40%, and 50% (see Figure 3). However, individual chi-square tests between each problem's solution rate for this experiment and Kershaw and Ohlsson's (2004) data were all non-significant (*p*s > .05).

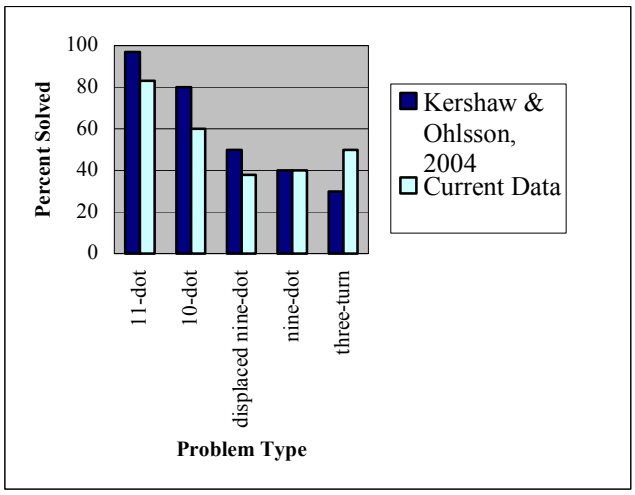


Figure 3: Comparison of Solution Order and Magnitude for the Five Problem Types

Effect of the number of non-dot turns Despite differences in exact solution rates, the new data are similar to those reported by Kershaw and Ohlsson (2004). First, there are overall differences in solution rate between the problems, χ^2 (4, N=150) = 17.02, $p < .05$, $\lambda = .15$. The standardized residuals were examined. The participants who solved the 11-dot problem (25/30 or 83%) caused the greatest standardized residual, 2.2; therefore, this cell made the greatest contribution to the chi-square. When the 11-dot problem was removed from the analysis, the differences between the other problems were not significant, χ^2 (3, N=120) = 4.02, $p > .05$, $\lambda = .09$. Therefore, once a non-dot turn was introduced, all problems became equally difficult.

An alternative way to examine the influence of the number of non-dot turns is to determine the probability of making a non-dot turn (cf. Kershaw & Ohlsson, 2004). The percentage of participants who made any non-dot turns versus correct non-dot turns was calculated for the nine-dot and three-turn problems, both of which require two unassisted (not affected by the first line) non-dot turns. Sixty-five percent (39/60) of the participants made one non-dot turn. Of these participants, 100% (39/39) made two non-dot turns. In contrast, 52% of the participants (31/60) made one correct non-dot turn. Of these participants, 94% (29/31) made two correct non-dot turns.

Effect of drawing lines beyond the dots The nine-dot problem forms a good Gestalt, but the dot groups that make up the other problems do not. The tendency to draw lines that extend beyond the boundary of the dots, the classic explanation of difficulty for the nine-dot problem, was measured across the problem types. The 11-dot problem was excluded from this analysis because drawing lines outside the dots is unnecessary for solution.

Eighty-three percent (99/120) of the participants drew lines outside of the dots. In the 10-dot (24/30), displaced nine-dot (28/30), and three-turn (29/30), participants were equally likely to draw lines outside of the dots, χ^2 (2, N=90) = 5.19, $p > .05$, $\lambda = .07$, despite differences in solution rate. In contrast, participants who attempted the nine-dot problem were less likely to draw lines outside of the dots (18/30). This effect is striking when the nine-dot problem is compared to the three-turn problem, both of which required two unassisted non-dot turns, χ^2 (1, N=60) = 11.88, $p < .05$, $\lambda = .26$. Although the solution rate for these two problems did not differ, the three-turn problem led to a greater rate of drawing lines outside the dots than the nine-dot problem.

Discussion

The results of Experiment I are comparable to Kershaw and Ohlsson (2004) in solution magnitude, the probability of making a non-dot turn, and the prevalence of drawing lines outside the dots. The order of solution rates did differ in that the three-turn problem had the third-highest solution rate in the current data, compared to the lowest solution rate in Kershaw and Ohlsson's (2004) data. However, as in Kershaw and Ohlsson's results, the problems that required non-dot turns did not differ significantly from each other. In addition, individual comparisons between each problem across the two data sets were not significant.

The current data do not support any of the predictions fully. Providing the first line of the solution did not affect the magnitude of solution rate, as predicted, but did affect the order of the solution rate. The solution percentages appear to support the third prediction, that solution magnitude would be affected differentially, but the rate increased for the three-turn problem yet decreased for the displaced nine-dot problem. However, as mentioned above, individual comparisons between the problem types across data sets did not reveal any significant differences.

The current data give further support to the non-dot turn as the key action necessary for solving the nine-dot problem and its variants. As soon as a non-dot turn was introduced, the solution rate dropped by at least 20%. In addition, drawing lines that went outside the dots was not enough to solve the problem. Eighty-three percent of the participants who attempted the 10-dot, displaced nine-dot, nine-dot, or three-turn problems drew lines outside of the dots, but only 47% of the participants correctly solved one of these four problems.

As noted previously, giving participants the first line did not increase the solution rate, as compared to Kershaw and Ohlsson (2004). This finding is interesting compared to similar manipulations used by Weisberg and Alba (1981) and MacGregor et al. (2001). Weisberg and Alba (1981) achieved a solution rate of 62% by giving participants the first line and telling them to go outside the dots. MacGregor et al. (2001, Experiment 4), in contrast, achieved a 6% solution rate after the first 10 attempts, and 47% after 10 additional attempts by giving participants the first line of the nine-dot problem. One explanation, in light of the current data, is that the extensive training used in Experiment I overshadowed any benefit of the first line for the problem variants. Although the solution rate was raised for the three-turn problem, its solution rate was not significantly different than the rate found for the three-turn problem by Kershaw and Ohlsson (2004), nor were there any differences in solution rate across the two experiments for any of the nine-dot problem variants. Untrained participants, in contrast, would most likely benefit from being given the first line, and would thus show differences in comparison to the control group in Kershaw and Ohlsson (2004, Experiment 3).

Experiment II

Experiment I further established non-dot turns as the key action required for solving the nine-dot problem. Experiment I also showed that the classic conception of difficulty for the nine-dot problem, the inability to draw lines beyond the boundary of the dots, did not hold up as a difficulty for the other problem types. However, participants were less likely to draw lines outside the dots for the nine-dot problem, thus supporting the Gestalt factor.

The aim of Experiment II was to examine how participants explore the search space of the nine-dot and 10-dot problems. Both Kershaw and Ohlsson (2004) and Experiment I showed that making non-dot turns is important, but did not show the process that participants go through when making a non-dot turn. Experiment II used a think-aloud methodology to examine the individual thoughts and actions that lead to the making of non-dot turns. Verbal protocols and other trace methods, such as eye movements, have been used effectively to understand the processes involved in achieving insight in problems such as the mutilated checkerboard (Kaplan & Simon, 1990) and in matchstick arithmetic (Knoblich, Ohlsson, & Raney, 2001).

Half of the participants received the training used in Experiment I, and the other half were not trained. The participants received either the nine-dot or 10-dot problem as their target problem. One prediction for Experiment II is that participants who received training will be more likely to solve their target problems, and will show a greater incidence of behaviors that lead to non-dot turns. Based on solution rates found in Experiment I and Kershaw and Ohlsson (2004), no difference in solution rate is expected between the 10-dot and nine-dot problems.

Method

Participants and Design Twenty undergraduates from UIC's participant pool participated in the experiment for course credit. No demographic data were collected about the participants.

The design of Experiment II was a 2 x 2 factorial. The two independent variables were type of training (control and training) and target problem (nine-dot and 10-dot).

Materials The training materials used in Experiment II were the same materials used in Experiment I. The control group did not receive any training. In addition, all participants were given a long division problem as a practice for thinking aloud while solving the target problem. A video camera was used to record each participant's verbalizations and actions.

Procedure Participants were seen individually. As in Experiment I, the participants who received training learned to distinguish the shape of the nine-dot problem solution from other shapes, and learned how to connect dots. They were shown the correct answer for each training exercise and were reminded that the material learned in training would be useful for solving the target problem. Participants in the control group did not receive any training.

Before beginning the target problem, participants practiced thinking out loud by solving a long division problem. Participants were then given four minutes to attempt the target problem. They were told to connect all the dots by using four straight lines, without lifting their pens from the page or retracing any lines. They were instructed to talk out loud while working on the problem. If the participant stopped verbalizing while working on the problem, the experimenter reminded the participant to

continue talking. Each participant's verbalizations, as well as his or her actions, were recorded using a video camera.

Protocol transcription Each participant's verbalizations and actions were transcribed into a verbal protocol by either the author or a research assistant. Protocols were constructed so that the participant's words and actions were grouped together. Actions were described in terms of drawing or simulating lines, and were transcribed by using a map that numbered the dots in each problem.

Results

Effects of training and problem type Solution rates for the problems across training types are as follows: 10-dot training, 60% (3/5); 10-dot control, 20% (1/5); nine-dot training, 40% (2/5); nine-dot control, 0% (0/5). A chi-square analysis was conducted to determine the effect of problem type. There was no significant difference between the number of solvers for the 10-dot and nine-dot problems, χ^2 (1, N=20) = .952, $p > .05$, $\lambda = .13$, as predicted.

A second chi-square analysis was conducted to determine the effect of the training. Participants were more likely to solve their target problem when they had received training than when they had not, χ^2 (1, N=20) = 3.81, $p = .05$, $\lambda = .25$, as predicted.

Analysis of behaviors that lead to non-dot turns Participants' verbal protocols were examined to determine the behaviors that led to making non-dot turns, the key action necessary for solving the 10-dot and nine-dot problems. Based on our previous work (Kershaw & Ohlsson, 2001, 2004; Kershaw et al., 2003), we hypothesized that several actions would show that participants were affected by the training and understood the requirements of the problem: 1) making diagonal lines, 2) making triangle shapes, 3) making the arrow-like shape of the nine-dot problem solution, and 4) making lines that extended beyond the boundary of the dots. For the purposes of this paper, the two actions that will be analyzed are making arrow shapes and making lines that extend beyond the boundary of the dots. In addition, participants' verbalizations may reveal attention to particular areas of the problem, or a rehearsal of strategies.

The participants' verbalizations were surprisingly unhelpful in determining what thoughts preceded making non-dot turns. The majority of participants limited their verbalizations to keeping track of the number of lines they had drawn so far. Only four of the 20 participants verbalized anything about going outside of the dots. Examples of these verbalizations include: "outside the line here" (said while moving a pen from the bottom right dot to the top left dot) and "let's see, I should probably think more about going outside," which was not accompanied by an action.

The use of arrow shapes and lines that extended beyond the dots illustrated the effect of the training in the solution attempts of the participants. Eight participants in the training group made at least one arrow shape, while only three participants in the control group made an arrow shape; this difference was significant, χ^2 (1) = 5.05, $p < .05$, $\lambda = .47$. Likewise, all 10 participants in the training group made lines that extended beyond the dots, while only two participants in the control group attempted such dots. This difference was also significant, χ^2 (1) = 13.33, $p < .05$, $\lambda = .78$.

In addition, these actions were better indicators of events that precede non-dot turns than participants' verbalizations. Participants who solved their target problems drew arrow shapes and extended lines beyond the dots in the correct places before making non-dot turns. In contrast, some participants who did not solve their target problems also drew arrow shapes, but drew them exclusively inside the dots. Other non-solving participants drew arrow shapes and extended lines, but did not make non-dot turns. As in Experiment I, drawing lines that extended beyond the dots was not enough to solve the target problems. Participants needed to extend their lines in the correct places, and make non-dot turns.

Discussion

The results of Experiment II followed up those of Experiment I and Kershaw and Ohlsson (2004) by showing the importance of non-dot turns in solving the nine-dot (and 10-dot) problem(s). In addition, Experiment II showed, like Kershaw and Ohlsson (2004), the effectiveness of training for raising the solution rate for the nine-dot and 10-dot problems.

Experiment II contributes to this line of research by providing a means to analyze the process of attempting the nine-dot problem (or one of its variants). This initial analysis of the verbal protocols revealed that participants who receive training are more likely to produce actions that are necessary for solving the problem, such as drawing an arrow shape, extending a line beyond the dots, and making a non-dot turn. However, as shown in Experiment I and in Kershaw and Ohlsson (2004), making non-dot turns is a difficult key action to execute. Participants must extend lines beyond the dots in the correct place and form the arrow shape of the solution correctly. Merely extending any line beyond the boundary of the dots will not lead to solution.

General Discussion

Key actions can be identified in many different types of problems and in everyday life, from using pliers as a pendulum weight in Maier's (1930) two-string problem to learning to continually stir a roux. In the nine-dot problem, the key action is making a non-dot turn (Kershaw & Ohlsson, 2001, 2004; Kershaw et al., 2003). While some key actions are easily discovered and produced, making a non-dot turn is hindered by interacting factors of difficulty: perceptual, knowledge, and process.

In Experiments I and II, making non-dot turns was compared to the classic conception of the key action necessary for solving the nine-dot problem, drawing lines that extend beyond the dots (Maier, 1930; Maier & Casselman, 1970; Scheerer, 1963). In both experiments, drawing lines outside the dots was not sufficient to solve a target problem. As a striking example, nearly all the participants in Experiment II readily drew lines outside of the dots in the 10-dot, displaced nine-dot, and three-turn problems. However, less than half of the participants actually solved one of these problems.

Experiments I and II provided further support for Kershaw and Ohlsson's (2004) analysis of the importance of making non-dot turns. Other insight and everyday problems are best solved through different key actions. Further study will allow for the identification of these key actions, and the determination of what cognitive factors underlie the production of such actions.

Acknowledgments

Thanks to Stellan Ohlsson for his assistance in planning the experiments and his ever-helpful comments. I also thank Colleen Coyne for her help in collecting some of the data for Experiment I and transcribing some of the protocols in Experiment II.

References

Birch, H.G., & Rabinowitz, H.S. (1951). The negative effect of previous experience on productive thinking. *Journal of Experimental Psychology, 41*, 121-125.

Burnham, C.A., & Davis, K.G. (1969). The nine-dot problem: Beyond perceptual organization. *Psychonomic Science, 17*(6), 321-323.

Chronicle, E.P., Ormerod, T.C., & MacGregor, J.N. (2001). When insight just won't come: The failure of visual cues in the nine-dot problem. *Quarterly Journal of Experimental Psychology, 54A*(3), 903-919.

Kaplan, C.A., & Simon, H.A. (1990). In search of insight. *Cognitive Psychology, 22*(3), 374-419.

Kershaw, T.C., & Ohlsson, S. (2001). Training for insight: The case of the nine-dot problem. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 489-493). Mahwah, NJ: Lance Erlbaum Associates.

Kershaw, T.C., & Ohlsson, S. (2004). Multiple causes of difficulty in insight: The case of the nine-dot problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 3-13.

Kershaw, T.C., Ohlsson, S., & Coyne, C. (2003). The fallacy of single-source explanations: The multiple difficulties of the nine-dot problem. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* [CD-ROM]. Cognitive Science Society.

Knoblich, G., Ohlsson, S., & Raney, G.E. (2001). An eye movement study of insight problem solving. *Memory & Cognition, 29*(7), 1000-1009.

Lung, C.T., & Dominowski, R.L. (1985). Effects of strategy instructions and practice on nine-dot problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 804-811.

MacGregor, J.N., Ormerod, T.C., & Chronicle, E.P. (2001). Information-processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 176-201.

Maier, N.R.F. (1930). Reasoning in humans: I. On direction. *Journal of Comparative Psychology, 10*, 115-143.

Maier, N.R.F. (1931). Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology, 12*, 181-194.

Maier, N.R.F., & Casselman, G.G. (1970). Locating the difficulty in insight problems: Individual and sex differences. *Psychological Reports, 26*, 103-117.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and non-insight problem solving. *Memory & Cognition, 15*(3), 238-246.

Ormerod, T.C, MacGregor, J.N., & Chronicle, E.P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 791-799.

Scheerer, M. (1963) Problem solving. *Scientific American, 208*(4), 118-128.

Weisberg, R.W., & Alba, J.W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. *Journal of Experimental Psychology: General, 110*(2), 169-192.

Fear of Isolation, Cultural Differences, and Recognition Memory

Kyung Il Kim (kyungil@psy.utexas.edu)
Department of Psychology, University of Texas, Austin
1 University Station A8000, Austin TX 78712-0187 USA

Arthur B. Markman (markman@psy.utexas.edu)
Department of Psychology, University of Texas, Austin
1 University Station A8000, Austin TX 78712-0187 USA

Abstract

Previous research suggests that members of East Asian cultures show a greater sensitivity to context (vs. target) information than do members of Western Cultures. We suggest this difference is rooted in a greater chronic fear of isolation (FOI) in East Asians than in Westerners. To support this hypothesis, we first compare chronic levels of FOI between East Asian and Western participants. Then we manipulate FOI in a group of Western college students and assess their recognition memory for object as a function whether the background is the same or different from when the picture was first seen. Consistent with our proposal, the manipulation affected people's sensitivity to background context in picture recognition in a manner consistent with previous studies of cultural differences.

Introduction

Previous research has uncovered cultural differences in reasoning and decision making performance between East Asian and Western populations (Nisbett, Peng, Choi, & Norenzayan, 2001). Clearly the study of cultural differences has practical implications for international commerce and theoretical implications for claims about the universality of cognitive processing.

This work is based on the observation that cognitive and perceptual orientations can differ in the degree to which they are analytic versus holistic. For instance, Masuda and Nisbett (2001) showed Japanese and American subjects pictures of animals and fish with a surrounding background. Later, subjects were shown pictures of animals or fish they had seen as well as new animals and fish appearing either with the same background or in a new background. Japanese (but not American) subjects were more likely to correctly recognize an old animal when it appeared with the original background than when it appeared in a new context.

Findings like this suggest there are significant differences in reasoning between cultures, but only a few studies in cross-cultural research have manipulated potential causal variables in studies. For example, Briley and Wyer (2002) manipulated the degree of group membership and cultural identity in Asian and Western college students. They found that experimentally induced feelings of being part of a group produced a greater preference for equality and compromise in individual choice tasks in both populations. Similarly, Gardner, Gabriel, and Lee (1999) examined the causal role of self-construal by investigating whether priming independent or interdependent self-construals within a culture could result in differences in psychological worldview that mirror those traditionally found between cultures. For instance, in one experiment of the study, they found that European-American participants primed with interdependence displayed shifts toward more collectivist social values and judgments that were mediated by corresponding shifts in self-construal. These studies provide insight into our understanding causes of observed cultural differences (Chiu, Morris, Hong, & Menon, 2000; Hong et al., 2003).

In previous work, we proposed that these cultural differences may be caused by a higher chronic fear of isolation (FOI) in East Asian populations than in Western populations (Kim & Markman, 2003). FOI is the degree to which people are anxious or afraid about being cut off from peers and relatives (Gilbert, Fiske, & Lindzey, 1998; Noelle-Neumann, 1984). Communication theories define FOI as a force that leads people to conceal their views when they believe they are in a minority (Noelle-Neumann, 1984). This pressure is assumed to be related to their fears of being negatively evaluated by others. The theory maintains that mass media works simultaneously with majority public opinion to silence minority beliefs on cultural issues. On this view, FOI prompts those with minority views to examine the beliefs of others and to conform to what they perceive to be a majority view. In this paper we discuss a difference in chronic FOI between cultures and then present a study that addresses the relationship between FOI and attention/memory respectively.

Different sensitivities to FOI between East Asian and Western culture

Before discussing how a difference in degree of FOI can influence judgment and decision making, we must first show that members of different cultures are likely to differ in their chronic level of FOI.

As a measure of FOI, we used the Fear of Negative Evaluation scale (Watson & Friend, 1969).[1] This 30-item instrument measures social anxiety about receiving negative evaluations from others. Scores on this scale reflect a fear of the loss of social approval. Items on the measure include signs of anxiety and ineffective social behaviors that would lead to disapproval by others. We gave this scale to 41 Asian students enrolled in University of Texas Austin and their spouses participated for the measurement of FOI in East Asian population. The participants were all born in East Asia (31 Korean 6 Japanese, 4 Chinese) and had a native language other than English. The length of stay in the US was less than 5 years before their participation ($\underline{M}$ = 3.1 years). Western participants were 49 undergraduate students at the University of Texas (all born in the US). Both groups filled out Fear of Negative Evaluation scale along with questions for demographic information. The East Asian Group showed significantly higher scores on the FNE ($\underline{M}$ = 17.54) than did the Western group ($\underline{M}$ = 11.54), $\underline{t}$ (88) = 11.56, $p < .01$. This finding supports the proposal that members of East Asian culture have a higher chronic FOI than do members of Western culture.

The social anxiety literature provides some insight about why different cultures have different levels of FOI. Cross-cultural differences in social anxiety have been studied in various ways, and the consensus among researchers is that members of relatively society-oriented cultures have more social anxiety than do those in individual-oriented cultures. For example, Okazaki and colleagues (Okazaki, 1997; Okazaki, Liu, Longworth, & Minn, 2002) found that Asian-American report higher distress on various measures of social anxiety. A practical merit of such studies is that they enable a comparison of two cultures controlling out other confounding variables such as language or culture-specific patterns in reporting /interviewing. For this reason these results provide insight into chronic differences in social anxiety between cultures. Other cross-cultural comparisons assessed the difference in social anxiety between Asian and Western populations. For example, Sato and McCann (1998) studied Japanese and American students and found a positive relation between social anxiety and interdependent self-construals (which are typical of collectivistic cultures). Similarly, Dinnel, Kleinknecht, and Tanaka-Matsumi (2002) shows that TKS (Taijin Kyofusho, a Japanese variant of social anxiety) - like symptoms (e.g., fear of offending others) were more likely to be reported Japanese university students than by their American counterparts. It is unlikely that social anxiety discussed in the current study is more related to other subcomponents then FOI.

Given that members of East Asian and Western culture differ in chronic FOI, it is important to discuss why this difference might lead to differences in reasoning. Social anxiety, especially FOI, motivates people to focus on social activity, to interact with other members in the social network and to consider others' responses seriously (Gilbert, 2001; Scheufele, Shanahan, & Lee, 2001). Thus, members of collectivistic cultures are expected to be generally more interested in relations among items in the environment than do members of individualistic cultures (Morris & Peng, 1994; Nisbett et al., 2001). It is also possible to observe such differences within a single culture by manipulating a potential cause. For example, as discussed earlier in the previous section, Gardner (1999) showed that manipulation of self-construals by priming interdependence induced a more collectivistic thinking in Western participants. Note that such patterns of behavior and thinking caused by primed interdependence are consistent with observed patterns rooted in a greater level of FOI (Gilbert et al., 1998; Noelle-Neumann, 1984).

In sum, social anxiety is higher in Eastern than Western populations. Increased social anxiety leads to increased attention to relations among items and to context. We connected these two and suggested that levels of FOI are positively related to the degree of dialectical thinking which has been treated as characteristic reasoning mode of collectivism culture (Kim & Markman, 2003).

We developed this idea further in the current study by examining the influence of a manipulation of FOI on recognition memory. As discussed above Masuda and Nisbett (2001) found that members of collectivist culture were more holistic in their analysis of scenes than were members of an individual culture. If a high level of FOI indeed makes people to attend to interpersonal relationships (and more broadly to relations between objects and their environments), then inducing a high level of FOI should make Americans less likely to attend to target information, which in turn should increase their memory for context vs. target information.

Manipulation and measurement of FOI

In our studies, FOI was manipulated as an independent variable following the method used by Kim and Markman (2003). Participants in the High

[1] There are other scales that have been used to measure FOI, but these scales also ask questions about physical rather than social isolation.

FOI group described experiences in which they were socially isolated from others (e.g., "you might have been anxious once when your friends were not talking to you at all, or when you went to a new place where you didn't know anyone and had difficulty meeting new people"). The Low FOI group described experiences in which they caused someone else to be socially isolated from other (e.g., "you might have been at a party and you didn't talk to one of your friends who did not know many people at the party and you felt bad about it later"). Many clinical techniques such as prolonged exposure treatment that is used to treat post-traumatic stress disorder are based on the premise that asking a patient to recall and describe a previous experience and associated emotion will activate and retrieve relevant feelings and memories, and put the person into that state again (Foa, Cashman, Jaycox, & Perry, 1997). Then we measured a person's FOI with the Fear of Negative Evaluation (FNE) scale as a manipulation check.

Experiment

Method

Participants

Eighty nine American undergraduate students (all born in the US) of the University of Texas participated in the study. Half of participants were randomly assigned to the High FOI condition and the other half were to the Low FOI condition.

Materials

In the first phase of the study, 24 animal pictures were presented. Each picture has an animal and a particular background (see Figure 1).

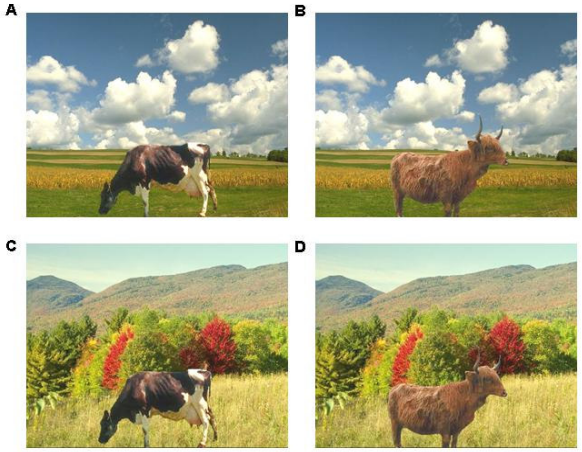


Figure 1. Sample pictures used in this study. (A) A study picture. (B) A new animal in the old background. (C) An old animal in a new background. (D) A new animal with a new background.

In the second phase, participants saw 96 pictures. 24 of them were same with the pictures seen in the first phase (old animal and old background). To create the rest of the 72 pictures, an additional 24 animals (new animal) and 24 backgrounds (new background) were used and the combination between the animal and the background information was manipulated. Because each animal could have one of two different backgrounds – the original background or a novel background, there were four different conditions: (a) old background and old animal, (b) old background and new animal, (c) new background and old animal, and (d) new background and new animal (see Figure 1). All of these pictures were used in Masuda and Nisbett's (2001) study.

Procedure

Participants were asked to describe their previous experiences relating to an anxiety producing situation. In the High FOI condition, participants wrote about being socially isolated from others. In the Low FOI condition, participants wrote about socially isolating someone else from them or other people. After completing this self-descriptive priming task, participants in both conditions responded to the Fear of Negative Evaluation scale as a manipulation check.

Then participants viewed 24 photos of animals in naturalistic environments. After a 2-minute delay, participants viewed 96 photos in a recognition memory test that varied whether the animals were old or new and whether the background was old or new. Subjects responded whether they had seen the animal in the photo regardless of the background of the test photo.

Results

First, we checked the effectiveness of our FOI manipulation. Average values on the Fear of Negative Evaluation scale were significantly higher in the High FOI condition ($\underline{M}$ = 15.61) than in the Low FOI condition ($\underline{M}$ = 10.60), $\underline{t}$ (87) = 3.92, $p < .01$. Note that the mean score of High FOI group approaches that seen in the East Asian group ($\underline{M}$ = 17.54) we measured.

For "old" and "new" responses, we subtracted people's accuracy for the pictures with the new background from their accuracy with the original background. Positive scores indicate sensitivity to the context.

The pattern of data in this study shows the same pattern observed by Masuda and Nisbett (2001) (see Figure 2). There was a significantly higher index ($\underline{M}$ = 2.39) for the High FOI condition than for the Low FOI condition ($\underline{M}$ = 1.22), $\underline{F}$ (1, 87) = 6.01, $\underline{p}<.05$. This effect is mediated by level of FOI. In an ANCOVA including FNE score, there is a significant correlation between FNE and the response index ($\underline{r}$ = .33, $\underline{p}$ <.01) and the main effect of FOI is reduced in significance, $\underline{F}$ (1, 87) = 1.91, $\underline{p}$ = .171.

For "new" responses there is also a marginally significant difference between the High (M = -1.43) and Low FOI conditions (M = -2.01), F (1, 87) = 1.18, p<.28. (Masuda and Nisbett (2001) found no significant difference between their Japanese and American subjects for "new" responses.) This finding is consistent with the hypothesis that participants in the High FOI condition attend more on background information than do those in the Low FOI group even when misleading cues of original background interfered with recognition.

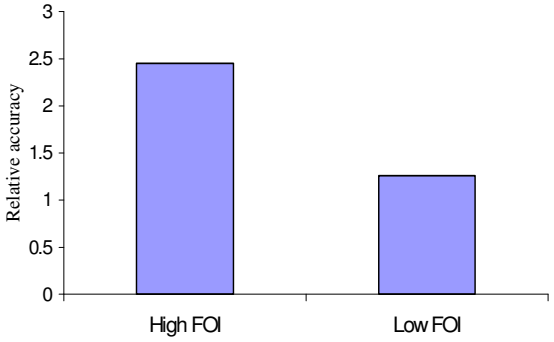


Figure 2. Participants' relative accuracy for "old" response.

Interestingly, this combination of results does not indicate greater overall accuracy between groups. A calculation of d' shows no difference for the High FOI condition (M = 1.74) and the low FOI condition (M = 1.76), F (1, 87) = .01, p = .916.

General discussion

This experiment demonstrated the influence of fear of isolation on attention and memory. Inducing a higher level of FOI in American college students made their attention more similar to that of East Asian students observed in previous studies. Participants in the High FOI condition showed greater accuracy for the memory of background information than did those in the Low FOI condition. When Fear of Negative Evaluation scale values were incorporated into the analyses as a covariate, they were significantly related to the degree of sensitivity to context, and the strength of the effect of FOI manipulation was decreased.

However, an alternative interpretation of the current results is that the induction of low FOI primarily induced the feeling of guilt, and hence a more *negative mood* than in the condition designed to induce high FOI. Some previous studies showed that negative mood leads to more analytic thinking (Bolte, Goschke, & Kuhl, 2003). For example, according to the personality systems interaction theory (Bolte et al., 2003; Kuhl & Kazen, 1999), an increase in negative affect supports a analytic processing mode whereas positive affect induce a relatively more holistic thinking. We tested this possibility in another study, in which participants' relative preference for dialectical proverbs were measured, and found that there was no meaningful difference on emotion scales (e.g., hedonic tone and general arousal) between the two FOI groups and that only FNE scale values explained the variation in the relative preference for dialectical proverbs within/between group (Kim & Markman, in preparation). Thus it is unlikely that emotion systematically influenced the current results.

It is also important that, as discussed earlier, East Asian participants exhibited a significantly greater FOI than did Western participants. Note that East Asian without FOI manipulation showed a greater average values on the Fear of Negative Evaluation scale than did those in the High FOI group in the current experiment.

These findings are consistent with the hypothesis that chronic differences in FOI in members of East Asian and Western cultures lead to the differences in attention observed in the previous studies (Masuda & Nisbett, 2001). We are not claiming that FOI is the only cause of cultural differences in reasoning. Indeed, differences in culturally accessible concepts such as collectivism and individualism may influence cognition either by affecting level of FOI through some other route (Aaker & Lee, 2001; Hsee & Weber, 1999). This issue has been much discussed in communication theories, which have yielded no clear consensus on whether FOI is an antecedent or an intervening variable. For example, Shoemaker, Breen, and Stamper (2000) tested whether FOI is antecedent to opinion formation or an intervening variable between opinion formation and willingness to voice the opinion. Their path analysis suggested that FOI is an antecedent variable, but they could not exclude possibility that it is an intervening variable. However, it seems that FOI is a robust causal factor explaining previously observed difference between cultures. Kim and Markman (2003) found that a manipulation of FOI induced a difference in degree of dialectical reasoning.

Thus, chronic levels of Fear of Isolation may be a causal factor underlying observed cultural differences in reasoning. The mechanisms that relate FOI to these reasoning differences will be the subject of future research, but we speculate that high levels of FOI lead people to think more about their relationship to others, and hence are more open to compromise in reasoning and more attentive to contextual and situational factors that guide behavior.

The current study also has implications for Cognitive Science in general. Most behavioral research assumes that the average performance of participants

reflects the basic functioning of the cognitive architecture. However, work on cultural differences points out dimensions along which performance on cognitive tasks may reflect learned strategies rather than constraints of the cognitive architecture itself. In line with this viewpoint, the study we present in this paper indicates that “some” of the observed cultural differences may reflect straightforward differences in chronic social anxiety rather than fundamental differences in knowledge gathered over years of experience within a culture.

Finally, it is important to bear in mind that we induced significant differences in memory for objects based on a simple manipulation of a participant's level of fear of isolation. As these findings demonstrate, a straightforward change in motivational state can lead to a large difference in basic cognitive functioning. This work highlights the need to include more research on the influence of motivation on cognitive processing within the canon of research in Cognitive Science.

Acknowledgements

This research was supported by NIDA grant NIH #1 R21 DA015211-01A1 to the second author. The authors thank Hunt Stilwell, Lisa Narvaez, Levy Larkey, Serge Block, Nathan Janak and Leora Orent for helpful discussion about and/or aid in conducting experiments. They also thank Takahiko Masuda and Richard Nisbett for their providing experimental materials for the current study.

References

Aaker, J. L., & Lee, A. Y. (2001). "I" seek pleasures and "we" avoid pains: The role of self-regulatory goals in information processing and persuasion. Journal of Consumer Research, 28(1), 33-49.

Bolte, A., Goschke, T., & Kuhl, J. (2003). Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. Psychological Science, 14(5), 416-421.

Briley, D. A., & Wyer, R. S. (2002). The effect of group membership salience on the avoidance of negative outcomes: Implications for social and consumer decisions. Journal of Consumer Research, 29(3), 400-415.

Chiu, C. Y., Morris, M. W., Hong, Y. Y., & Menon, T. (2000). Motivated cultural cognition: The impact of implicit cultural theories on dispositional attribution varies as a function of need for closure. Journal of Personality and Social Psychology, 78(2), 247-259.

Dinnel, D. L., Kleinknecht, R. A., & Tanaka-Matsumi, J. (2002). A cross-cultural comparison of social phobia symptoms. Journal of Psychopathology and Behavioral Assessment, 24(2), 75-84.

Foa, E. B., Cashman, L., Jaycox, L., & Perry, K. (1997). The validation of a self-report measure of posttraumatic stress disorder: The Posttraumatic Diagnostic Scale. Psychological Assessment, 9(4), 445-451.

Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). "I" value freedom, but "we" value relationships: Self-construal priming mirrors cultural differences in judgment. Psychological Science, 10(4), 321-326.

Gilbert, D., Fiske, S. T., & Lindzey, G. (1998). The handbook of social psychology (4th ed.). New York: McGraw-Hill.

Gilbert, P. (2001). Evolution and social anxiety. The role of attraction, social competition, and social hierarchies. The Psychiatric Clinics Of North America, 24(4), 723-751.

Hong, Y. Y., Chan, G., Chiu, C. Y., Wong, R. Y. M., Hansen, I. G., Lee, S. L., Tong, Y. Y., & Fu, H. Y. (2003). How are social identities linked to self-conception and intergroup orientation? The moderating effect of implicit theories. Journal of Personality and Social Psychology, 85(6), 1147-1160.

Hsee, C. K., & Weber, E. U. (1999). Cross-national differences in risk preference and lay predictions. Journal of Behavioral Decision Making, 12(2), 165-179.

Kim, K. I., & Markman, A. B. (2003). The effect of cultural differences in fear of isolation on dialectical reasoning. Paper presented at the 27th Annual Conferences of the Cognitive Science Society, Boston, USA.

Kim, K. I., & Markman, A. B. (in preparation). Differences in fear of isolation as an explanation of cultural differences: Evidence from memory and reasoning.

Kuhl, J., & Kazen, M. (1999). Volitional facilitation of difficult intentions: Joint activation of intention memory and positive affect removes Stroop interference. Journal of Experimental Psychology-General, 128(3), 382-399.

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. Journal of Personality and Social Psychology, 81(5), 922-934.

Morris, M. W., & Peng, K. (1994). Culture and Cause - American and Chinese Attributions for Social and Physical Events. Journal of Personality and Social Psychology, 67(6), 949-971.

Nisbett, R. E., Peng, K. P., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. Psychological Review, 108(2), 291-310.

Noelle-Neumann, E. (1984). The spiral of silence : public opinion, our social skin. Chicago: University of Chicago Press.

Okazaki, S. (1997). Sources of ethnic differences between Asian American and White American college students on measures of depression and social

anxiety. Journal of Abnormal Psychology, 106(1), 52-60.
Okazaki, S., Liu, J. F., Longworth, S. L., & Minn, J. Y. (2002). Asian American-white American differences in expressions of social anxiety: a replication and extension. Cultural Diversity & Ethnic Minority Psychology, 8(3), 234-247.
Sato, T., & McCann, D. (1998). Individual differences in relatedness and individuality: An exploration of two constructs. Personality and Individual Differences, 24(6), 847-859.
Scheufele, D. A., Shanahan, J., & Lee, E. (2001). Real talk - Manipulating the dependent variable in spiral of silence research. Communication Research, 28(3), 304-324.
Shoemaker, P. J., Breen, M., & Stamper, M. (2000). Fear of social isolation: testing an assumption from the spiral of silence. IRISH communication review, 8, 65-78.
Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. Journal of consulting and clinical psychology, 33(4), 448-457.

Asymmetries in the Bidirectional Associative Strengths Between Events in Cue Competition for Causes and Effects

Deanah Kim (kmkim@usc.edu)
Department of Psychology, University of Southern California
Los Angeles, CA 90089-1061

Stephen J. Read (read@usc.edu)
Department of Psychology, University of Southern California
Los Angeles, CA 90089-1061

Abstract

Two experiments using social stimuli tested a recurrent neural network model's predictions for cue competition for causes and effects. The delta-rule based model predicts the presence of cue competition for effects as well as for causes as a result of an asymmetry in the bidirectional associative strengths between the relevant cue-outcome pairs. This model can capture cue competition for effects when cues are encountered in the cause-effect direction, unlike associative and feed-forward models. Results support the model's prediction of cue competition for both effects and causes. The implications of these results for causal model theory and for various associative accounts of cue competition are discussed.

Introduction

Since the advent of research on causal induction many researchers have focused on causal models that explain the competitive nature of learning cues that predict or indicate the occurrence of an event (e.g., Rescorla & Wagner, 1972; Gluck & Bower, 1988; Shanks, 1991; Waldmann & Holyoak, 1992). The first account of competitive learning of cues was the Rescorla-Wagner (1972) model. When multiple cues are present preceding an event, these cues compete with each other for predictive strength, resulting in the competitive learning of cues. A classic example of this is blocking, whereupon learning that stimulus A predicts outcome X during initial training, there is a deficit in learning that B also perfectly predicts X when AB are presented together preceding X in the second training phase.

Although cue competition of causes is well established in both the animal and human causal learning literature, the question of whether competition occurs among multiple effects of a common cause has produced somewhat inconsistent findings and has resulted in a heated debate (e.g., Shanks, 1991; Waldmann & Holyoak, 1992; Van Hamme & Wasserman, 1993; Price & Yates, 1995; Matute, Arcediano & Miller, 1996; Shanks, & Lopez, 1996). Cue competition for effects describes a two-stage conditioning phenomenon whereupon first learning that cause A perfectly predicts the occurrence of effect X during Phase 1 of training, there is a deficit in subsequently learning that cause A also perfectly predicts the occurrence of effect Y when the XY compound is presented together after the presentation of cause A during the Phase 2 training.

A number of researchers have provided evidence for cue competition for effects. Some researchers (e.g., Shanks, 1991; Shanks & Lopez, 1996; Price & Yates, 1995; Cobos, Lopez, Cano, Almaraz & Shanks, 2002) interpret the findings as consistent with associative learning theories. Others (e.g., Matute et al., 1996; Miller & Matute, 1998) assert that the findings are more consistent with contiguity theory, which assumes that associations are learned noncompetitively and bi-directionally through simple contiguity, and that cue competition takes place at judgment.

In contrast, proponents of causal model theory (e.g., Waldmann & Holyoak 1992; Melz, Cheng, Holyoak & Waldmann, 1993; Waldmann, 2000) deny the evidence for cue competition for effects, suggesting that whereas causes compete, effects do not. They argue that multiple effects should not compete with each other because they provide new information about the effects of a common cause.

The goal of this paper is to contrast the predictions of a recurrent neural network, with associative learning theory, causal model theory and contiguity theory's predictions for cue competition for causes as well as for effects. Two experiments designed to test the different predictions for the occurrence of cue competition between effects will be presented. These experiments use various social behaviors as target stimuli for cues and outcomes in an attempt to extend the research in multiple cue contingency learning beyond the traditional settings of biological, physical or abstract events and their consequences.

In addition, these studies assess both directions of reasoning between causes and effects. Typical research in this domain only investigates one direction of reasoning and thus cannot say anything about the relative strength of the forward link from causes to effects and the backward link from effects to causes.

Cue Competition in Associative Models

The Rescorla-Wagner (R-W) model (Rescorla & Wagner, 1972) formally describes the change in associative strength during learning by: $\Delta V_{cs(n)} = \alpha_{cs}\beta_{us(n)}(\lambda_{us(n)} - V_{total(n)})$, where $\Delta V_{cs(n)}$ is the change in the associative strength (V) of CS as a result of a pairing with US on trial n; α_{cs} is the learning rate parameter of the CS; $\beta_{us(n)}$ is the learning rate parameter of the US on trial n; $\lambda_{us(n)}$ is the asymptote of learning or the maximum associative strength that the US can support on trial n; and $V_{total(n)}$ is the sum of associative strengths of all CSs present on trial n, or the extent to which the US is predicted on trial n. The basic principle behind the R-W model is that associative learning is determined by the extent to which an US is surprising, represented by the difference

between the US that is actually presented on trial n and the US that is expected on the basis of the summed predictive value of all the cues that are present on trial n.

With respect to blocking and cue competition, the R-W model predicts the consequences of presenting multiple causes. Cue competition for causes is observed in blocking experiments because by the end of Phase 1 training, the animal has learned that CS_1 perfectly predicts the US. During Phase 2 training, when CS_1 and CS_2 are presented together with the US, no learning occurs for CS_1CS_2 because changing the associative strength of the CS_2 cannot improve the already perfect predictability of the US.

However, the R-W model is unable to naturally handle cue competition between multiple effects of a common cause, because it is a predictive model that assumes a cause-to-effect directionality in learning associations. In other words, the difference term (– V_{total}) only applies to the ability of cue (CS) to predict outcome (US), but not outcome to cue. However, several researchers have obtained cue competition for effects and have provided associative accounts for them (e.g., Shanks, 1991; Shanks & Lopez, 1996; Price & Yates, 1995; Cobos et al, 2002). In order to do so, these researchers have had to resort to the somewhat convoluted procedure of presenting participants with the effects *preceding* the cause (for instance, where symptoms predict the disease that caused them). In other words, the R-W model can accommodate cue competition for effects by using a diagnostic learning procedure, where the multiple effects can be presented as antecedent events, which are understood to occur after their cause, even though the effects are presented prior to the cause. Thus the effects (antecedents) compete with each other in predicting the cause. However, the R-W rule cannot handle the more typical situation in which causes occur before effects.

Gluck and Bower (1988) and Shanks (1991) demonstrated that a simple two layer feedforward network using the delta rule, which is closely related to the R-W rule, correctly predicted competition of cues (symptoms) for a common outcome (a rare disease). The delta rule is an error-correcting learning rule that says that the changes in weights, Δw_{ij}, from input node i to output node j is given by the following equation: $\Delta w_{ij} = \partial (t_j - o_j) a_i$, where a is the activation on input node i; t is the target activation on output node j; o is the observed or actual activation on output j; and ∂ is the learning rate (constant). As in the R-W rule, the change in weight between the input and output nodes, Δw, depends on the extent to which the target activation of the output differs from the observed activation of the output.

However, this network can only learn forward links from cues to outcomes. Thus, as with the R-W rule they could capture cue competition for effects only by assuming diagnostic learning where the effects precede the cause.

We will show that a recurrent network model with delta-rule learning does not have this limitation, but can handle cue competition for effects when causes precede effects.

Cue Competition in Causal Model Theory

Causal model theory argues that people use meaningful world knowledge about the basic characteristics of causal relations in conjunction with contingency information to build causal models of the relations between causes and effect (Waldmann & Holyoak, 1992). The causal model theory uses a contingency rule to deal with a multiple cue situation, where the contingency is the difference between the proportion of cases in which the effect and cause co-occur and the proportion of cases in which the effect occurs in the absence of the cause. When the causal model is predictive, cue competition between causes is expected in the classic blocking paradigm because during Phase 2 training, the new cue, Cue_2, always co-occurs with first cue, Cue_1. Waldmann and Holyoak (1992) argue that because "it is impossible to determine whether the observed unconditional contingency between Cue_2 and the effect is genuine or spurious," this should lead to uncertainty, which should further lead to a lowered predictiveness of Cue_2, or partial blocking (p. 226). On the other hand, they assert that effects do not compete with each other, because each effect provides further information about the cause and there is no uncertainty (Waldmann & Holyoak, 1992; Waldmann, 2000).

Waldmann has done a number of studies that fail to find cue competition for effects (with the exception of Study 2 in Waldmann and Holyoak (1992)). However, he typically uses complicated learning tasks where effects temporally precede causes (diagnostic learning). This has apparently been motivated by the necessity of using diagnostic learning to compare the predictions of the R-W rule.

Cue Competition in Contiguity Models

Some researchers propose that a noncompetitive, contiguity theory of learning may better accommodate cue competition for effects by asserting that cue competition does not arise during learning, but during later judgment (e.g., Matute et al., 1996). Further, Matute et al. (1996) found that the wording of test questions moderates the observance of cue competition for effects. They obtained cue competition for causes when they used test questions that implicitly probed the conditional probability of an effect given a cause compared to its probability given an alternative cause (p[E|C] with p[E|C']), and they found cue competition for effects when they probed the conditional probability of a cause given an effect compared to its probability given an alternative effect (p[C|E] with p[C|E']). Their work seems to suggest that the direction of the relationship queried may be related to whether evidence is found for cue competition.

Cue competition in Recurrent Neural Networks

Recently, Read (2003) demonstrated that a recurrent network, based on McClelland and Rumelhart's (1988) auto-associator, with bidirectional links between the input and output nodes and using a modified version of the delta rule, can predict cue competition for both causes and effects. Unlike a feed-forward model, the recurrent model acquires bidirectional links or associations between the input and output nodes, and thus is able to accommodate cue competition for effects with predictive learning, where the cause *precedes* the effects.

One of the reasons this literature has become so confusing is that in order to use the R-W rule or a feed-forward network as a model of cue competition for effects, one must test cue competition for effects with diagnostic learning,

where effects are encountered before causes. However, with the recurrent network with delta-rule learning, this is not necessary. The current model allows one to test an associative model, closely related to R-W, with the more natural situation in which causes temporally *precede* effects.

In a recurrent neural network, the associative strengths of the bidirectional links between any two events may differ, and this possible asymmetry can be illustrated in the classic blocking paradigm. In the recurrent model, the observation of blocking depends on the direction of the association between the redundant cause B and the common effect X. If the association between cause A and X is trained in phase 1, then when A and B are subsequently paired preceding X, the link from B→X should exhibit competitive learning because, when B is activated, X is already activated due to the simultaneous presence of the previously trained cause A, which is a perfect predictor of X. Thus, there is little change in the link from B→X during Phase 2 training because B does not provide any new information about X. However, the link from X→B should not exhibit cue competition during Phase 2 training, as activating X does not predict B because B is not initially predicted by anything. Therefore, there is a great discrepancy between the target activation of B and the actual activation of B, which results in a greater weight change in the link from X→B. The result is that the link from X→B will be stronger than the link from B→X, suggesting that cue competition should only be observed in the link from B→X. As seen in Figure 1, Read demonstrated this asymmetry in weights using the recurrent model, with delta rule learning, with a learning rate of .15, with 10 passes through 10 learning instances with the same contingencies as in the current experimental stimuli.

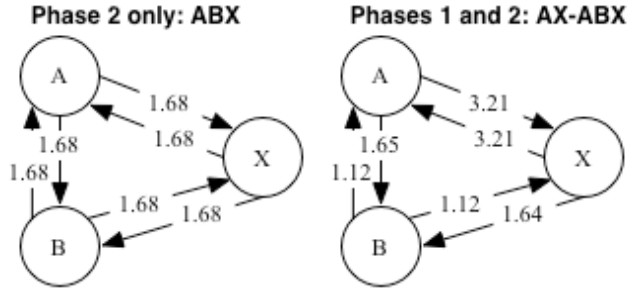


Figure 1: Weights demonstrating cue competition for causes after Phase 1 and 2 training (but not after Phase 2 only training). A and B are causes, X is the common effect.

Similarly, the recurrent model predicts cue competition for effects using the same rule. Again, the model predicts an asymmetry between the associative strengths of the links from A→Y and Y→A. The associative link from Y→A should exhibit competitive learning because when Y is activated during Phase 2 training, A is already highly activated from the simultaneous presence of X, which is a perfect predictor of A. Thus, there is very little weight change in the link from Y→A. However, the link from A→Y should not exhibit cue competition in learning because when A is activated during Phase 2 learning, Y is not initially predicted by anything. Therefore, the discrepancy between the target and the actual activations of Y are large, resulting in a bigger weight change in the link from A→Y. Thus, the associative strength from A→Y should be stronger than the associative strength of Y→A. Read's simulation results in Figure 2, with the same parameters, reflect this as well.

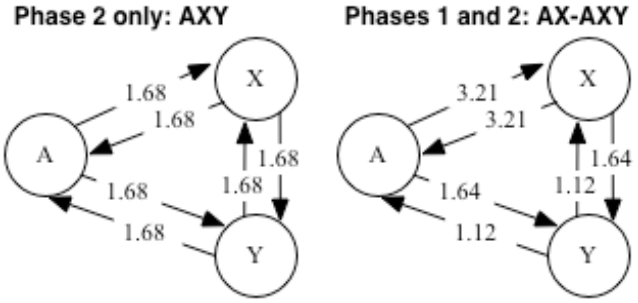


Figure 2: Weights demonstrating cue competition for effects after Phase 1 and 2 training (but not after Phase 2 only training). A is the cause and X and Y are the effects.

Purpose. Our purpose is threefold. First, we will provide further evidence that cue competition for effects occurs. Second, we will demonstrate that this effect can be obtained using social stimuli. Finally, we will test the predictions of the recurrent network model for cue competition and the asymmetry in the associative strengths of the bidirectional links between the cue and the target outcome. We will ask subjects to make judgments about the associative strengths of each of the two possible links between a cue and an outcome. Doing this in the same study has not previously been done. It is expected that cue competition for effects will be observed as well as the asymmetry in the associative strengths of the two possible links: the target outcome, Y, will exhibit cue competition with outcome, X, for the weight from Y→A, but not for the weight from A→Y.

We will get directional measures of strength by asking subjects to make conditional probability judgments between all pairs of nodes. Interpreting these judgments depends on the relationship between weights and conditional probabilities. As O'Reilly and Munakata (2000) show, in neural networks, the weight from an input i to an output o is a function of the conditional probability of the input given the output (p [i|o]). The output node can be thought of as corresponding to a hypothesis and the input node to data concerning the hypothesis. Thus, the weight from input to output captures the conditional probability of the data given the hypothesis. The critical implication is that judgments of the conditional probability p[Y|A] will be sensitive to the strength of the weight from Y to A whereas judgments of p[A|Y] should be sensitive to the weight from A to Y.

Study 1

Stimuli were four unrelated behaviors that had no known preexisting relationship with each other: breaking a glass (cause A); shaving one's head (effect X); lighting a tree on fire (effect Y); and meditating (filler effect Z). Information was presented simultaneously in list format (as Van Hamme et al. (1993) and Matute et al. (1996) did) with a dashed line break in between the antecedent and subsequent events to separate the common cause from the multiple effects.

Participants were instructed to learn the causal relationships between the antecedent and subsequent events. Finally, questions designed to assess the conditional probabilities and probe the associative strengths between all four behaviors were used (A→X; X→A; A→Y; Y→A; X→Y; Y→X; A→Z; Z→A).

Method

Participants. 93 undergraduates from the University of Southern California volunteered for extra credit. The study was a between subjects design with repeated measures on judgments, where the experimental group received both Phase 1 and Phase 2 of training, and the control group only received the Phase 2 training.

Materials and Procedure. Subjects were randomly assigned to the experimental or control group and seated in front of a computer, on which the entire experiment was done. The cover story asked subjects to imagine that they were anthropologists in the distant future traveling to a long lost human colony on a faraway planet to study their culture and social customs. They were instructed that their goal was to learn the various behavioral patterns of the colonists by observing individual instances of sets of behaviors. They were instructed to learn the causal relationships among the behaviors. Before seeing the behaviors, all subjects made initial judgments about the extent to which the four stimulus behaviors were related to each other to establish that they had no known relationship. They were asked to rate the extent to which the occurrence of one behavior affected the likelihood of another behavior on a scale from -10 to 10, where -10 indicates "strongly inhibits," 10 indicates "strongly promotes", and 0 indicates "no relationship." These questions and rating scale were also used for the testing after the training phase(s).

After the initial ratings, subjects in the experimental condition received Phase 1 training, where they saw 10 behavior sets exhibited by 10 individuals. Each set was presented individually on separate screens along with the name of the individual exhibiting the behaviors. Each set was displayed until the subject pressed the space bar. Eight of the 10 sets involved "breaking a glass" (cause A) followed by "shaving one's head" (effect X); 2 of the 10 sets involved "breaking a glass" (cause A) followed by "meditating" (filler effect Z). The order of the 10 sets was randomized for each participant. After Phase 1, subjects started Phase 2 training, where they saw 10 more behavior sets exhibited by 10 new individuals. Eight of the 10 sets involved "breaking a glass" (cause A) followed by "shaving one's head" (effect X) and "lighting a tree on fire" (effect Y). As in Phase 1, two of the 10 sets involved "breaking a glass (cause A) followed by "meditating" (filler effect Z).

Subjects in the control condition only received Phase 2 training. Immediately after training, they made judgments about the extent to which one behavior affects the likelihood of another behavior, for all four behaviors in both directions. Thus, subjects made judgments about the extent to which "breaking a glass" affects the likelihood of "shaving one's head" (A→X); "shaving one's head" affects the likelihood of "breaking a glass" (X→A); "breaking a glass" affects the likelihood of "lighting a tree on fire" (A→Y); "lighting a tree on fire" affects the likelihood of "breaking a glass" (Y→A); "shaving one's head" affects the likelihood of "lighting a tree on fire" (X→Y); "lighting a tree on fire" affects the likelihood of "shaving one's head" (Y→X); "breaking a glass" affects the likelihood of "meditating" (A→Z); and "meditating" affects the likelihood of "breaking a glass" (Z→A).

Results and Discussion

The mean of the eight initial ratings was -.62 for the experimental condition, and -.49 for the control, indicating that the four behaviors had no preexisting causal relationships. The mean of the eight final ratings was 5.77 for the experimental condition, and 6.19 for the control, indicating that subjects learned the causal contingencies.

A between-groups comparison for each of the eight final ratings found all of them to be non-significant with one exception. As predicted, the difference between experimental and control group judgments of the Y→A rating (p[Y/A]) was found to be highly significant, $t(91)=3.02$, $p=.003$ (experimental M= 5.29 vs. control M=7.21) (See Figure 3). This provides evidence for cue competition between effects in the direction predicted by the recurrent network.

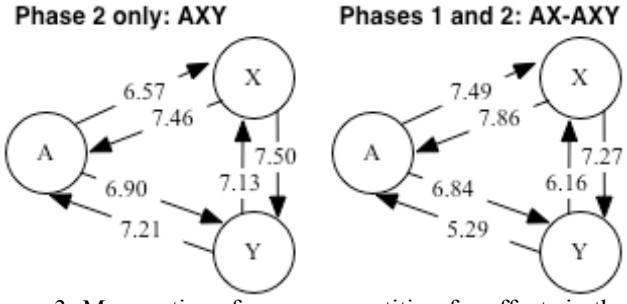


Figure 3: Mean ratings for cue competition for effects in the AXY (control) and AX-AXY (experimental) conditions. A is the cause, X and Y are the two competing effects.

Although the results of Study 1 were consistent with our predictions, we wondered whether the test questions did a good job of measuring conditional probabilities. After learning the behavioral contingencies, subjects were asked to make some "final estimates about the extent to which the occurrence of the first behavior affects the likelihood of the occurrence of the second behavior." However, for the backward reasoning test questions (e.g., X→A, Y→A, and Z→A) the actual test questions asked subjects to indicate the likelihood that an individual doing X (or Y or Z) had previously done A. In other words, the task instructions implicitly asked subjects to make forward casual judgments from the first listed behavior to the second, while the actual test questions asked subjects to make backward causal judgments from the first listed behavior to the second (at least for the backward reasoning questions). Thus, it is unclear whether we were successful in measuring the bidirectional associative strengths among the four behaviors.

Study 2

Study 2 involved several changes. First, the wording of the test questions was changed to more clearly measure

conditional probabilities. Test questions probing the associative strength of the link from A→X were more clearly phrased as the probability of A given X (p[A|X]), X→A was more clearly phrased as the probability of X given A (p[X|A]), and so forth (The phrasing of the new questions is presented later). Next, because the recurrent network model also predicts cue competition for causes as a result of asymmetrical associative strengths in the links between the redundant cue and outcome, the design from Study 1 was used to investigate cue competition for causes as well as for effects. For the Effects condition, the design and stimulus behaviors (Cause A, Effects X, Y and Z) were identical to those of Study 1. For the Causes condition, Cause A and Effect X remained the same with the addition of a new redundant Cause B, ("ringing a bell") in the Phase 2 training portion, and changing the previous filler effect Z to filler cause Z. Finally, Study 2 was conducted on the web.

Method

Participants. 168 adults ranging from ages 18 to 67 participated in this study on the Internet. Mean age was 39.28 (SD = 12.792). Participants were from previous on-line studies, unrelated to causal reasoning, who indicated that they were interested in future on-line studies. They were recruited by email and were residents of the US, with three exceptions. They were entered into a lottery for a $50 cash prize, with the odds of winning at 1/50. The study was a between subjects design with repeated measures on judgments. As before, the experimental groups for causes and effects received both Phase 1 and Phase 2 training, and the control groups only received Phase 2 training.

Materials and Procedure. Subjects clicked a link in their email directing them to the study. Upon clicking the button that initiated the experiment, subjects were randomly assigned to one of four conditions: experimental and control conditions for Causes and experimental and control conditions for Effects. Subjects were presented with the same cover story as in Study 1. Because Study 1 showed that there were no preexisting relationships between the behaviors, the initial judgments were dropped for Study 2.

The procedures for the Effects conditions were identical to Study 1. Procedures for the Causes conditions were identical to those for Effects, with the exception of changes in the behaviors. In Phase 1 for the experimental group, 8 of the 10 sets involved "breaking a glass" (cause A) followed by "shaving one's head" (effect X); 2 of the 10 sets involved "meditating" (filler cause Z) followed by "shaving one's head" (effect X). In Phase 2, 8 of the 10 sets involved "breaking a glass" (cause A) and "ringing a bell" (redundant cause B) followed by "shaving one's head" (effect X). As in Phase 1, 2 of the 10 sets involved "meditating" (filler cause z) followed by "shaving one's head" (effect X). Control subjects only received Phase 2 training.

The order of the eight test questions was randomized for each subject. Immediately following training, subjects were asked to make final judgments about the extent to which the occurrence of one behavior affects the likelihood of the occurrence of another behavior in terms of conditional probabilities for all four stimulus behaviors both directions. They made their judgments by clicking on a radio button on a rating scale from 0 to 10, where 0 indicates "No chance", 5 indicates "50-50", and 10 indicates "Certain". The test questions were phrased: for p[A|X], "Assuming that someone shaves their head, how likely is it that they had broken a glass," and for p[X|A], "Assuming that someone has broken a glass, how likely is it that they will shave their head?," and so forth.

Results and Discussion

As presented in Figure 4, the results for the Effects condition in study 2 replicate those of Study 1.

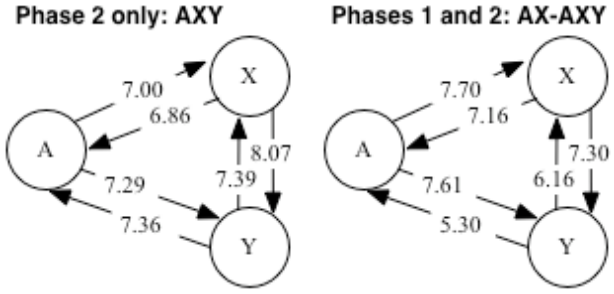


Figure 4: Mean ratings for cue competition for effects in the AXY and AX-AXY conditions in Study 2.

Between groups comparison of the final ratings for the experimental and control conditions showed a difference in the critical variable of Y→A rating, (p [Y|A]) $t(70)=3.75$, $p=.00$ (M=5.30 for experimental vs. M=7.36 for control). No other comparisons were significant. Cue competition for effects was asymmetric in the predicted direction.

For the Causes condition, there was a significant difference between the experimental and control groups for the critical variable of B→X rating (p[B|X]), $t(76) = 2.57$, $p = .01$. However, the difference for the non-critical variable of X→B (p[X|B]) was also significant, $t(76) = 2.88$, $p = .00$, as presented in Figure 5. Cue competition for causes is obtained, but contrary to the network model's predictions, there is no asymmetry in the associative strengths.

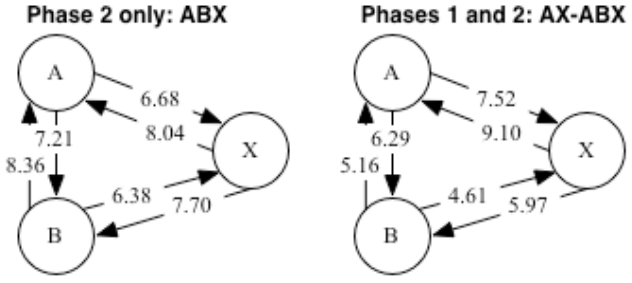


Figure 5: Mean final ratings for cue competition for causes in the ABX and AX-ABX conditions in Study 2. A and B are the competing causes, and X is the common effect.

General Discussion

These results replicate the well-established phenomenon of competition between causes (e.g., Van Hamme et al., 1993; Waldman & Holyoak, 1992) as well as the more controversial competition between effects (Shanks, 1991; Price &

Yates, 1993; Matute et al., 1996). Further, these studies show that this effect can be obtained with social behaviors and is not limited to biological or physical events.

The present study is the first to study cue competition for causes and effects by systematically exploring all possible directional links between causes and effects. Studies 1 and 2 demonstrated that cue competition between effects occurs on the weight from the redundant effect Y to the cause A, rather than on the weight from the cause A to the redundant effect Y. However, Study 2 seemed to indicate that cue competition between causes occurs on both the weight from the redundant cause B to effect X as well as on the weight from effect X to the redundant cause B.

The results clearly contradict causal model theory, which states that effects do not compete. As for the associative learning and the neural network models, the results support their prediction of competitive learning and the presence of cue competition between effects.

One advantage of the recurrent model is that it provides an account of cue competition for effects without the necessity of requiring diagnostic learning (effects precede causes). Further, the recurrent model predicts and the current results confirm that the extent of cue competition depends on the direction of the weight or relationship between cue and effect. Previous models would have been unable to make such a prediction.

The results for cue competition between effects are consistent with the idea that the weights are sensitive to the conditional probabilities between causes and effects. These studies show that cue competition between effects occurs on the weights from Y to A and not from A to Y after AX-AXY training. This makes sense in terms of conditional probability, in that, taking both Phase 1 and 2 into account, every time Y was presented, it was always preceded by A, and thus p(A|Y) is 100%. (note the weight from A to Y should encode this conditional probability). However, whenever A was presented, Y followed A only half the time (during Phase 2), and thus p(Y|A) is 50% (the weight from Y to A should encode this conditional probability). Thus, the asymmetry in cue competition between effects is consistent with the conditional probabilities.

However, with regard to cue competition for causes the same does not seem to apply. A conditional probability analysis should predict a similar asymmetry. Instead, the results indicate no asymmetry in the bi-directional associative links between the redundant cause and the common effect; cue competition occurs when reasoning both from B→X and from X→B. It is unclear why we do not get weight asymmetry for cue competition for causes. However, in further research, using different social stimuli, we did find the predicted asymmetry, suggesting that the current results may be specific to the current stimuli.

We note one other caveat. Research in this area has not separated the effects of learning from the judgment process. Future studies in cue competition should be designed to examine the various types of processes that participants may use to arrive at their judgments of contingency.

References

Cobos, P. L., Lopez, F. J., Cano, A., Almaraz, J. & Shanks, D. R. (2002). Mechanisms of predictive and diagnostic causal induction. . *Journal of Experimental Psychology: Animal Behavior Processes*, *28*, 331-346.

Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 225-244.

Matute, H., Arcediano, F. & Miller, R. R. (1996). Test question modulates cue competition between causes and between effects. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 182-196.

McClelland, J. L. & Rumelhart, D. E. (1988).*Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Cambridge, MA: MIT Press/Bradford Books.

Mels, E. R., Cheng, P. W., Holyoak, K. J., & Waldmann, M. R. (1993). Cue competition in human categorization: Contingency or the Rescorla-Wagner learning rule? Comment on Shanks (1991). *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 1398-1410.

Miller, R. R., Barnet, R. C. & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin, 117*, 363-386.

Miller, R. R. & Matute, H. (1998). Competition between outcomes. *Psychological Science, 9*,146-149.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.

Price, P. C. & Yates, J. F. (1995). Associative and rule-based accounts of cue interaction in contingency judgment. J*ournal of Experimental Psychology: Learning, Memory and Cognition, 21*, 1637-1655.

Read, S. J. (2003). *An integrative model of causal learning and causal reasoning using a feedback neural network.* Unpublished manuscript, U. of Southern California.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology, 37B*, 1-21.

Shanks, D. R. & Lopez, F. J. (1996). Causal order does not affect cue selection in human associative learning. *Memory & Cognition, 24*, 511-522.

Van Hamme, L. J. & Wasserman, E. A. (1993). Cue competition in causality judgments: The role of manner of information presentation. *Bulletin of the Psychonomic Society, 31*, 457-460.

Waldmann, M. R. (2000). Competition among causes but no effects in predictive and diagnostic learning. . *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 52-76.

Waldmann, M. R. & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Feature- vs. Relation-Defined Categories: Probab(alistic)ly Not the Same

Aniket Kittur (nkittur@ucla.edu)
John E. Hummel (jhummel@lifesci.ucla.edu)
Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology, 1285 Franz Hall
University of California, Los Angeles
Los Angeles, CA 90095

Abstract

Relational categories underlie many uniquely human cognitive processes including analogy, problem solving, and scientific discovery. Despite their ubiquity and importance, the field of category learning has focused almost exclusively on categories based on features. Classification of feature-based categories is typically modeled by calculating similarity to stored representations, an approach that successfully models the learning of both probabilistic and deterministic category structures. In contrast, we hypothesize that relational category learning is analogous to schema induction, and relies on finding common relational structures. This hypothesis predicts that relational category acquisition should function well for deterministic categories but suffer catastrophically when faced with probabilistic categories, which contain no constant relations. We report support for this prediction, along with evidence that the schemas induced in the deterministic condition drive categorization of novel and even category-ambiguous exemplars.

Relational and Feature-Based Categorization

Most mathematical models of human category learning start with the assumption that people represent categories as lists of features, and assign instances to categories by comparing the features of an instance to the features stored with the mental representation of the category (either a prototype or stored exemplars; e.g., Bruner, Goodnow, & Austin, 1956; Kruschke, 1992; Kruschke & Johansen, 1999; Nosofsky, 1992; Rosch & Mervis, 1975; Shiffrin & Styvers, 1997). Accordingly, most studies of human category learning in the laboratory investigate how people learn categories with exemplars consisting of well-defined (to the experimenter, at least) features.

In the real world, as some researchers have forcefully pointed out (e.g., Barsalou, 1993; Keil, 1989; Murphy & Medin, 1985; Rips, 1989; Ross & Spalding, 1994) categories are less often defined in terms of lists of features than in terms of *relations* between things: either relations between the features or parts of an exemplar (e.g., the legs need to be in a particular kind of relation to the seat in order for an object to serve as a chair), or relations between the exemplar and the user's goals (e.g., any object that affords sitting can, in some circumstances, be considered a chair), or relations between the exemplar and other objects in the world (e.g., what makes an object a "conduit" is a relation between that object and whatever thing flows through it, whether it be water, light, electricity, information, or karma). In spite of their importance in human cognition, comparatively little is known about how people learn relational categories.

Relational category learning is important because relational concepts (i.e., mental representations of relational categories) play an essential role in virtually all aspects of human thinking, including our ability to make and use analogies, problem solving, scientific discovery, and even aspects of perception (see, e.g., Gentner, 1983; Gentner et al., 1997; Green, 2004; Hesse, 1966; Holyoak & Thagard, 1995; Hummel, 2000). The utility of relational representations is that they permit generalization from a small (often as few as one or two) number of examples to a large (potentially infinite) number of new cases (as in the case of inferences generated through the use of analogies, schemas and rules; Gick & Holyoak, 1983; Pirolli & Anderson, 1985; Ross, 1987).

Relational concepts cannot be adequately represented as lists of features (as assumed by most current models of category learning), but instead must be mentally represented as relational structures such as schemas or theories (Gentner, 1983; Holland, Holyoak, Nisbett, & Thagard, 1986; Hummel & Holyoak, 2003; Keil, 1989; Murphy & Medin, 1985). This observation suggests that the operations governing relational schema induction may also underlie the acquisition of relational categories (see, e.g., Kuehne et al., 2000).

At least one theory of schema induction, Hummel and Holyoak's, 2003, LISA model, predicts that a schema induced from two or more examples retains (roughly) the structured intersection of what the examples have in common. For example, consider two analogous stories about love triangles. In the first, Abe loves Betty, but Betty loves Chad, so Abe is jealous of Chad; in the second Alice loves Bill, but Bill loves Cathy, so Alice is jealous of Cathy. Drawing an analogy between these stories maps Abe to Alice, Betty to Bill, and Chad to Cathy (along with the roles of the *loves* and *jealous-of* relations). The schema LISA induces from this analogy retains what the examples have in common, and de-emphasizes the ways in which they differ. For example, since the analogy maps males to females and vice versa, the resulting schema effectively discards the actors' genders, stating (roughly) "person1 loves person2 but person2 loves person3, so person1 is jealous of person3," where persons1…3 are generic people, rather than being specifically males or females (see Hummel & Holyoak, 2003).

Importantly, this intersection discovery process also takes place at the level of whole propositions. For example, if the second story contained a proposition stating that, as a

result of her jealousy, Alice was mean to Cathy, but the first story had no corresponding proposition, then LISA would simply drop this proposition in its entirety from the resulting schema.

If we assume that relational category learning is a process of relational schema induction, then this property of dropping unmapped propositions (i.e., unmapped relations) from the induced schema (i.e., category representation) leads to a counterintuitive prediction: If a relational category has a probabilistic structure, such that every member of the category shares some relations with every other member of the category, but there is no relation that *all* members share, then category learning should fail catastrophically. The reason is that the process of schema induction will drop any relation that is absent from any exemplar from the emerging schema. If every relation is absent from some exemplar (i.e., no relation is present in every exemplar), then schema induction will eventually drop every relation from the schema. By the end, the induced schema will be the empty set.

To clarify, consider a simple relational category with four exemplars, each with three relations chosen from the set r1, r2, r3 and r4 (for our current purposes it does not matter what r1…r4 are, only that they are relations of some sort). Let exemplar 1 (e1) contain the relations r1, r2 and r3. That is, e1 = [r1, r2, r3]. Similarly, let e2 = [r2, r3, r4]; e3 = [r1, r3, r4]; and e4 = [r1, r2, r4]. Note that mapping, for example, e1 to e2 results in a schema (s1,2) that contains relations r2 and r3 (which e1 and e2 share), but lacks r1 (which e1 possesses but e2 does not) and r4 (which e2 possesses but e1 does not): s1,2 = [r2, r3]. Mapping s1,2 onto, say, e3, produces a schema containing only r3, and mapping that schema onto e4 produces a schema containing no relations. The resulting schema is clearly not a useful basis for classifying exemplars as members of the category.

The point is that relational category learning is predicted to be extremely difficult when the categories have a strictly probabilistic structure (i.e., with no relation shared by all exemplars). By contrast, if there is even a single relation that is shared by all exemplars, then category learning should improve dramatically relative to the purely probabilistic case. Categorization performance should also improve dramatically, even with purely probabilistic categories, if the relational structure is replaced with a feature-based structure. Learning of feature-based categories is well known to be robust to probabilistic category structures, a fact that underlies prototype effects (e.g., Posner & Keele, 1968).

In summary, we predict a sharp dissociation between relational and feature-based category learning with respect to their robustness to probabilistic category structures: Both relational and feature-based categories should be learnable when they have a deterministic structure, even if only a single relation or feature reliably predicts category membership; similarly, feature-based categories should be learnable whether they have a deterministic structure or a probabilistic one. By contrast, relational categories should be extremely hard to learn from examples when those examples are presented in a probabilistic structure.

We tested this hypothesized dissociation between relational and feature-based category learning using a 2x2 design, in which relational vs. feature-based categories were crossed with probabilistic vs. deterministic category structures. In order to control all extraneous sources of potential effects, the same basic stimulus set was used in all four conditions; only the assignment of stimuli to categories varied.

Method

Subjects. 33 UCLA undergraduate students participated for course credit.

Instructions. Participants were read a cover story describing a computer manufacturer trying to determine the function of accidentally unlabelled computer chips. Subjects then engaged in a training phase followed by a transfer phase. During both phases, subjects were instructed to indicate the category to which the onscreen stimulus belonged by pressing one of two keys. The categories were labeled "math" chips and "graphics" chips.

Materials. On each trial, the subject saw an exemplar consisting of an octagon and a square, arranged on a fixed background designed to resemble a computer chip (see Figure 1). Each exemplar had both relational properties (e.g., octagon *bigger than* square) and featural properties (e.g., octagon of size 3).

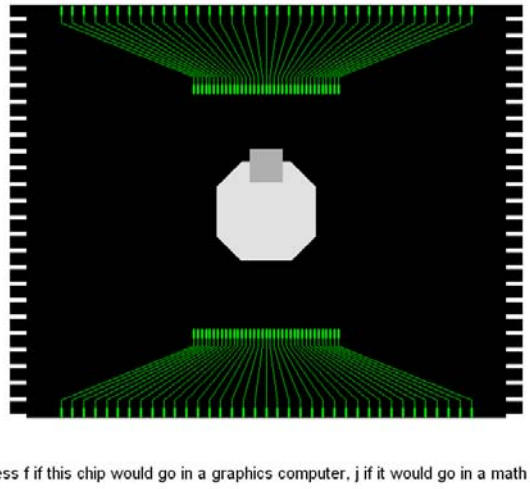


Figure 1: Example stimulus.

The properties of each exemplar were determined by an identical family resemblance category structure (see Table 1). The prototypes of the two categories were defined as (1,1,1,1) and (0,0,0,0), and distortions were made by chang-

ing the value of one or more dimensions to its opposite[1]. Each column in Table 1 represents an exemplar, and the particular value on each dimension (1 or 0) defines the value of a relation (in the relational condition) or a feature (in the featural condition) for each exemplar. The values for both the relational and featural properties are listed in Table 2. For example, the relational prototype with structure (1,1,1,1) would have an octagon *bigger, darker, above,* and *in front of* a square, while the prototype with structure (0,0,0,0) would be the exact opposite. The properties were set up so that using features could not result in learning to criterion in the relational condition, and using relations in the feature condition would also lead to sub-criterion responding.[2] Stimulus generation and display as well as response collection were done with a program written in Matlab.

Design. The experiment used a 2 (category structure: probabilistic vs. deterministic) X 2 (relevant property: features vs. relations) between-subjects design. The only difference between the conditions in terms of the stimuli used was that, in the deterministic condition, a single distorted exemplar from each category was not presented during training, so that one dimension was constant for all exemplars of a category. The choice of which dimension was held constant was counterbalanced across subjects.

Table 1: Family resemblance category structure. Each column represents an exemplar, and each row a dimension.

Category A					Ambiguous						Category B				
1	1	1	1	0	1	1	1	0	0	0	0	0	0	1	0
1	1	1	0	1	1	0	0	1	0	1	0	0	1	0	0
1	1	0	1	1	0	1	0	1	1	0	0	1	0	0	0
1	0	1	1	1	0	0	1	0	1	1	1	0	0	0	0

Procedure. During the training phase subjects classified only distorted exemplars of each category (depicted in the light gray columns of Table 2). All distortions for each category were shown in random order exactly once per block. Responses were followed by accuracy feedback, during which the exemplar remained on the screen. Subjects pressed the space bar to proceed to the next trial. The training phase continued until the subject responded correctly on at least seven out of eight trials for two consecutive blocks[3], or until they had finished 75 blocks (600 trials) without reaching this criterion.

Following the training phase, subjects were informed that they would be tested on chips for which feedback could not be given. During this transfer phase subjects classified all 16 possible exemplars, including the prototypes and ambiguous exemplars. Subjects completed five blocks, with each block showing all 16 exemplars in random order exactly once.

After the transfer phase, each subjects completed a questionnaire in which they were asked to write down the criteria they used to categorize the exemplars.

Table 2: Category definitions.

Relational categories

Exemplar	Relation	Exemplar	Relation
1	Bigger	0	Smaller
1	Darker	0	Lighter
1	Above	0	Below
1	In Front	0	Behind

Feature-based categories

Exemplar	Feature	Exemplar	Feature
1	O *size* 3	0	O *size* 2
1	O *shade* 4	0	O *shade* 3
1	S *size* 1	0	S *size* 2
1	S *shade* 1	0	S *shade* 2

Note: Prototype exemplars are shown with their defining properties on each dimension. In the relational condition, each dimension defines how the octagon (O) in the stimulus relates to the square (S). For the featural condition each dimension defines specific feature values.

Results

Training. Only 5 of the 7 subjects (71%) in the relational probabilistic (RP) condition learned to criterion within 600 trials. 25/25 subjects (100%) in the other conditions learned to criterion within the 600 trial limit. In the analyses that follow, the 2 subjects in the RP condition who never learned to criterion are treated as though they reached criterion on trial 601. Given that our hypothesis predicts that learning in the RP condition will be harder (and therefore take longer) than learning in the other conditions, this assumption is extremely conservative.

The mean number of trials to criterion is shown in Figure 2 for each condition. Subjects in the RP condition took more trials to reach criterion than those in the FD (featural deterministic), FP (featural probabilistic), and RD (rela-

[1] Note that the exemplars marked "Ambiguous" are equal distance between the two prototypes, having exactly two values different from each.

[2] In the relational condition, stimuli from different categories could have the same features (and stimuli from the same category could have different features) as long as the specified relations held; features were thus non-diagnostic. For the featural condition, the relations *in front* and *above* had no relevance to the category structure, and were pseudo-randomized. The relations *bigger* and *darker* were made irrelevant by choosing values such that the octagon was never smaller or lighter than the square (though it could be the same size, since only three sizes were used). See the Discussion section for further analysis of feature and relation values.

[3] This criterion level (87.5%) was selected because strategies involving tracking only one or two relations in the probabilistic condition would not meet the criterion level (both would result in 75% correct responding).

tional deterministic) conditions. A planned contrast comparing the RP condition to the other three revealed that this difference was statistically reliable ($p < 0.01$). There was also a significant main effect of category type (relational vs. featural, $F(1,33)=4.64$, $p < 0.05$). The main effect of category structure (deterministic vs. probabilistic) and the interaction were both marginally reliable ($0.05 < p < 0.15$).

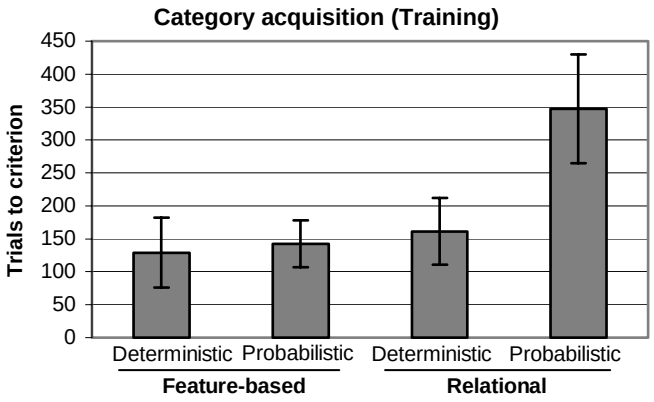


Figure 2: Average number of trials required by subjects in each condition to reach criterion during training.

Transfer. The key prediction for the transfer phase was that subjects in the deterministic condition would categorize exemplars based on whatever dimension was held constant during training. This prediction applied especially to the relational condition, which could not rely on holistic processing. To test this hypothesis we analyzed classification of the ambiguous exemplars, which were equidistant between the two prototypes. Subjects who used all category dimensions equally should be unsystematic in their classification of these ambiguous exemplars. By contrast, subjects who attend to a single dimension should classify ambiguous exemplars according to that dimension only (as detailed in Table 3). If a classification response for a dimension that was held constant during training matched the response pattern in Table 3, then +1 was scored for that response; classifications that did not match Table 3 response patterns were scored as -1. Under this scoring system, consistently responding to ambiguous exemplars in the direction predicted by the constant training dimension results in a positive score; consistently responding in the direction opposite the constant dimension results in a negative score; and unsystematic responding results in a score near zero.

Classification of ambiguous exemplars in accordance with the dimension that was constant during training was significantly above chance ($p < 0.01$). Breakdown into featural and relational conditions showed a non-significant trend for the relational condition to evoke classifications based on the constant training dimension more often than the featural condition (Figure 3).

Discussion

The results showed that acquisition of relational probabilistic categories takes significantly longer than acquisition of deterministic relational categories, or featural categories of any kind (probabilistic or deterministic). Importantly, the ease of acquisition in the deterministic relational condition shows that this effect is not due strictly to the relational nature of the task. Instead, the catastrophic failure represents an interaction between the relational nature of the stimuli and the probabilistic structure of the categories. This interaction is consistent with the hypothesis that relational category learning is a process akin to relational schema induction by intersection discovery: When the intersection is the empty set (as it is in the probabilistic condition but not the deterministic condition), relational category learning suffers markedly. By contrast, feature-based category learning is much more robust to the probabilistic category structure, presumably because feature-based category learning is not a process of relational schema induction; instead, as predicted by models of feature-based category learning, it may be that learning feature-based categories can be accomplished simply by cataloging and matching features.

Table 3: Classification of exemplars based on single dimensions

Exemplar	Dim 1	Dim 2	Dim 3	Dim 4
1 1 0 0	A	A	B	B
1 0 1 0	A	B	A	B
1 0 0 1	A	B	B	A
0 0 1 1	B	B	A	A
0 1 1 0	B	A	A	B
0 1 0 1	B	A	B	A

Note: Table entries indicate how each exemplar would be categorized by a subject who attended only to a single dimension (columns in the table). For example, a subject who attended only to the first dimension (Dim 1) would classify the first, second and third exemplars as As since their values on that dimension are all one, and B for the fourth, fifth and sixth exemplars (the values of which are zero on that dimension).

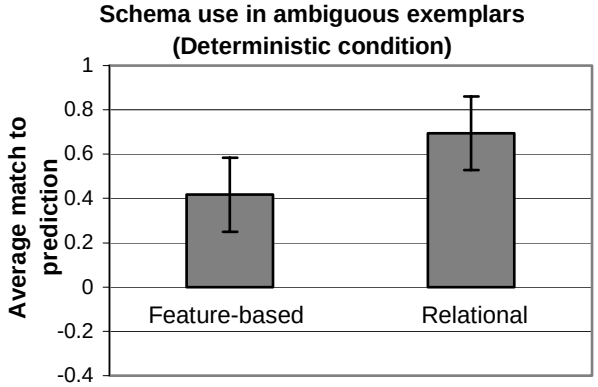


Figure 3: Average match to predicted classification pattern. Positive values indicate schema-based classification; zero corresponds to unsystematic responding.

The significant match of subjects' responses with single-dimension classification predictions in the deterministic condition also shows that subjects do preferentially use dimensions that are constant during training to classify novel and even category-ambiguous exemplars.

Is it possible to explain these results in other ways? One possibility is that rather than attending to the relations, subjects in the relational conditions may instead be tracking the feature values of certain dimensions. On this hypothesis, there is no schema induction going on in any condition; instead, responding is based on the values of particular features. This account obviates the need for a separate process to explain relational categorization.

However, analysis reveals that subjects tracking a feature of a single dimension would only classify 5/6 correct in the deterministic condition, and 2/3 correct in the probabilistic condition. Both of these values are below the 7/8 criterion, suggesting that subjects who reach criterion were not doing so by tracking a feature of a single dimension.

The possibility remains that subjects were tracking the values of multiple features or dimensions, although these seem unlikely strategies for a number of reasons. First, even when tracking the values of a single feature the subject must hold in mind three or four values and their associations with each category (for example, each size of the octagon and its corresponding category). Each additional feature or dimension would double the number of values necessary to track. This strategy does not seem plausible given the well-known limits on the capacity of working memory. Also, subjects' responses to the debriefing questionnaires in the relational conditions did not suggest such strategies were being used; instead, they generally reported the use of one or two relations as diagnostic, often along with some exception exemplars. Thus it seems more likely that subjects were indeed attending to the relations between the components of each stimulus rather than tracking feature values of those components.

Another hypothesis to explain the difference between the featural and relational conditions is that subjects were memorizing all the possible exemplars, and a difference in the number of distinct exemplars made the relational condition harder. This view must also hold that the deterministic conditions do not rely on such memorization, in order to explain the results. This view has some merit, though two factors reduce its likelihood.

First, the total number of distinct exemplars in the featural condition is not very different than the relational condition: 128 vs. 144. While this is a difference between the categories, it is difficult to ascribe the extra difficulty of the relational probabilistic condition to its having an extra 16 exemplars.

Second, although debriefing questionnaires did indicate subjects were memorizing some of the exemplars in the relational probabilistic condition, these were of very limited number (usually ~2 exemplars) and were memorized as exceptions to a more general classification rule. Thus while it remains a theoretically possible explanation, the "number of exemplars" view is not very compelling.

Preliminary analysis of the debriefing forms for subjects who learned to criterion in the relational probabilistic task suggest that what is learned is often a classification rule (such as might result from a schema induction process) along with a few memorized exceptions. Subjects often mentioned one or two relations in their classification rules; only one subject reported attending to all four dimensions; unsurprisingly, this subject was the only one who deduced the formal category structure (that is, that three out of four of the dimensions are necessary for category membership).

Subjects in the featural probabilistic condition also failed to show feature-tracking strategies in their debriefing questionnaires. Instead, their responses often showed a reliance on emergent properties of the stimuli such as high vs. low contrast. Questionnaires from the deterministic conditions tended to show a focus on the dimension that was constant during training, and mentioned particular features in the featural condition and relations in the relational condition. Thus subjects' explicit responses often fit well with the predictions about processing.

Why should relational categories rely on schema induction processes? One possibility is that feature-based categories tend to give rise to emergent properties, since their features are fixed at some value or limited range of values. However, it is much more difficult for emergent properties to arise in relational categories, because they can take on many different and overlapping values. The lack of emergent properties may explain the dependence of relational categorization on deterministic dimensions. This view is consistent with subjects' self-reported strategies.

Another interpretation of the present results is that people are either unwilling or unable to perceive, predicate and categorize patterns across four relations. This deficit may be due to working memory constraints, strategy choice, or low prior experience with similar situations. Studies of working memory suggest that we can hold about four chunks or role bindings in working memory (e.g., Halford, Wilson, & Phillips, 1998); holding four two-place relations exceeds this limit. It may be that people can learn some probabilistic relational categories with experience by recoding relations as features; others may be learned by dividing the probabilistic category into deterministic subcategories, or by perceiving a unifying causal relation for the entire category.

In conclusion, the results of the present experiment suggest that relational category learning relies heavily on finding common relations across exemplars. In contrast, feature-based category learning appears to function robustly whether common elements are present or not. These findings are consistent with the view that relational category learning is a kind of relational schema induction that depends on intersection discovery. Performance on the transfer trials also support this conclusion in that dimensions that were constant during training dominated classification of novel exemplars, even those that were category-ambiguous. Such findings suggest that relational category

learning may be fundamentally different from feature-based category learning, though more work is needed to distinguish these modes of category learning.

References

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins & S. E. Gathercole & M. A. Conway & P. E. Morris (Eds.), *Theories of memory*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. Oxford, England: John Wiley and Sons.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D., Brem, S., Ferguson, R. W., Markman, A. B., Levidow, B. B., Wolff, P., & Forbus, K. D. (1997). Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences. Special Issue: Conceptual change, 6*, 3-40.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Green, C. B., & Hummel, J. E. (2004). Relational perception and cognition: Implications for cognitive architecture and the perceptual-cognitive interface. In B. H. Ross (Ed.), *The psychology of learning and motivation*. San Diego: Academic Press.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral & Brain Sciences, 21*, 803-864.

Hesse, M. B. (1966). *Models and analogies in science.* Notre Dame, IN: University of Notre Dame Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA, US: The MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: The MIT Press.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in human shape perception. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220-264.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA, US: The MIT Press.

Kuehne, S., Forbus, K., Gentner, D. and Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*, 1083-1119.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes, Vol. 1: From learning theory to connectionist theory; Vol. 2: From learning processes to cognitive processes*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Pirolli, P. L., & Anderson, J. R. (1985). The role of practice in fact retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*, 136-153.

Posner, M. I., & Keele, S. W. (1968). On the Genesis of Abstract Ideas. *Journal of Experimental Psychology, 7*, pp. 353-363.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. O. Vosniadou, Andrew (Ed.), *Similarity and analogical reasoning*. New York, NY, US: Cambridge University Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 629-639.

Ross, B. H., & Spalding, T. L. (1994). Concepts and categories. In R. J. Sternberg (Ed.), *Thinking and problem solving. Handbook of perception and cognition (2nd ed.)*. San Diego, CA: Academic Press, Inc.

Shiffrin, R. M., & Styvers, M. (1997). A model for recognition memory: REM--retrieving effectively from memory. *Psychonomic Bulletin & Review*, 145-166.

Are natural kinds psychologically distinct from nominal kinds? Evidence from Learning and Development

Heidi Kloos (kloos.6@osu.edu)
The Ohio State University, Center for Cognitive Science
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir Sloutsky (sloutsky.1@osu.edu)
The Ohio State University, Center for Cognitive Science
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

Known theories of categorization operate under the assumption that most concepts are fundamentally similar. The current research argues that this assumption is unwarranted: Different types of concepts may differ in how they are represented and learned. We specifically focus on natural-kind and nominal-kind concepts, concepts that differ in their statistical structure. Natural kinds consist of highly redundant and intercorrelated features, whereas nominal kinds consist of isolated rules that do not correlate with other features. If these types of concepts are fundamentally different, they should exhibit important dissociations in how they are learned. Two learning regimes were contrasted: one in which participants were shown instances of the concept without being given a definition of the concept (implicit learning regime), and one in which participant were given a definition of the concept without being shown individual instances (explicit learning regime). Preschoolers and adults participated. The results show a strong dissociation between the two kinds of concepts in terms of acquisition, indicating that existing theories of categorization are incomplete.

Introduction

The ability to form categories by overlooking differences for the sake of generality is a critically important component of cognition. While the importance of concepts and categories is widely accepted, a number of puzzling questions remain unanswered. How do categories arise? Which processes underlie categorization? And how are categories represented in the cognitive system?

Several influential approaches have emerged in an attempt to answer these questions. According to the "classical view," categories are represented by sets of features that are individually necessary and jointly sufficient to determine category membership (Bruner, Goodnow, & Austin, 1956; Vygotsky, 1986/1934; for a review see Smith & Medin, 1981). For example, the concept *prime number* includes two features: an integer, and no remainder when divided by one or by itself. Each feature is necessary and they are jointly sufficient to determine whether or not a number is a prime.

By the 1980s, the classical view came under severe attack for its inability to predict and account for several key phenomena in the study of concepts, such as, for example, the gradedness of category membership. (Mervis & Rosch, 1981; see also Murphy, 2002, Smith & Medin, 1981, for extensive reviews).

With the demise of the classical view, two theoretical approaches to conceptual development have emerged: the naïve-theory approach and the similarity-based approach. The naïve-theory approach argues that even if there are no truly defining features, features differ in their conceptual centrality, this centrality being often determined by a feature's causal status (Medin, 1989; Gelman & Coley, 1991; Keil, Smith, Simons, & Levin, 1998). For example, apples and basketballs are round, but the feature "roundness" is more central for basketballs than it is for apples.

On the contrary, the similarity-based approach suggests that categorization decisions are made on the basis of similarity between a to-be-categorized entity and existing categories (see Murphy, 2002; Sloutsky, 2003, for reviews). Categories could be represented as best examples or prototypes (Posner & Keele, 1968, Rosch & Mervis, 1975) or as sets of encountered exemplars (e.g., Nosofsky, 1986, 1992). In the former case, an entity is categorized as A if it is similar to A's prototype, whereas in the latter case an entity is categorized as A if it is similar to individual exemplars of A encountered previously.

Despite the differences among these theoretical approaches, there is an important commonality – they implicitly assume that all (or at least most the concepts) concepts are fundamentally the same, and therefore, that concepts have to be learned and represented in the same or a similar way.

However, it is possible that there are different classes of concepts that give rise to different types of representation. The particular distinction considered here can be mapped onto the normative distinction between *natural kinds* and *nominal kinds* (Kripke, 1972; see also Keil, 1989, for a review). Natural kinds refer to classes of entities that exist in nature (e.g., *bird*), and nominal kinds refer to more arbitrary groupings based on a small set of necessary and sufficient properties (e.g., *triangle, acceleration*).

Natural kinds may differ in several ways from nominal kinds. However, the difference highlighted in the current experiments pertains to the difference in their statistical

structure. Natural kinds have a rich correlational structure, meaning that the relevant features correlate among each other. For example, creatures that lay eggs also have feathers and fly. Nominal kinds, on the other hand, lack such correlations among relevant features. They are based instead on an isolated rule. For example, accelerated motion does not have any common features with other motions except the change in velocity or the change in vector of the moving body.

It seems that the classical view of categorization considered nominal kinds as most representative concepts, whereas the similarity-based positions considered natural kinds as the most representative ones. The current study asks whether the normative distinction between natural and nominal kinds is accompanied by a psychological distinction between these two types of concepts. If true, such psychological distinction should manifest itself in how natural kinds and nominal kinds are represented and learned. The goal of this research is to examine dissociations in learning of natural and nominal kinds.

Statistical Structure of Concepts

To reiterate, natural-kind concepts often have multiple correlations among features of category members. Nominal kinds, on the other hand, are typically based on a small set of features uncorrelated with other features. It could be said then that natural kinds are statistically dense, embedded in multiple redundancies, whereas nominal kinds are statistically sparse, that is based on a single rule embedded in irrelevant variance.

Because natural kinds are statistically dense, it is possible that natural kinds are acquired spontaneously and do not require explicit training. Even infants are sensitive to multiple correlations and can spontaneously acquire categories based on multiple correlations. (e.g., Younger, 1993). Therefore, is likely that the mere exposure to instances of a natural kind could suffice for the acquisition of the concept. For example, infants may learn to group dogs together after seeing a variety of dogs (Quinn, Eimas, & Rosenkrantz, 1993). The basis for this learning is extraction of statistical information from a set of exemplars (Mareschal & Quinn, 2001). In fact, explicit training of a natural-kind category may hurt the acquisition of the natural kind. Billman and Knuston (1996) showed that in an unsupervised-learning setting, adults could learn the concept that was based on redundant relations, while failing to learn the concept when it was based on an isolated or orthogonal relation.

Nominal kinds are statistically sparse, meaning that they lack redundancy, and that only a limited set of features or feature relations is relevant. Because only a small portion of total information is relevant for the membership in a concept, it might be difficulty for the learner to spontaneously determine what is relevant, without having explicit instruction. This might be especially true for relational concepts, those that are based on a relation among features, not the features themselves (e.g., the concept of *ratio*). There are few reasons to believe that mere exposure to a limited set of instances would result in an acquisition of a relational concept. On the contrary, even feedback-based learning of relational concepts proved to be a challenge (e.g., Bruner, et al., 1956).

Based on these considerations, we hypothesize that there might be an acquisitional dissociation between natural and nominal kinds, with the former requiring unsupervised exposure (i.e., implicit learning regime), and the latter requiring an explicit instruction about the relevant rule (i.e., an explicit learning regime).

Overview of Experiments

In the three reported experiments, we systematically manipulated two factors: the type of the concept to be learned ("natural kind" vs. "nominal kind") and the learning regime (implicit learning regime vs. explicit learning regime). Preschool children (Experiment 1) and adults (Experiment 2 and 3) participated in the four resulting conditions: implicit or explicit learning of a concept that mimics natural kinds, and implicit or explicit learning of a concept that mimics a nominal kind.

For both kinds of concepts, the same animal-like artificial stimuli were created, such that none of the single features were predictive of the category membership. Only the relations between features mattered. Similar to natural kinds in the real world, the natural kind of the current experiment was based on multiple correlations among features (e.g., creatures that had a dark body also had a long tail and lots of wings). Conversely, in the nominal kind of the current experiment only one, arbitrary selected, relation was predictive of category membership.

In the implicit learning regime, the learners were presented with instances of the target category without being told the defining rule of the category. Conversely, in the explicit learning regime, the learners were given the defining rule of the category without being shown specific instances.

We predicted an interaction between learning regime and kind of concept. The concept that is based on redundant relations (i.e., "natural kind") should be best learned in the implicit learning regime, and the concept that is based on an isolated relation (i.e., "nominal kind") should be best learned in the explicit learning regime.

Experiment 1

The goal of the first experiment was to examine the acquisition of natural-type and nominal-type concepts under different learning regimes by young children.

Method

Participants Participants were 61 5-year-olds (32 girls and 29 boys), recruited from suburban middleclass preschools. The mean age in each condition (natural/implicit, natural/ explicit, nominal/implicit, nominal/ implicit) in months was 58.1 (*SD* = 4.6), 61.9 (*SD* = 2.7), 60.4 (*SD* = 5.2), and 60.3 (*SD* = 4.7), respectively. Additional 28 children were tested

(n = 8, 5, 7, and 8 in the respective conditions) and omitted from the sample because their performance in the catch trials did not meet the criterion (see Procedure).

Stimuli The stimuli were colorful drawings of unfamiliar animals presented on a computer screen. Each instance had the following six parts: a body, antennas, horizontal and vertical wings, a tail, and buttons on the body (Figure 1). These six parts could vary in at least one characteristic. They could vary in size (e.g., long or small tail), in shade (e.g., dark or light body), or in number (e.g., few or lots of buttons).

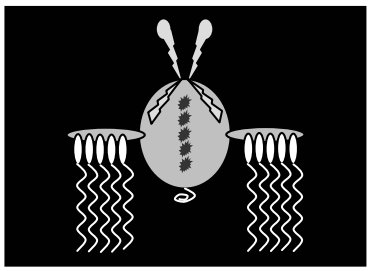


Figure 1: Example of the stimuli.

Two types of categories were created, one that included multiple correlations of features (i.e., they approximated a natural kind), and the other that were based on a single arbitrary selected relation (i.e., they approximated a nominal kind). Each type of category consisted of a target category and a contrasting category. Table 1 shows examples of items to illustrate how stimuli differed between natural and nominal kinds, and between target category and contrasting category.

For the natural kind, the sizes, shades, and number of parts correlated systematically. In the target category, a light body had light antennas, short horizontal wings, a short tail, few vertical wings, and few buttons. And a dark body had dark antennas, long horizontal wings, a long tail, many vertical wings, and many buttons. In the contrasting category the correlations were reversed. For example, a light body went with dark antennas, short horizontal wings, a long tail, few vertical wings, and many buttons. No single feature was predictive of the category.

For the nominal kind, only the number of parts mattered, while the correlations among sizes and shades were varied randomly. In the target category, there were *more* buttons than tails and vertical wings together, and in the contrasting category, there were *fewer* buttons than tails and vertical wings together. The numbers of buttons, tails, and vertical wings were chosen in such a way that neither the number of a single part nor the correlation between two of the parts were predictive. This ensured that no other information (e.g., difference in quantity) was redundant with the rule.

An additional set of stimuli was created that functioned as catch items. These items were from the contrasting category but with very salient changes. They had a diamond shaped body, no buttons, and no horizontal wings.

Table 1: Structure of Exemplar Items Used in Experiments 1 and 2

	Target Cat.		Contrasting Cat.	
	Item 1	Item 2	Item 1	Item 2
Natural Kind				
Parts				
Body	light	dark	light	dark
Antennas	light	dark	dark	light
Horiz. wings	short	long	short	long
Tails	short	long	long	short
Vert. wings	few	many	few	many
Buttons	few	many	many	few
Nominal Kind				
Parts				
Tails	1	3	3	5
Vert. wings	4	2	4	6
Buttons	7	9	5	9

Note. For the nominal kind, the numbers refer to the actual number of a particular part for the nominal kind (e.g., 1 = one tail).

Procedure The cover story presented to children involved the creature Fritz who lives on planet Elbee and who would like to get a pet. Pets on planet Elbee are called Ziblets and come from a magical store that carries both pets and dangerous wild animals. Children's task was to determine whether or not an animal from this magical store is a Ziblet.

The procedure had two phases: a training phase and a testing phase. In the training phase, children were given information about Ziblets (target category in Table 1). In the implicit learning regime, they were shown 24 pictures of Ziblets, presented one by one. They were told: "I will show you the Ziblets that other families on planet Elbee have as pets. Can you look at them and try to remember them?" In the explicit learning regime, children were presented with the defining rule. They were either told (for the natural kind) "A Ziblet with a dark body has dark antennas, long horizontal wings, a long tail, one or two short vertical wings and two or three light buttons; and a Ziblet with a light body has light antennas, short horizontal wings, a short tail, four or five long vertical wings and five or six dark buttons", or they were told (for the nominal kind) "For a Ziblet, the number of buttons is smaller than the number of tails and vertical wings together". Each separate part mentioned in the rule (e.g., a long tail) was depicted on the computer screen.

The testing phase was identical in both learning regimes. Sixteen testing trials were presented in random order, half of them being instances of the target category (Ziblets) and half of them being instances of the contrasting category (Non-Ziblets). Children's task was to determine whether an

instance is a Ziblet or not. Six catch trials followed intermixed with instances of the target category. To be included in the study, children had to reject four of the catcher trials.

Results and Discussion

Accuracy scores were calculated for each participant by subtracting the number of correctly accepted Ziblets from the number of incorrectly accepted Non-Ziblets and transforming the difference into a proportion. An accuracy score of zero (i.e., no difference between proportion of hits and proportion of false alarms) would be expected by chance.

Figure 2 shows the mean accuracy scores for each condition. A 2 (concept: natural, nominal) by 2 (learning regime: implicit, explicit) between-subjects ANOVA revealed a significant interaction ($F(1,53) = 14.46$, $p < .001$), with the mean accuracy scores being above chance in the conditions natural/implicit ($t(15) = 4.07$, $p < .01$) and nominal/explicit ($t(15) = 3.2$, $p < .01$) but not in the conditions natural/ explicit and nominal/implicit.

An analysis of individual pattern of responses corroborated this trend. Eleven children in the natural/implicit condition (69%) and 12 children in the nominal/explicit condition (75%) had an accuracy score above 0.20. Conversely, only 5 children in the natural/explicit condition (31%) and only 3 children in the nominal/implicit condition (23%) had an accuracy score above 0.20.

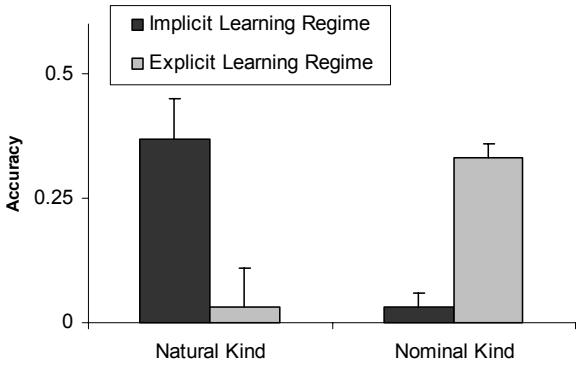


Figure 2: Accuracy Scores for Children.
Error bars represent standard errors.

In short, as predicted, children could learn the natural-type concept in the implicit learning regime, but not in the explicit learning regime; and they could learn the nominal-type concept in the explicit learning regime but not in the implicit learning regime. These findings cannot be due to differences in stimuli, as the same cartoon animals were used for both natural and nominal kinds. Furthermore, the findings cannot be due to differences in procedure, given that the learning regime for the natural kind (either implicit or explicit) was closely matched with the learning regime for the nominal kind.

These results reveal an important dissociation: while the implicit learning regime favored acquisition of concepts resembling natural kinds, the explicit learning regime favored acquisition of concepts resembling nominal kinds, thus supporting the contention that there is a psychological distinction between natural kinds and nominal kinds.

Experiment 2

The goal of this experiment was to extend the findings of Experiment 1 to adult participants. Adults participated in the same four conditions that were used for children in Experiment 1: natural/implicit, natural/explicit, nominal/ implicit, and nominal/explicit.

Method

Participants Participants were 54 introductory psychology students at a large mid-western university who participated for class credit. Additional nine adults (two or three in each condition) were tested and omitted from the sample because their performance in the catch trials did not meet the criterion (see Procedure).

Stimuli The stimuli were identical to the ones used in Experiment 1.

Procedure Adults were asked to learn about creatures called Ziblets in order to distinguish them from creatures that are not Ziblets. In the implicit learning regime, they were presented with 24 instances and asked to remember them. In the explicit learning regime, they were given the same defining rule that was presented to children. Again, no instances were presented; only pictures of the parts accompanied the rule.

Thirty-two testing trials followed in which instances were presented on the screen, and adults had to determine whether they see a Ziblet or not. Half of the instances were from the target category (Ziblets) and half of them were from the contrasting category (Non-Ziblets).

Eight catch trials followed intermixed with three trials from the target category. To be included in the study, adults had to respond correctly in at least 6 catch trials. At the end of the procedure, adults were asked to give a verbal description of the difference between Ziblets and other animals presented on the screen.

Results and Discussion

Mean accuracy scores are presented in Figure 3 (with standard error as error bars). A 2 (concept: natural, nominal) by 2 (learning regime: implicit, explicit) between-subjects ANOVA revealed a significant effect of learning regime ($F(1,50) = 13.15$, $p < .01$) with accuracy scores being higher in the explicit learning regime ($M = 0.57$, $SD = .45$) than in the implicit learning regime ($M = 0.23$, $SD = .34$), and a significant interaction ($F(1,50) = 11.9$, $p < .01$). When presented with the natural kind, participants performed above chance in both learning regimes ($t_{implicit}(13) = 5.3$, $p < .001$; $t_{explicit}(13) = 3.68$, $p < .01$), but when presented with

the nominal kind, they performed above chance only in the explicit learning regime ($t(10) = 6.12$, $p < .001$).

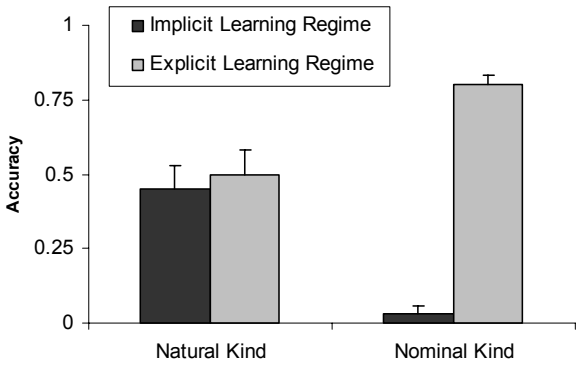


Figure 3: Accuracy Scores for Adults.
Error bars represent standard errors.

Adults' verbal responses were analyzed in terms of whether or not they contained the defining rule. For the natural kind, a response was coded as correct when the statement included at least one of the correlations. For the nominal kind, a response was coded as correct when the statement included the numerical relation. Table 2 shows the pattern of results. As expected, adults could verbalize the defining rule of the nominal kind in the explicit but not in the implicit learning regime. For the natural kind, even though adults' categorization accuracy did not differ as a function of learning regime, their verbal responses did. Only three the adults could verbalize the rule of natural kinds in the implicit learning regime whereas seven adults could verbalize the rule in the explicit learning regime[1].

Table 2: Number of Correct Verbal Statements
(Percentage correct in parentheses).

	Concept	
Learning Regime	Natural	Nominal
Implicit	3 (21%)	0
Explicit	7 (50%)	8 (73%)

Overall, learning of nominal kinds in adults exhibited tendencies similar to those in young children: participants ably learned the concept when presented with the defining rule of the concept, and they performed poorly when they were presented with instances of the category.

At the same time, unlike young children for the natural-kind concept, adults performed equally well under different learning regimes. This was surprising, given that the rule of the natural concept was rather lengthy, involving statements about the characteristics of six parts. We content that real natural kinds involve more than six simple correlations, thus making explicit learning of real natural kinds more difficult than explicit learning of current categories. This contention, however, remains speculative, and it will be examined in future research.

[1] The finding that some adults failed to verbalize the correct rule even in the explicit learning regime may be an artifact of the procedure. Instead of describing the difference between Ziblets and Non-Ziblets, a large majority of the adults described the difference between test items and catch items.

Experiment 3

The goal of this experiment was to document that the dissociation found in Experiments 1 and 2 is not limited to the particular nominal kind used in those experiments. Recall that the nominal kind used in Experiments 1 and 2 was based on a mathematical relation – a relation that may differ considerably from the correlations relevant for the natural kind. This difference was minimized in Experiment 3 by using the same target category that was used for the "natural kind" in the previous experiments. The contrasting category was new. It was constructed in such a way that only one correlation – rather than multiple correlations – was violated. To distinguish between target category and contrasting category, adults had to keep in mind all correlations. Therefore, the category to be learned was statistically sparse (all correlation mattered, no redundancy was present) without differing in content form the correlations of the "natural kind".

Method

Participants A new group of 28 students participated in this experiment (14 in the implicit learning regime, and 14 in the explicit learning regime). Additional 3 adults were tested and omitted from the sample because their performance in the catch trials did not meet the criterion.

Stimuli The stimuli of the target category were identical to the ones used in Experiments 1 and 2. Table 3 shows in abstract notation the characteristics of the contrasting category (Non-Ziblets) used in this experiment. These items differ from the Ziblets in only one of the correlations, rather than in all three correlations.

Table 3: Exemplar Items of the Contrasting Category used in Experiment 3

	Contrasting Category					
	Item	Item	Item	Item	Item	Item
Parts	1	2	3	4	5	6
Body	0	1	1	1	1	1
Antennas	1	0	1	1	1	1
Horiz. wings	1	1	0	1	1	1
Tails	1	1	1	0	1	1
Vert. wings	1	1	1	1	0	1
Buttons	1	1	1	1	1	0

Note. The numbers refer to the values of the respective characteristics (1 = light, small, or few; 0 = dark, large, or many). The target category is the same as in Experiment 2 (shown in Table 1)

Procedure The procedure was identical to the procedure used for the natural-kind concept in Experiment 2. Participants were presented either with the implicit or the explicit learning regime.

Results and Discussion

Mean accuracy scores were calculated for each learning condition. A significant difference was found ($t(31) = 3.76$, $p < .001$), with adults in the implicit-learning condition performing worse than adults in the explicit-learning condition (implicit: $M = 0.17$, $SE = .05$; explicit: $M = 0.44$, $SE = .06$). These results further indicate that the dissociation between natural kinds and nominal kinds reflects the structure of the to-be-learned categories rather than the property of the particular relation used in Experiments 1 and 2.

General Discussion

The results reported here support a psychological distinction between concepts that differ in their statistical structure. Concepts that are based on highly redundant features were best learned through an implicit learning regime (especially for children); and concepts that are based on non-redundant features were best learned through an explicit learning regime. The latter findings applied whether the concept was based on a mathematical relation (Experiment 2) or on a set of correlations between two features (Experiment 3). This suggests that the dissociation in acquisition reflects the statistical structure of the category rather than the particular relation.

Though not directly investigated in these sets of experiments, we argue that statistically dense concepts resemble natural kinds, while statistically sparse concepts resemble nominal kinds. It is likely then that natural kinds (e.g., the concept of *bird*) require a different learning environment than nominal kinds (e.g., the concept of *acceleration*). Furthermore, it is possible that this learning dissociation reflects itself in the way the concepts are represented. For example, it is possible that effects of gradedness are more likely to be found with natural kinds than with nominal kinds.

Finding dissociation in learning between different types of concepts indicates that a theory of categorization is incomplete if it pertains only to one kind of concept. A more complete account would address the processes of categorization for both natural and nominal kinds.

Acknowledgments

This research is supported by a grant from the National Science Foundation (REC # 0208103) to Vladimir M. Sloutsky.

References

Billman, A. J., & Knutson, D. (1996). Unsupervised concept learning. *Journal of Experimental Psychology, 22*(2), 458-487.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. Wiley.

Gelman, S. A., & Coley, J. (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman, & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development*. New York: Cambridge University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65*, 103-135.

Kripke, S. (1972). Naming and necessity. In D. Davidson & G. Harman (Eds.), *Semantics of natural language*. Dordrecht, Holland: Reider.

Mareschal, D., & Quinn, P. C. (2001). Categorization in infancy. Trends in Cognitive Sciences, 5, 443-450.

Medin, D. L., (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1481.

Mervis, C. G., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-116.

Murphy, G. L. (2002). *The big book of concepts*. MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39-57.

Nosofsky, R. M. (1992) Similarity scaling and cognitive process models. *Annual Review of Psychology, 43*, 25-53 1992

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.,

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception, 22*, 463-475.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences, 7*, 246-251.

Smith, E. E., & Medin, D, L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986/1934). *Thought and language*. MIT Press.

Younger, B. (1993). Understanding category members as "the same sort of thing": Explicit categorization in ten-month infants. *Child Development, 64*, 309-320.

Visual Imagery in Deductive Reasoning: Results from experiments with sighted, blindfolded, and congenitally totally blind persons

Markus Knauff (markus.knauff@tuebingen.mpg.de)
Max-Planck-Institute for Biological Cybernetics
Tübingen, Germany

Elisabeth May (elisabeth.may@mail.uni-oldenburg.de)
Department of Psychology, University of Oldenburg
Oldenburg, Germany

Abstract

We report three experiments on visual mental imagery in deductive reasoning. Reasoning performance of sighted participants was impeded if the materials were easy to envisage as visual mental images. Congenitally totally blind participants did not show this visual-impedance effect. Blindfolded participants with normal vision showed the same pattern of performance as the sighted. We conclude that irrelevant visual detail can be a nuisance in reasoning and impedes the process.

Introduction

Various sorts of evidence are compatible with the conjecture that visual mental imagery is a vital part of human cognition, including the famous studies of the mental rotation and the mental scanning of images (Shepard & Cooper, 1982; Kosslyn, 1980). The aim of the present paper, however, is to show that visual mental imagery is *not* necessary in reasoning. It can even be a nuisance in reasoning and impedes the process. The article is motivated by the distinction between visual and spatial representations and processes that has been introduced by Ungerleider and Mishkin (1982). In addition, the article is motivated by studies showing that congenitally totally blind persons are as good as sighted in the construction and application of spatial representations (e.g. Kerr, 1983), but differ from sighted people in their use visual images.

The paper begins with a brief summary of previous findings on imagery and reasoning. We focus on deductive reasoning, in which the truth of the premises ensures the truth of the conclusion. We then outline our hypothesis regarding the connection between visual images, spatial representations, and congenital blindness. We report three experiments that test this hypothesis. Finally, we draw some general conclusions about visual imagery, spatial representations, and reasoning.

An influential study of imagery and deductive reasoning was carried out by DeSoto, London, and Handel (1965), who investigated so-called three-term series problems, such as "Ann is taller than Beth.", "Cath is shorter than Beth", "Who is tallest?" and argued that reasoners represent the three individuals in a visual image, and then "read off" the answer by inspecting the image. There are several other authors who argued in the same vein (e.g. Huttenlocher, 1968; Shaver, Pierson, & Lang, 1976; Clement & Falmagne, 1986). Other authors did not find evidence that imagery plays a role in reasoning (e.g. Sternberg, 1980, Richardson, 1987; Johnson-Laird, Byrne, & Tabossi; 1989; Newstead, Pollard, & Griggs, 1986). In Knauff and Johnson-Laird (2000; 2002) we argued that a possible resolution of the inconsistency in the previous results is that these studies have overlooked the distinction between visual images and spatial representations (e.g. Ungerleider & Mishkin, 1982; Logie, 1995; Smith et al., 1995). We conducted a series of experiments to test this hypothesis. We initially accomplished rating studies to identify a set of verbal relations that varied in the ease of constructing visual images and spatial representations from it. Their results yielded three sorts of verbal relations:

1. *visuospatial relations* that are easy to envisage visually and spatially; e.g. above – below;
2. *visual relations* that are easy to envisage visually but hard to envisage spatially, e.g. cleaner – dirtier;
3. *control relations* that are hard to envisage both visually and spatially, e.g. smarter – dumber

From the three sorts of verbal relations we constructed a set of three-term- and four-term-series problems. In three experiments, visual relations such as *cleaner* and *dirtier* significantly impeded the process of reasoning in comparison with control relations such as *smarter* and *dumber*. In contrast, visuospatial relations, such as *front* and *back*, which are easy to envisage visually and spatially, speeded up the process of reasoning in comparison with control relations (Knauff & Johnson-Laird, 2002). In a subsequent neuroimaging study (fMRI) we showed that in the absence of any correlated visual input, all types of reasoning problems evoke activity in spatial areas of the brain (right superior parietal cortex, and bilaterally in the precuneus), but that only the problems based on visual relations also activated early visual areas corresponding to V2 (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003). We explained the findings by an interplay of visual images and spatial representations. For example, given the premises:

The cat is above the ape.
The dog is below the ape.

the participants construct a spatial array representing the relative positions of the three individuals:

cat
ape
dog

They evaluate a possible conclusion by checking whether it holds in the representation. Perhaps the ability to envisage spatial representations is a precursor to many forms of abstract reasoning (Johnson-Laird, 1996). Likewise, relational terms that lead naturally to spatial representations should speed up the process of reasoning. In contrast, a visual relation, such as *dirtier*, may elicit irrelevant visual detail. One imagines, say, a cat caked with mud, but such a representation is irrelevant to the transitive inference. It takes additional time to replace this vivid image with one in which dirtiness is represented in degrees. In other words, the visual relations, which are hard to envisage spatially, lead to a mental picture, but the vivid details in this picture impede the process of thinking.

If visual relations impede reasoning in sighted people, what happens if congenitally totally blind people reason with the same materials? In the last two decades, comparisons between blind and sighted people have been made on a large variety of visuospatial tasks, involving mental scanning, mental rotation, memory for paths and words, etc. (e.g. Kerr, 1983; Marmor & Zaback, 1976; Zimler & Keenan, 1983). They always reported the same results: people who are blind from birth are able to envisage abstract spatial arrangements, but unable to envisage visual mental images. Most of the explanations rely on the distinction of two different neural pathways associated with the processing of "what" and "where" information (Ungerleider & Mishkin, 1982). The distinction is well-established in numerous fields of cognitive science (e.g. Kosslyn, 1994; Landau & Jackendoff, 1993), and is supported by investigations with brain damaged patients (e.g., Newcombe, Ratcliff, & Damasio, 1987), neuroimaging studies (e.g., Smith et al., 1995), and experiments on visual and spatial working memory (c.f. Logie, 1995).

The "what" and "where" distinction in mental imagery has also been studied with congenitally blind participants. Vecchi (1998) conducted experiments in the dual-task paradigm with participants who were blind from birth and report that mental imagery can rely on purely spatial representations without a visual component. In a PET study, Büchel, Price, Frackoviak, and Friston (1998) demonstrated that congenitally blind people show task-specific activation in parietal association areas, whereas blind participants who lost their sight after puberty show additional activation in the primary visual cortex in the same task (Braille reading). Luzzatti et al. (1998) in a case study showed that visual and spatial imagery can be differentially impaired after brain injuries. All these studies clearly show that visual and spatial imagery are functionally independent processes which must rely on different neural systems.

What does that mean for the hypothesis that the ability to envisage spatial representations is a precursor to reasoning, but visual imagery can impede the process? The results concerning visual and spatial imagery in the congenitally blind motivate the following hypothesis:

Relations that elicit visual images containing details that are irrelevant to an inference should impede the process of reasoning in sighted people. They, however, should not hinder the reasoning of congenitally totally blind people, because they are able to construct spatial representations without being sidetracked by irrelevant visual images.

The aim of the following experiments is to test this hypothesis. In Experiment 1 sighted students solved three-term-series problems with the three sorts of verbal relations, in Experiment 2 people who were blind from birth, and in Experiment 3 sighted people who were blindfolded to remove any visual input.

Experiment 1: Sighted Participants

In our previous experiments (Knauff & Johnson-Laird, 2000; 2002) the reasoning problems were presented visually as sentences on the screen. The aim of the first experiment was to replicate the visual-impedance effect with sighted people but with an auditory presentation of the materials.

Participants. We tested 24 sighted undergraduate students from the University of Oldenburg (mean age 22.7; 18 female, 5 male), who received a course credit for their participation.

Materials. The experiment used the set of verbal relations that has been identified in Knauff and Johnson-Laird (2000; 2002). The three sorts of relations were:

1. *visuospatial relations:* above – below, front – back
2. *visual relations:* cleaner – dirtier, fatter — thinner
3. *control relations:* better — worse, smarter – dumber

From these verbal relations we constructed a set of three-term series problems which all concerned the same terms (*dog*, *cat*, *ape*). Here is an example of a problem with a valid conclusion:

The dog is cleaner than the cat.
The ape is dirtier than the cat.
Does it follow:
The dog is cleaner than the ape?

All sentences of the reasoning problems were recorded as audio files, edited for similar length and normalized for loudness and peak gain. Half of the problems had valid conclusions and half had invalid conclusions. The participants were told to evaluate whether the conclusion followed from the premises. In the example, cleaner and dirtier are used once in each premise, and cleaner occurs in the conclusion. But, in the experiment as a whole, each relation and its converse occurred equally often in each premise and in the conclusion.

Design. The participants acted as their own controls and evaluated 8 inferences of all three sorts (visuospatial, visual, and controls), making a total of 24 three-term series problems. The relations in these problems were those given above. The problems were presented in a randomized order over the set of participants.

Procedure. The participants were tested individually in a quiet room, and they sat at a PC that administered the experiment. The reasoning problems were presented in auditory format via headphones. The participants were told to evaluate whether or not the conclusion followed necessarily from the premises. They were instructed to respond as accu-

rately and quickly as possible. They made their response by pressing the appropriate key on the keyboard, and the computer recorded their response and latency. Prior to the experiment, there were four practice trials.

Results and Discussion. Table 1 presents for all three of our experiments the percentages of correct conclusions and their mean latencies for the different sorts of relational inferences. The present inferences were relatively easy (92.5% correct overall) and there was no significant difference between accepting valid conclusions (94.5% correct) and rejecting invalid conclusions (95.2% correct). Thus, we pooled the results from these conditions.

The MANOVA for dependent measures on the accuracy data revealed a reliable difference across the three sorts of problems, $F(2, 46) = 5.30$, $p < .01$. There was also a difference in the mean number of correct responses between the visual problems and the control problems, $F(1, 23) = 6.57$, $p > .02$, and the (not orthogonal) contrast between visuospatial and visual problems was significant, $F(1, 23) = 6.57$, $p < .02$. The response latencies also showed the predicted trend (visual relations resulted in longer response latencies than control problems and reasoning with these problems in turn took longer than with visuospatial problems), but the main effect across the three sorts of problems was statistically not significant, $F(2, 46) = .53$, $p > .6$.

The main goal of the experiment was to test a new experimental setup that can be used with the blind people later on. The experiment was successful in showing that the visual-impedance-effect also appears with the auditory presentation of the materials. Thus, we can use the same experimental setting in the later studies. A second corollary from the findings is that the visual-impedance effect does not depend on the visual presentation of the materials. It rules out that the impedance is simply due to interference between the visual process of reading the premises and conclusions and the mental activity of envisaging a visual mental image to solve the problem. Instead, the findings again emphasize the importance of distinguishing between visual and spatial representations. Visual relations such as *fatter* and *thinner* impeded the process of reasoning in comparison with control relations such as *smarter* and *dumber*. In other words, the visual relations, which are hard to envisage spatially, lead to a mental picture, but the vivid details in this picture impede the process of thinking.

Experiment 2: Congenitally Totally Blind Participants

This experiment directly tests the hypothesis that the visual relations do not impede the reasoning of congenitally totally blind people, because they are able to construct spatial representations without being sidetracked by irrelevant visual images.

Participants. We tested 10 congenitally totally blind participants (mean age 24.8; 7 female, 3 male). According to the German Law, a person is congenially totally blind, if she or he has less than 5% of normal vision and got blind before the age of 2. Most of the participants were blind from birth due to retinopathy of prematurity. The participants we recruited from two self-helping groups for the blind. All participants gave their informed consent prior to the participation in the study.

Materials, Design, and Procedure. The materials and the design were identical to Experiment 1. The instructions were read to the participants by one of the experimenters. The participants were tested individually in a quiet room at the institutions, and they sat in front of a Laptop that administered the experiment. Except of the two keys associated with "yes" and "no" and the spacebar, all other keys were removed from an external keyboard.

Results and Discussion. The second row of Table 2 presents the mean latencies and correct responses to the three sorts of relational inferences. Overall, the blind responded correctly to 76.2 % of the inferences and again there was no significant difference between accepting valid conclusions (76.7% correct) and rejecting invalid conclusions (75.8% correct). Hence, we pooled the results from these conditions.

The ANOVA showed no significant difference in reasoning accuracy between the three sorts of problems, $F(2, 18) = .36$, $p > .70$, and none of the single contrast revealed a significant difference (visual vs. control: $F(1, 9) = 2.38$, $p > .63$; visuospatial vs. control: $F(1, 9) = .70$, $p > .79$; visuospatial vs. visual: $F(1, 9) = .74$, $p > .41$). In the response latencies there was also no difference between the three sorts of problems, $F(2, 18) = .928$, $p > .41$, and not one of the single contrast showed a significant difference (visual vs. control: $F(1, 9) = 1.51$, $p > .24$; visuospatial vs. control: $F(1, 9) = 1.32$, $p > .28$; visuospatial vs. visual: $F(1, 9) = .35$, $p > .56$).

The results shed new light on the role of visual images and spatial representation in reasoning. Mental representations must be derived from perception and thus the representations of persons who are blind from birth must be different to that of sighted persons. In particular, haptics or auditory perceptions lead to spatial representations without a visual component. This account is supported by several studies that report the same *pattern of performance* in highly spatial tasks in sighted and congenitally blind persons. In a classical study by Kerr (1983) congenitally totally blind and sighted showed almost the same pattern of response times depending on imagined distance, image size, etc. Kerr concluded that "picturability" does not affect the recall of "mental images" in the blind. The only difference was that sighted participants reported forming the images while the blind did not (or at least significantly slower). Marmor and Zaback (1976) explored Shepard and Metzler's mental rotation tasks and found that blind people also show longer reaction times for larger rotation angles. Zimler and Keenan (1983) found similar results in congenitally blind children and adults. In addition, they reported that the haptic images of the blind maintain the same *spatial* information just as the visual images of the sighted do. Obviously, people who are blind from birth do not tend to construct visual mental images. But they are able to construct and to employ spatial representations. In fact, most of our blind partici-

pants reported not using visual images. Instead, they reported that they located the objects of the inference on a spatial scale or in degrees, representing, say "dirtiness". Although such introspections certainly can be wrong, they agree with the experimental findings: The blind are able to construct spatial representations without being irritated by irrelevant visual images.

Experiment 3: Blindfolded Participants

Is there an alternative explanation for the different patterns of results in sighted and blind participants? One possible account is that the visual-impedance effect in the sighted is simply due to interference between the visual input from the surrounding and the mental activity of envisaging a visual mental image. To rule out this explanation in the third experiment the participants had normal vision, but were blindfolded to eliminate any visual input. If the visual-impedance-effect is due to interference between visual imagery and visual perception, they should be also resistant to the impedance effect of visual relations—much as the congenitally blind people are. If, in contrast, the tendency of sighted people to construct visual images is responsible for the visual-impedance effect, then the pattern of results should be similar to that in Experiment 1.

Participants. We tested 30 sighted undergraduate students of University of Oldenburg (mean age 23.3; 18 female, 10 male). They were completely blindfolded to remove any visual input. They received a course credit for their participation.

Materials, Design, and Procedure. The design, the materials and the procedure were identical to Experiment 1 and 2. As in Experiment 2, the instructions were read to the participants by one of the experimenters and except for the two keys associated with "yes" and "no" and the spacebar, all other keys were removed from an external keyboard.

Results and Discussion. Overall, there were 90.8% correct responses and there was no significant difference between accepting valid conclusions (90.2% correct) and rejecting invalid conclusions (91.6% correct). They were pooled again. The MANOVA showed a reliable difference in accuracy across the three sorts of problems, $F(2, 58) = 3.71$, $p < .03$. There was also a significant difference in the mean number of correct responses between the visual problems and the control problems, $F(1, 29) = 5.80$, $p < .03$ and the visual problems and the visuospatial problems, $F(1, 29) = 4.29$ $p < .05$

The response latencies showed that visual problems were slower than control problems, which in turn were slower than the visuospatial problems. The main effect across the three types of problems is statistically significant, $F(2, 58) = 4.22$, $p < .02$. The difference between visual and control relations did not reach statistical significance, $F(1, 29) = 2.25$ $p > .14$, but the difference between visual and visuospatial problems was reliable, $F(1, 29) = 7.01$, $p < .02$.

The pattern of performance in the blindfolded participants is almost identical to that of the sighted participants in Experiment 1. There was again the trend visual > control > visuospatial (although the single contrast between visual and control problems did not reach statistical significance in the lantencies). These data clearly show that the characteristics of the reasoning problems lead to the visual-impedance effect. It is not simply due to interference between the visual input from the surrounding and the mental activity of envisaging a visual mental image.

Table 1: Percentages of correct responses and their mean response latencies (in s) in the three experiments as a function of the different sorts of relations: visual relations, control relations, visuospatial relations.

	Visual inferences	Control inferences	Visuospatial inferences
sighted (Exp 1)	86.9%	94.8%	94.8%
	1.26	1.01	.97
blind (Exp 2)	75.0%	73.6%	79.9%
	5.24	5.29	6.06
blindfolded (Exp 3)	86.7%	92.9%	92.9%
	1.42	1.08	.86

General Discussion

The starting point of our studies was the distinction between visual and spatial modes of representation in reasoning. Previous studies enabled us to identify visuospatial relations, such as *above-below*, which are easy to envisage both visually and spatially, visual relations, such as *cleaner-dirtier*, which are easy to envisage visually but hard to envisage spatially, and control relations, such as *better-worse*, which are hard to envisage both visually and spatially. Our former studies showed that visual relations significantly impede the process of reasoning by slowing it down. We refer to that as *visual-impedance-effect* (Knauff & Johnson-Laird, 2002).

In the present studies we tested a group of sighted participants, one group of congenitally totally blind participants, and one group of blindfolded participants with normal vision. In the sighted participants the visual relations significantly impeded the process of reasoning by slowing it down. The blindfolded did show the same impedance effect. But, the participants who were blind from birth were not affected by the ease of envisaging the verbal relations visually. They showed the same reasoning performance across all sorts of problems.

What might cause visual imagery to impede reasoning in the sighed participants? A theory that relies on visual imagery as the medium for reasoning is implausible, because individuals can reason about relations that they cannot visualize. Similarly, such a theory cannot readily explain why relations that are easy to envisage visually impeded reasoning. One could object that the ability to visualize the visual relations was impeded by the concurrent visual perception. In fact, several studies have shown that visual imagery and visual perception interfere and that visual imagery performance is impaired under this condition (c.f Logie, 1995). Our

third experiment with the blindfolded participants, however, clearly falsifies this hypothesis. An explanation based on formal inference rules (Rips, 1994; Braine & O'Brien, 1998), is also questionable, because it does not account for the effects of content, and does not immediately suggest an explanation of the visual impedance effect.

The present findings support a spatial account of reasoning. The initial idea has been introduced by Huttenlocher (1968) and was further elaborated in the mental models theory of reasoning (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). The model theory does not rely on linguistic processes like rule-based-approaches of reasoning (Braine & O'Brein, 1998; Rips, 1994). Such processes are relevant only to transfer the information from the premises into a spatial array and back again, but the reasoning process itself totally relies on non-linguistic processes for the construction and inspection of spatial mental models. The mental models mirror the spatial relations between the represented objects. In contrast to visual images, mental models can represent any possible situation and can abstract away from such visual details as colors, textures, and shapes (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Mental models can also represent class-inclusion, temporal order, and abstract relations such as ownership (c.f. Johnson-Laird & Byrne, 1991). Several studies have shown that the content can facilitate inferences in certain cases and impede them in other cases (e.g. Johnson-Laird & Byrne, 2001). Likewise, a visual relation, such as *dirtier than*, can elicit a vivid visual detail, such as an animal caked with mud that is irrelevant to an inference. It will then take additional time to retrieve the information needed to construct the appropriate spatial mental model for making the inference.

The blind participants, however, did not show this visual impedance effect. This provides additional support for the spatial account of reasoning. Clearly, people who are blind from birth do not tend to use visual mental images, unless they are forced to do so, as in the studies from the literature. But they are able to construct and to employ spatial representations. Such models represent the objects of the inference in degree or on a spatial scale. For this reason, they are not sidetracked by irrelevant visual images and thus perform *relatively* better than sighted persons in the visual problems.

There are, however, some ambiguities in the data from the blind. First, the data are in line with other studies that compared blind and sighted people on a large variety of visuospatial tasks. They consistently reported that blind persons in *absolute* terms perform less accurately or more slowly than the sighted on such tasks (e.g. Kerr, 1983). Such an overall deficit of the blind participants is also visible in the present studies. The sighted participants solved on average 92.2% of the inferences correctly, but the blind participants only 76.2%. The sighted needed 1.08 seconds on average to respond to a problem; the participants who were blind from birth needed 5.3 seconds. Even the blindfolded participants from Experiment 3 performed much better than the blind persons. The dominant approach to explain such findings runs somewhat counter to our own account. It is usually seen as a *visual imagery deficit* of the congenitally blind. In particular, haptics or auditory perceptions also lead to spatial representations, but it is argued that these representations might be sub-optimal compared to vision-based representations (a recent discussion can be found in Fleming, Ball, Collins, & Ormerod, in press). From this view, our blind participants show less good performance, because they are less good in visual mental imagery. However, this account cannot readily explain why the impedance effect of visual relations disappears in the blind. If a visual imagery deficit is responsible for the overall performance deficit of the blind, the impedance effect should be even more pronounced in the blind compared to the sighted.

A second critical aspect in the data is that 9 out of 10 blind participants showed a minor increase of response latencies in the visual problems. The differences were particularly small and apparent only in the visual inspection of the data. However, one could argue that under these conditions the non-effect in the MANOVA (and in the post-hoc computed nonparametric tests) is just due to the small number of participants. If the effect would turn out to be reliable it could have two causes: It could either indicate that even the blind try to envisage the visual relations in a visual mental image (what is implausible due to the above reasons) or that the relations differ in the degree to which they imply transitivity. Spatial relations are unequivocal, but visual relations might be more dubious. Given, say, the following premises:

The cat is fatter than the ape.
The ape is fatter than the dog.

Reasoners might have wondered whether the fatness of cats, apes, and dogs, is commensurable. They claim that, say, an elephant is thin is relative to elephants, and so it is sensible to assert that a thin elephant is fatter than a fat dog. The criterion for fatness shifts from one animal to another. This factor might have confused reasoners in our experiment, and impeded their inferences with the visual relations. A related factor is the degree to which the premises accord with the participants' existing beliefs. For example, the preceding premise (The cat is fatter than the ape) might strike some individuals as implausible. However, these explanations are unlikely, because in the experiments as a whole, each such plausible premise is matched with one using the converse relation (The cat is thinner the ape), and so this factor seems not likely to account for our results.

A last point is that we did not use purely spatial relations, i.e., those that are hard to envisage visually but easy to envisage spatially. If, as our findings suggest, the visual character of the materials leads to an impairment of reasoning performance, whereas the possibility of spatially envisaging the materials speeds up reasoning, then tasks based on purely spatial relations should be processed most quickly. Indeed, we found a (not significant) trend in this direction in Knauff & Johnson-Laird (2002). However, we encountered some technical problems with these relations and some colleagues overall doubt the existence of such relations.

In conclusion, our results suggest that the content of verbal relations can affect the process of inference. If the con-

tent yields information relevant to an inference, as it does with visuospatial relations, then reasoning proceed smoothly, and may even be slightly faster than with other sorts of content. But, if the content yields visual images that are irrelevant to an inference, as it does with visual relations, then reasoning of sighted persons is impeded and takes reliably longer. People who are blind from birth are *immune* to such impedance effects, since they do not tend to use disrupting visual images. A word of caution, on the other hand, is that the visual and spatial nature of representations in reasoning also depends on the nature of the problem. Transitive inferences might elicit spatial representations, but that does not rule out that other problems rely on visual images in addition.

Acknowledgments

M.K. is supported by a Heisenberg Award from the Deutsche Forschungsgemeinschaft (DFG) and by the grants Kn465/2-4 and BACKSPACE in the Transregional Collaborative Research Center on Spatial Cognition (SFB/TR 8). The authors are grateful to Jack Loomis, who had the initial idea of testing the visual-impedance-hypothesis with blind people. We are also indebted to three anonymous reviewers and to Mrs. Backsmann and Mr. Becker from the self-helping groups in Oldenburg and Hannover.

References

Büchel C, Price C, Frackowiak R.S, & Friston K (1998). Different activation patterns in the visual cortex of late and congenitally blind subjects, *Brain*, 121, 409-19.

Braine, M. D. S., & O'Brein, D. P. (Eds.) (1998). *Mental logic*. Mahwah, NJ: Erlbaum.

DeSoto, L B., London, M., & Handel, M.S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2, 513–521.

Fleming, P., Ball, L.J., Collins, A.F., & Ormerod, T.C. (in press). Spatial representation and processing in the congenitally blind. Chapter in S. Ballesteros, & M.A. Heller (Eds.): *Touch, Blindness, and Neuroscience.* Madrid: UNED Press.

Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review*, 75, 550–560.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. (1991). *Deduction*. Hove, UK: Erlbaum.

Johnson-Laird, P.N., & Byrne, R. M.J. (2001) Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, in press.

Johnson-Laird, P. N., Byrne, R., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantifiers. *Psychological Review*, 96, 658-673.

Kerr, N.H. (1983). The role of vision in "visual imagery" experiments: Evidence from the congenitally blind. *Journal of Experimental Psychology: General*, 112, 265-277.

Knauff, M., Fangmeier, T., Ruff, C. C. & Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*, 4, 559-573.

Knauff, M. & Johnson-Laird (2000). Visual and spatial representations in spatial reasoning. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 759-765). Mahwah, NJ: Erlbaum.

Knauff, M. & Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. *Memory & Cognition*, 30, 363-371.

Kosslyn, S.M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1994). *Image and brain*. Cambridge, MA: MIT Press.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences,* 16, 217-265.

Logie, R.H. (1995). *Visuospatial working memory*. Hove, UK: Erlbaum.

Luzzatti C., Vecchi T., Agazzi D., Cesa-Bianchi M., & Vergani C. (1998). A neurological dissociation between preserved visual and impaired spatial processing in mental imagery. *Cortex*, 34, 461-469.

Marmor, G., & Zaback, L. (1976). "Mental rotation by the blind: Does mental rotation depend on visual imagery?" *Journal of Experimental Psychology: Human Perception and Performance,* 2, 515-521.

Newcombe, F., & Ratcliff, G. (1989). Disorders of visuospatial analysis. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology* (pp. 333-356). Amsterdam: Elsevier.

Newcombe, F., Ratcliff, G., & Damasio, H. (1987). Dissociable visual and spatial impairments following right posterior cerebral lesions: Clinical, neuropsychological and anatomical evidence. *Neuropsychologia*, 18, 149-161.

Newstead, S.E., Pollard P., & Griggs, R. A. (1986). Response bias in relational reasoning. *Bulletin of the Psychonomic Society*, 2, 95-98.

Richardson, J.T.E. (1987). The role of mental imagery in models of transitive inference. *British Journal of Psychology*, 78, 189-203.

Rips, L. J. (1994). The psychology of proof. Cambridge, MA: MIT Press.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.

Smith, E. E., Jonides, J., Koeppe, R. A., Awh, E., Schuhmacher, E. H., & Minoshima, S. (1995). Spatial versus object working memory: PET investigations. *Journal of Cognitive Neuroscience*, 7, 337-356.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, 109, 119-159.

Ungerleider, L.G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A.Goodale, & R. J. W. Mansfield (Eds*.), Analysis of visual behaviour* (pp. 549-586). Cambridge, MA: MIT Press.

Vecchi, T. (1998). Visuospatial imagery in congenitally totally blind people, *Memory*, 6, 91-102.

Zimler, J., & Keenan, J.M. (1983). Imagery in the congenitally blind: How visual are visual images? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 9, 269-282.

Stored Knowledge versus Depicted Events: what guides auditory sentence comprehension?

Pia Knoeferle (knoeferle@coli.uni-sb.de)
Department of Computational Linguistics,
Saarland University, 66041 Saarbrücken, Germany

Matthew W. Crocker (crocker@coli.uni-sb.de)
Department of Computational Linguistics,
Saarland University, 66041 Saarbrücken, Germany

Abstract

In a seminal article, Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy (1995) showed that eye-movements to real-world objects reflect a rapid interplay of utterance and visual environments in sentence comprehension. Further, Kamide, Scheepers & Altmann (2003) found that when *linguistic and world knowledge* constrain the domain of reference in a visual scene, people even anticipate as yet unmentioned arguments/referents. Studies by Knoeferle, Crocker, Scheepers & Pickering (2003) have since revealed that when linguistic and world knowledge did not disambiguate an initial syntactic and role ambiguity, *depicted agent-action-patient events* permitted anticipation of thematic role-fillers online. This paper opposes linguistic and world knowledge on the one hand, and visual scenes on the other hand in order to determine their relative importance in auditory comprehension. We observed a preferred reliance of auditory sentence comprehension processes on information that had to be extracted from depicted event scenes. Determining the nature and time-course of the interaction between linguistic/world knowledge and visual scenes is a first step towards developing a theory of real-time auditory sentence comprehension in visual environments. Our finding has implications for theories of the language faculty (e.g., Jackendoff, 2002).

Introduction

How do we understand utterances online in visual environments? As a first hypothesis, we might assume that the presence of visual environments does not affect language comprehension processes. Crucially, however, language refers to things in the world. It has further been demonstrated that in on-line comprehension reference to entities is established rapidly, and that referentially relevant non-linguistic information influences how the linguistic input is structured (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). In auditory comprehension in visual environments, the unfolding utterance provides linguistic, semantic, and world knowledge. The scene in turn *affords* information about entities and events in the immediate environment (e.g., an outstretched hand affords the prospect of shaking hands) (Gibson, 1966; see Steedman, 2002, for a formal description of object and event affordances). Mental representations that issue in parallel from distinct cognitive components such as the auditory and visual system have to be integrated in "real-time". Comprehension in visual scenes hence moves from single-mode understanding to bi- or even multi-modal understanding.

The architecture of the language faculty

Since visual scene and linguistic information interact in online comprehension (e.g., Kamide et al., 2003; Knoeferle et al., 2003; Tanenhaus et al., 1995), adequate description of auditory comprehension in visual environments requires that we embrace a theoretical account which situates language comprehension with respect to other cognitive systems such as the visual, auditory, or motor system. Jackendoff (2002) proposes one such framework. The individual levels in Jackendoff's architecture are modular in the sense of being domain-specific (i.e., their representational vocabulary is specialized), but unlike Fodorian modularity (Fodor, 1983), linguistic structures interact with one another, and with other cognitive sub-systems. This variety of modularity permits communication between phonological, syntactic, and conceptual structure. Moreover, it also allows communication between conceptual structure and perception or action via interface processors (Jackendoff, 2002, pp. 220f.).

While Jackendoff's theory provides for the interaction of linguistic/world knowledge and visual information, it is underspecified with respect to the precise nature and time-course of this interaction. In order to develop a theory of real-time comprehension on the basis of his architecture, further experimental work is required. As a first step in this direction, we need to clarify how linguistic and world knowledge is integrated with object and event affordances that have to be extracted from the visual environment.

Previous work emphasizes both the importance of visual scenes in determining online comprehension when the utterance was structurally ambiguous (Tanenhaus et al., 1995; Knoeferle et al., 2003), and the importance of stored linguistic/world knowledge in anticipating which object in a scene will be referred to next (Kamide et al., 2003)[1].

[1] We use the term 'stored knowledge' to refer to linguistic/world knowledge which is stored in memory. The terms 'world knowledge' and 'stereotypical knowledge' are used synonymously to mean stereotypical relationships between scene entities (e.g., a cat is a stereotypical agent for chasing mice).

Stored knowledge versus depicted event scenes

Kamide et al. (2003) have shown that unambiguous case-marking, lexical expectations and world-knowledge influence anticipation of post-verbal arguments depicted in the scene. In German, a case-marked article can determine the grammatical function and thematic role of the noun phrase it modifies. Both SVO (subject-verb-object) and OVS (object-verb-subject) orders are grammatical. Participants inspected images showing a hare, a cabbage, a fox and a distractor object while hearing sentences such as *Der Hase frisst gleich den Kohl* ('The hare (subj) eats soon the cabbage (obj)') and *Den Hasen frisst gleich der Fuchs* ('The hare (obj) eats soon the fox (subj)'). The subject and object case-marking on the article of the first noun phrase together with world knowledge extracted at the verb allowed anticipation of the correct post-verbal referent. This was evidenced by anticipatory eye-movements to the cabbage after participants had heard 'The hare (subj) eats …' and to the fox after having encountered 'The hare (obj) eats …'. Hence, when the utterance is unambiguous, and linguistic/world knowledge restricts the domain of potential referents in a scene, the comprehension system may anticipate mention of scene objects.

Knoeferle et al. (2003, accepted), in contrast, considered ambiguous utterances, where neither case-marking nor stereotypical knowledge could assist in disambiguation. Specifically, they examined the time course with which listeners were able to resolve an initial structural and thematic role ambiguity in German sentences. As the linguistic input did not determine the correct syntactic analysis and thematic role-assignment of the sentence, listeners had to rely on depicted events in the scene for interpretation of the utterance. The events showed, e.g., a princess washing a pirate, while a fencer painted her. The princess was thus determined as either patient or agent of an event depending on the depicted action (washing/painting respectively). Listeners heard *Die Prinzessin wäscht/malt den Pirat/der Fechter*. ('The princess (amb.) washes/paints the pirate (obj./patient)/the fencer (subj./agent)'). Once the verb had identified the relevant depicted action, anticipatory eye-movements to the appropriate other event participant (the pirate or the fencer) were observed. The anticipation of a patient and agent role-filler for initially ambiguous German subject-verb-object and object-verb-subject sentences respectively suggests rapid use of depicted events in resolving structural and thematic role ambiguity online. This finding was also shown for the English main clause (MC)/reduced relative (RR) ambiguity, and hence generalized to another language and construction.

Teasing apart the relative effects of visually perceived events and stored linguistic/world knowledge in online sentence comprehension is of relevance for theories of the architecture of the language system such as the one proposed by Jackendoff (2002). Such endeavor may ultimately allow us to propose a more concrete theory of processing mechanisms within such a framework and to hence develop it into a situated theory of real-time sentence comprehension. What the above studies have shown, is that stored knowledge and visual-scene information are both rapidly applied, and that each may guide comprehension processes online. It is further clear, that utterance, world knowledge and the immediate visual scene interact during online comprehension. What remains unclear is the nature and time-course of that interaction. Among other questions, we might ask: What is the relative importance, or priority of different information sources, such as linguistic, scene, and world knowledge? Does scene information guide comprehension, or is the use of scene information determined by stored knowledge?

To further investigate this empirical question, consider an example from German. As noted above, German has a rich case marking system where grammatical function is usually indicated by unambiguous case morphemes. Word order constraints are less rigid in German than in English, and both subject-verb-object (SVO) and object-verb-subject (OVS) order are grammatical with SVO being the preferred reading (e.g., Hemforth, 1993). On hearing an OVS sentence fragment such as *The pilot* (object/patient[2]) *jinxes*… while inspecting the example scene in Figure 1, a number of processes occur. When we hear *The pilot*, object case-marking permits assignment of a patient role to the noun phrase while establishing reference to the pilot in the scene. As there are no constraints on inspecting the scene, perceivers might notice a wizard holding a telescope, and a character resembling a detective who is serving some food. Having encountered an agent and the verb, we might expect post-verbal mention of an agent. At this point, our knowledge that a wizard is a likely jinxing-agent can combine with the fact that the wizard is the only entity whose affordances match the expectations raised by the verb. The combination of entity affordance and stereotypical knowledge allows us to anticipate the wizard as a likely-to-be-mentioned agent. The decisive contribution is, however, made by the utterance, as the verb provides knowledge of stereotypical thematic role-fillers of a jinxing-action (a wizard). The scene affords no information about a jinxing-*relation* between two event participants, as there is no depicted jinxing-action. In contrast, when we hear *The pilot* (object/patient) *serves food to*…, verb-based knowledge of stereotypical agents of a serving-food action (e.g., a cook), cannot provide any guidance, as the scene affords no such entity. However, the scene does afford a depicted food-serving event performed by the detective. While stereotypical knowledge does not allow determination of thematic role-relations in this case, the affordances of the depicted scene events do. Based on findings by Kamide et al. (2003) and Knoeferle et al. (2003, accepted), we would expect unmistakable resolution of the temporal uncertainty regarding the yet-to-be-mentioned agent in both of the above examples once people have heard the verb.

Imagine we heard instead *The pilot* (object/patient) *spies-on*… In this case, the scene affords both a stereotypical

[2] In our materials, a grammatical subject and object correspond to an agent and patient respectively in the scenes.

agent (the detective), and an immediately depicted agent of a spying event (the wizard) as potential agents (see Fig. 1). When we hear *spies-on*, lexical access makes available the meaning of the lexical item and stereotypical knowledge related to it (see Ferretti, McRae & Hatherell, 2001). After encountering the verb, word meaning, stereotypical knowledge of *spy-on*, and scene affordances are available to anticipate either a depicted spying-event and its agent (the wizard), or a stereotypical agent (the detective).

Do listeners rely more on extracting thematic role relations from stereotypical knowledge provided by the utterance (*spy-on* + WORLD KNOWLEDGE –> *detective*), or do they rely on thematic relations afforded by the scene (*spy-on* + WIZARD-SPYING-EVENT -> *wizard*) in incremental interpretation? For the ambiguous *spy-on* example thematic role-relations that are provided by the visual scene, conflict with stereotypical knowledge of who-does-what-to-whom. The comprehension system has to choose between two available, yet conflicting types of information in determining online thematic role-assignment.

While Jackendoff's framework does not make explicit predictions, there are reasons to expect a priority of stereotypical/world knowledge in online thematic role-assignment in such architecture. Jackendoff (2002, pp. 282) argues against a strict separation of linguistic meaning and world knowledge (see also Levinson, 2000). Experimental evidence confirms such an assumption. Ferretti et al. (2001) found that verbs immediately activated stereotypical knowledge of agents (*arresting-cop*) or patients (*arresting-criminal*), but not locations (*swam-ocean*). They conclude that this type of world knowledge is part of thematic-role knowledge, and immediately activated upon encountering the verb (see also McRae, Ferretti & Amyote, 1997). Stored knowledge about stereotypical agents is hence readily available for online thematic role-assignment processes. Non-stereotypical thematic role-relations afforded by scene events, however, must be newly acquired by a perceiver, and via a different perceptual system (the visual system). On the basis of experimental evidence for the tight coupling between word meaning and world knowledge (e.g., Ferretti et al., 1997), we expect comprehension processes to rely in preference on stored knowledge over scene information. Indeed, such a prediction would also appear to follow from traditional assumptions concerning the modularity of the language faculty with respect to other perceptual faculties such as the visual system (e.g., Fodor, 1983).

In summary, we expect the following: when the verb determines either a depicted or a stereotypical target agent only (Table 1, a1 and a2), both verb-derived knowledge of stereotypical role-fillers and affordances of the scene events should allow anticipation of the appropriate target agent. This would replicate - within a single study - findings by Kamide et al. (2003) and Knoeferle et al. (2003). When the verb is *serves-food-to* (a1), we expect a higher percentage of anticipatory looks to the only depicted food-serving agent (the detective) than to the respective other agent in the scene (the wizard). Conversely, when the verb is *jinxes* (a2), more looks should occur to the stereotypical agent (the wizard) than to the other agent in the scene (the detective). In the interesting case of competition, when the verb (*spy-on*) allows more than one potential scene entity as target agent, no interaction is expected. Rather, we should observe a main effect, with the point of interest being the direction of the main effect. The fact that stored stereotypical knowledge is readily available from memory, argues for guidance of online thematic role-assignment processes by stereotypical knowledge rather than by event affordances acquired from the scene in a Jackendoffian framework. This should be revealed in a higher percentage of inspections to the stereotypical spying-agent (the detective) than to the depicted spying-agent (the wizard) for sentences b1 and b2 (see Fig. 1 and Table 1). Crucially, these looks should occur before people hear the disambiguating second noun, and hence reveal online expectations of thematic role interpretation.

Experiment

Method

Participants Twenty-four German native speakers with normal or corrected-to-normal vision were paid 5 euro for taking part in the experiment.

Materials We created 48 images using commercially available clipart and graphic programs. For each of these images, a female native German speaker recorded 4 sentences, which described either a depicted event (e.g., wizard-spying) or a stereotypical event (e.g., detective-spying, see Fig. 1; Table 1).

Design A set of 24 items was created. Each item consisted of 8 spoken sentences and 2 images (Table 1 and Fig. 1 show examples for the 4 sentences for one image of an item set). The two versions of an image only differed in the actions performed by the respective characters. This ensured that each of the target agents (wizard, detective) was a stereotypical and a depicted agent in turn, and that each verb referred once to a depicted event, and once to a stereotypical event. Actions were typically depicted as a character holding an instrument. The way in which actions or characters were depicted did not differ between the two image versions. The middle character on each image (e.g., the pilot) was always a patient ('being acted upon'). The entities to the left and the right of the patient character were performing an action upon the patient entity, and hence always had an agent-role. The two agents were balanced for position (left vs. right). An example image (see Fig. 1) showed two such agent-action-patient events, e.g., wizard-spying-on-pilot and detective-serving-food-to-pilot.

Table 1: Sentences for the example image in Figure 1

Image	Condition	Sentence	PATIENT	VERB	ADV	AGENT
Fig. 1	No-Competitor & depicted target	a1	Den Piloten	*verköstigt*	gleich	der Detektiv.
			The pilot (PAT.)	serves-food-to	soon	the detective (depicted AGENT)
			'The detective will soon serve food to the pilot.'			
Fig. 1	No-Competitor & stereotypical target	a2	Den Piloten	*verzaubert*	gleich	der Zauberer.
			The pilot (PAT.)	jinxes	soon	the wizard (stereotypical AGENT)
			'The wizard will soon jinx the pilot.'			
Fig. 1	Competitor & depicted target	b1	Den Piloten	*bespitzelt*	gleich	der Zauberer.
			The pilot (PAT.)	spies-on	soon	the wizard (depicted AGENT)
			'The wizard will soon spy on the pilot.'			
Fig. 1	Competitor & stereotypical target	b2	Den Piloten	*bespitzelt*	gleich	der Detektiv.
			The pilot (PAT.)	spies-on	soon	the detective (stereotypical AGENT)
			'The detective will soon spy on the pilot.'			

Figure 1: Example of an image for Experiment 1

In addition to the affordances of the depicted events (a telescope affording a spying action), each of the characters also had entity affordances (a detective affording a spying action). The event and entity affordances were always incongruous for any one item entity in this experiment (e.g., a detective was never depicted as performing a spying action). Rather, one agent on each image was a stereotypical competitor for the depicted event performed by the other agent (e.g., the detective was a stereotypical competitor for the depicted wizard-spying event), while carrying out a different action (serving-food) himself. Note, however, that 25 % of the fillers showed plausible events, e.g., a criminal being arrested by a cop). By manipulating the verb people heard, we created four conditions, crossing the factors *competitor* (competitor, no-competitor) with *information type* (depicted target, stereotypical target). For the no-competitor conditions (see Table 1, sentences a1 and a2), the verb permitted either a depicted or a stereotypical target only: "verköstigen" ('serve-food-to') determined the detective (Table 1, a1) as depicted agent; "verzaubern" ('jinxes') identified the wizard as stereotypical agent (Table 1, a2). For the competitor condition (see Table 1, sentences b1 and b2) the verb "bespitzeln" ('spy-on') allowed two scene entities as likely targets (Fig. 1): the wizard, being depicted as performing a spying-action, and the detective, a stereotypical agent for a spying-action. Sentences were unambiguous OVS sentences (see Table 1). They always started with an object case-marked noun phrase referring to a patient role-filler (Fig. 1, the pilot). The middle character was not engaged in an action, and its gaze and position did not bias towards either the left- or the rightward entity. Conditions were matched for length and frequency as much as possible (CELEX). For the image in Figure 1, the sentences in Table 1 were recorded.

Procedure An SMI Eye-Link head-mounted eye-tracker monitored participants' eye-movements. Images were presented on a 21″ multi-scan color monitor at a resolution of 1024 x 768 pixels together with the spoken sentences. The image appeared 1500 ms prior to utterance onset. Each participant saw only one condition of each item, and the order of appearance of items was randomized individually for every participant. It was further ensured that no participant heard any utterance or part of it more than once. There were eight experiment lists. Each consisted of 24 experiment and 48 filler items. Consecutive experiment trials were separated by at least one filler trial. Before the experiment, participants were instructed to listen to the sentences and to inspect the images. There was no other task. The entire experiment lasted approximately 30 min.

Analysis The critical time region we chose for the analysis extended from the late verb (200 ms prior to adverb onset) to the start of the second noun phrase (labeled 'ADV', see Table 1). For this region, participants had heard most of the verb, but had not heard the disambiguating second noun. The *X-Y* coordinates of participants' fixations were assigned to regions for the entities and scene background. Consecutive fixations within one object region (i.e., before a saccade to another region occurred) were added together, being counted as one *inspection*. For the inferential analysis hierarchical log-linear models were used, which combine characteristics of a standard cross-tabulation chi-square test with those of ANOVA. Log-linear models are adequate for count variables because they neither rely upon parametric

assumptions concerning the dependent variable (e.g., homogeneity of variance), nor require linear independence of factor levels (Howell, 2001). Entities were coded depending on their event role. For Figure 1, for instance, the wizard was coded as 'stereotypical agent', and the detective as 'depicted agent' for the no-competitor conditions (a1 and a2 respectively), and vice versa for the competitor conditions (b1 and b2, see Table 1). The inspections to a character within the ADV time region were a dependent variable in the statistical analysis. Inspection counts for the ADV analysis region were adjusted to factor combinations of *character* (stereotypical agent, depicted agent), *competitor condition* (competitor, no competitor), *information type* (depicted target, stereotypical target) and either participants (N = 24) or items (N = 24).

Results

Figures 2 and 3 show the proportion of inspections to the characters (depicted agent, stereotypical agent) during the ADV time interval. Figure 2 shows inspection percentages to the entities in the two information type conditions (depicted target, stereotypical target) for the No-Competitor condition, Figure 3 for the Competitor condition.

For the No-Competitor condition (a1, a2, see Table 1), when the verb singled out either a depicted (a1) or a stereotypical agent (a2), a significant interaction of information type (depicted target, stereotypical target) and character (depicted agent, stereotypical agent) revealed clear disambiguation using either depicted or stereotypical information (all ps < 0.0001). This was due to a significantly higher percentage of inspections to the depicted agent in the depicted (a1) than in the stereotypical target condition (a2), and a significantly higher percentage of inspections to the stereotypical agent in the stereotypical target condition (a2) than in the depicted (a1) (see Fig. 2).

In contrast, for the Competitor condition (b1, b2, see Table 1), when stored stereotypical knowledge and scene affordances competed and provided conflicting information, we expected no interaction since the stimuli between b1 and b2 were identical (see Table 1). Rather, we found a main effect. We observed more anticipatory looks to the agent of the depicted spying-event (the wizard), than to the stereotypical agent (the detective) for sentences b1 and b2 (see Fig. 3). These looks occurred after people had heard *The pilot* (object/patient) *spies-on*…. and before they heard the respective second noun phrase, which then disambiguated towards the depicted or stereotypical target type. Log-linear analyses showed that the main effect of depicted agent was significant ($p < 0.0001$ by part. and items) in the absence of a significant interaction (ps > 0.6).

Importantly the observed main effect for the Competitor condition (b1 and b2) is only meaningful in comparison to the significant interaction found in the No-Competitor condition (a1 and a2). The difference between the main effect for the Competitor condition (b1 and b2), and the significant interaction for the No-Competitor condition (a1 and a2, see Table 1) was significant. Analyses revealed a three-way interaction between character (depicted agent, stereotypical agent), competitor condition (competitor, no-competitor) and information type (depicted target, stereotypical target) ($p < 0.001$ by part. and items).

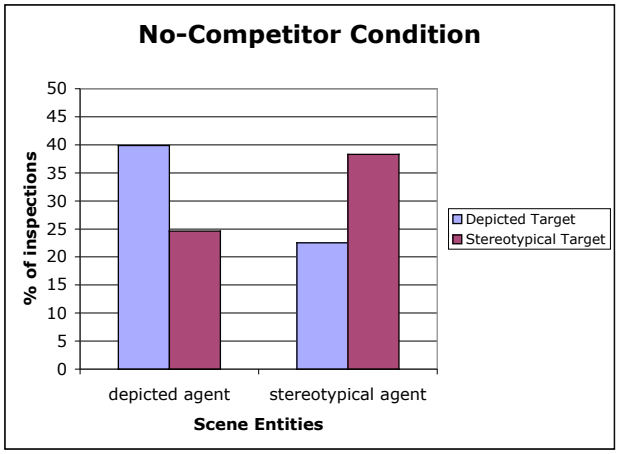


Figure 2: Percentage of inspections to characters during the analysis region ('ADV') in the No-Competitor Condition

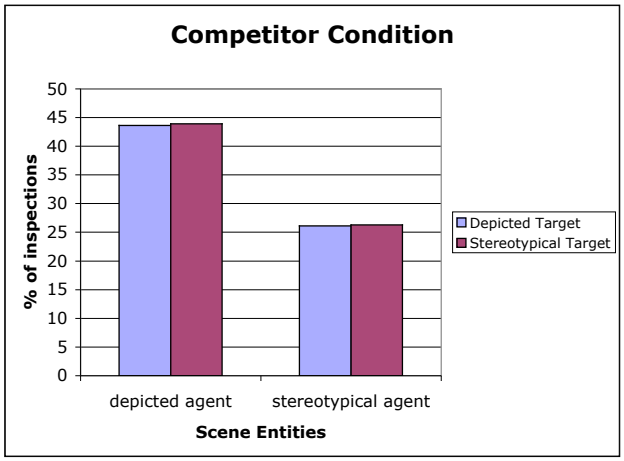


Figure 3: Percentage of inspections to characters during the analysis region ('ADV') in the Competitor condition

Discussion

Within a single study the observed pattern of eye-gazes has shown that both stereotypical knowledge and information that has to be newly acquired from depicted event scenes, allow rapid thematic role interpretation of an unfolding utterance. Our results thus confirm previous findings that thematic-role knowledge (e.g., Ferretti et al., 2001; Kamide et al., 2003) and depicted scene events (e.g., Knoeferle et al., accepted) each are readily available for online comprehension. In the face of competition, however, people have a clear preference for relying on thematic relations acquired from depicted events. We argue that our findings are an important step towards developing a fully specified theory of comprehension that is able to make explicit predictions of auditory sentence comprehension in visual environments. For architectures of the type proposed by Jackendoff (2002), they point to the necessity of incorporating a processing (rather than an architectural)

account for the preferred reliance of the comprehension system on thematic role-relations afforded by depicted event scenes.

Clearly, the observed preferred reliance of listeners on information that they had to newly acquire from depicted scenes, and via a different perceptual system (the visual system), counters our initial expectations of a priority of stored stereotypical knowledge. When stored knowledge and scene compete and provide conflicting information, it is not stereotypical knowledge, which influences our interpretation of the scene. Rather, when verb meaning identified relevant depicted events, the scene guided interpretation of the utterance. Our findings hence indicate an active contribution of thematic role-relations afforded by scene events in online thematic role-assignment. It should nonetheless be highlighted that scene events only influenced thematic role-assignment once they had been identified by the verb.

Such utterance-mediated influence of depicted events suggests a highly efficient interaction between the visual and comprehension systems, where lexical items single out scene entities and events, which then may influence online interpretation. Under this view, we would expect that reference from verbs to the depicted actions in the scene is a pre-requisite for the influence of scene events on online comprehension processes.

Studies by Knoeferle, Crocker, Scheepers & Pickering (accepted) found, however, that depicted scene events allowed online thematic role-assignment and structural disambiguation even before people had encountered a sentence-final verb. The insight that emerges from their finding, and the ones presented in this paper, is that visual scene information has great importance in online comprehension. Indeed, reference from verbs to depicted actions is not an indispensable pre-requisite for the guiding influence of depicted events in auditory comprehension. In sum, this speaks to a *strong* – albeit not necessarily unconstrained - guidance account of the influence of visual scenes on online comprehension processes. Further research is required to establish whether there are constraints on the influence of depicted event scenes on online thematic role-assignment, and under which conditions they apply.

An Adaptive Perspective

In many cases, people find themselves in relevant situations where both spoken language and immediate scene context are available. When watching movies, for instance, people seem to rapidly integrate their knowledge of the world with the movie events unfolding over time in front of their eyes. Clearly, cognitive processes such as understanding an unfolding utterance and an immediate event may occur simultaneously. Yet, when both types of information are available in our environment, the pattern of eye-gazes we observed provides strong evidence for a preference of the immediate environment over expectations of stereotypical events. Further, the rapid impact of the immediate situation identifies the comprehension system to be highly adapted towards acquiring new information from its environment rather than relying on linguistic and world knowledge, a conclusion which bears important implications for both developmental and evolutionary accounts of the language comprehension system.

Acknowledgements

We thank Martin J. Pickering for his comments on an earlier version of this experiment, and Ulrich Pfeiffer for his help in running it. This research was funded by a PhD scholarship to the first author and by SFB 378 project "ALPHA" to the second author, both awarded by the German research foundation (DFG).

References

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. Journal of Memory and Language, 44, 516-547.

Fodor, J. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Gibson, J. (1966). *The Senses considered as perceptual systems*. Boston, Mass.: Houghton-Mifflin Co.

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens.* Berlin: Infix.

Howell, D. C. (2001). *Statistical methods for psychology*. Belmont, CA: Duxbury Press.

Jackendoff, R. (1983). *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research, 32*, 37-54.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering, M.J. (2003). Actions and roles: using depicted events for disambiguation and reinterpretation in German and English. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 681-686), Boston, MA.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering M.J. (accepted). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*.

Levinson, S.C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, Mass.: MIT Press.

McRae, K., Ferretti, T. R. & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12, 137-176.

Steedman, M. (2002). Formalizing affordance. In: *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 834-839), Fairfax VA.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Does Irrelevant Information Play a Role in Judgment?

Boicho Kokinov (bkokinov@nbu.bg)[12]
Penka Hristova (phristova@cogs.nbu.bg)[1]
Georgi Petkov (gpetkov@cogs.nbu.bg)[1]
[1]Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology,
New Bulgarian University, 21 Montevideo Street
Sofia 1635, Bulgaria

[2]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev Street, bl.8
Sofia 1113, Bulgaria

Abstract

This paper presents an unusual prediction made by the DUAL-based model of judgment JUDGEMAP and its verification. The model is shortly presented as well as the simulation data obtained with it. These data predict that people will use the information on an irrelevant dimension when judging another dimension. This prediction is then tested in a psychological experiment and confirmed.

Introduction

Suppose that you are judging how tall a person is. Do you expect that the color of his or her eyes will play a role in that process? Or suppose you are judging the quantity of oil in the bottle you are buying, do you expect that the font used on its label will have an effect? Finally, suppose you are judging the length of a given line segment. Do you expect that the color of the line will make a difference?

Both our intuition and the theories of judgment would answer these questions negatively. Basically they would assume that when judging length we ignore all irrelevant features (including color) and only physical length plays a role. Of course, many other factors, like order of presentation and context, may play a role, but only the length of the lines will take part in the judgment.

This paper is challenging this assumption of standard theories of judgment and is trying to answer the above seemingly stupid and self-evident questions and surprisingly to show that all features (including the irrelevant ones) do matter or more precisely they may matter under certain circumstances.

Approaches to Judgment

There are a number of theories of judgment and a few running models. Most of the theories originate from psychophysics and are mathematical in their nature; they do not describe the process of judgment, but only characterize the end result. Since we are interested in describing the process of judgment we will briefly outline only the main approaches proposed so far in that direction.

Judgment as measuring similarity/dissimilarity with a standard. The classical ideal point approach proposed by Coombs (1964, Wedell & Pettibone, 1999) falls into this category. He believes individuals have their "ideal points" and therefore judging a stimulus can be described as comparing it to this standard and measuring the distance toward it. The Adaptation Level Theory (Helson, 1964) falls into the same category, however, here the standard (adaptation level) is changed depending on context. Finally, the Norm Theory (Kahneman & Miller, 1986) follows a similar approach, however, the standard here is called "norm" and what is more important is that this norm is constructed on the spot rather than retrieved from long-term memory. A comparison set is constructed in working memory consisting of known exemplars and its norm is computed. Thus all three theories can be described as relying on comparison of the target stimulus with a standard (Figure 1), but they differ in the degree to which they subscribe to the constructivist approach toward this standard.

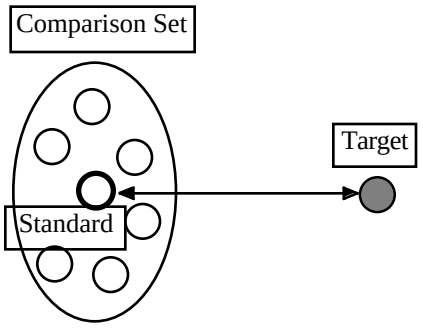


Figure 1. Judgment as comparison with a standard.

Judgment as classification task. Within this approach the comparison set is subdivided into subcategories each of them corresponding to a judgment label (or scale element) and the target stimulus is classified within one of these subcategories. The Range-Frequency Theory (Parduci, 1965, 1974) postulates the constraints which should be met by such category subdivision: the range of value variation within all subcategories should be about the same, and the number of examples in all subcategories should be about the same. The Theory of Criterion Setting (Treisman & Williams, 1984, Treisman, 1985) is a process model that explains how dynamically we change the boundaries of the subcategories. Finally, the ANCHOR model (Petrov &

Anderson, 2000, in press) describes the process of learning of these subcategories and solves the classification task by comparing the target stimulus to the prototypes of each subcategory, these prototypes are supposed to be hold in long-term memory and are called anchors (Figure 2). The comparison set represented by the set of anchors is dynamically formed.

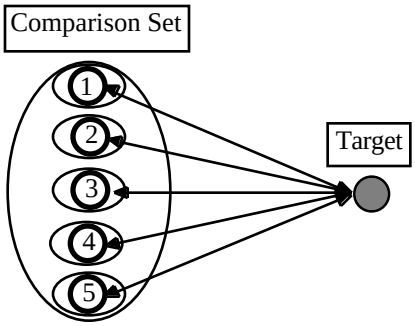


Figure 2. Judgment as classification task. Comparing the target to the standard of each of the subcategories.

Judgment as a mapping task. The DUAL-based model of judgment discussed in this paper follows a third approach: The target stimulus is not compared to the comparison set, but is rather included in it and then a mapping is established between the elements of the comparison set and the set of rating labels (or scale elements). This mapping should be as close as possible to a homomorphism, i.e. the relations among the elements of the comparison sets should be kept among their corresponding rating labels. Thus the process of judgment involves construction of the comparison set, joining the target to it, and mapping between the comparison set and the rating labels (Figure 3).

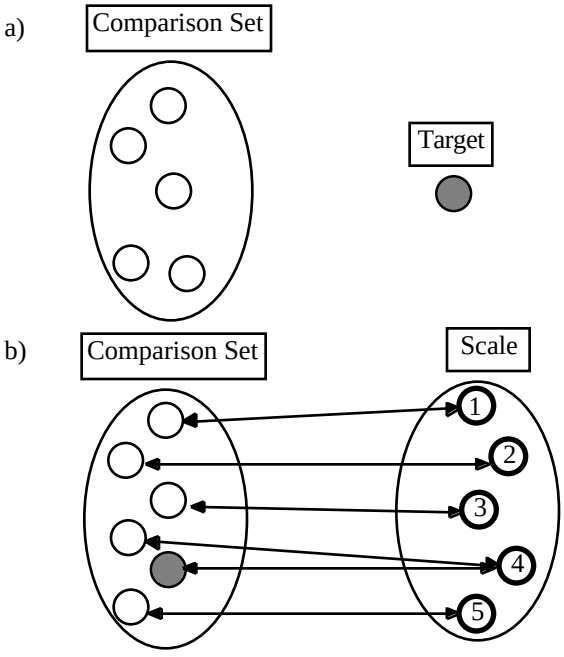


Figure 3. Judgment as mapping in the DUAL-based model.

DUAL-Based Model of Judgment

The current model – JUDGEMAP (Judgment as Mapping) – is based on a general cognitive architecture – DUAL (Kokinov, 1994b, 1994c). This architecture is a hybrid (symbolic/connectionist) one and is explicitly designed to model context-sensitivity of human cognition. It is based on decentralized representations of concepts, objects, and episodes and parallel emergent computations.

The AMBR1 (Kokinov, 1988, 1994a) and AMBR2 (Kokinov, 1998, Kokinov & Petrov, 2001) models are built on DUAL and integrate memory and analogy-making. Since the process of judgment, as described above, involves memory (construction of the comparison set in working memory) and mapping (which is a central mechanism in analogy-making) the JUDGEMAP model is most naturally integrated in DUAL and borrows many of the mechanisms developed for analogy-making in AMBR. Because of the lack of space the model is described only in broad strokes. Interested readers are invited to consult the literature on DUAL and AMBR for more details.

Construction of the comparison set. The comparison set is formed from perception (the target as well as potential context stimuli) and from long-term memory (familiar or recently presented exemplars as well as generalized prototypes, if such exist in LTM). The mechanism responsible for that construction is spreading activation. The sources of activation are the INPUT and GOAL nodes, i.e. the perceived target (and possibly context) stimuli and the goal to judge the stimuli on a scale predefined in the instruction (e.g. a scale from 1 to 7). Thus the representations of the target and the scale elements become sources of activation which is then spread through the network of micro-agents. Naturally, concepts related to the representation of the target become active, e.g. various features of the target – these include both relevant and irrelevant features (of course, relevant features receive more activation than irrelevant ones). The activation spreads further from the general concepts (like RED, GREEN, etc.) towards specific examples of the concepts (other red or green objects). However, there are only a few links from the general concepts to their exemplars – only to the most familiar (typical) exemplars or to recently experienced ones. Thus gradually a number of exemplars (and possibly prototypes) are activated and become part of working memory – all these form the comparison set (Figure 4).

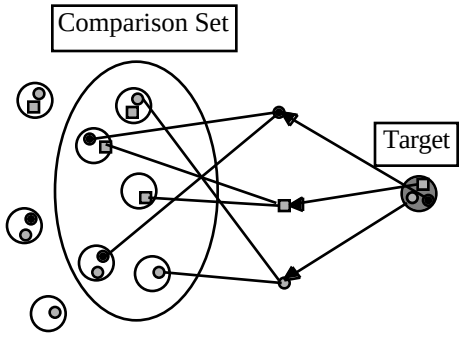


Figure 4. Formation of the comparison set in WM by the spreading activation mechanism of DUAL.

Mapping of the comparison set onto the scale elements. We can now consider the comparison set as a retrieved base and map it onto the scale elements which are the target. The mapping process should preserve the relations among the elements of the comparison set among their images on the scale. The mapping should also follow the range-frequency principle described in the previous section. How is the mapping achieved in JUDGEMAP? Similarly to AMBR, a constraint-satisfaction network is constructed by the marker-passing and structure-correspondence mechanisms. This network consists of temporal agent-hypotheses representing possible correspondences between members of the comparison set and elements of the scale. These initial hypotheses are formed according to the range principle. Excitatory and inhibitory links are constructed among the hypotheses and the spreading activation mechanism selects the winning hypotheses which form the mapping (Figure 5). The competition among the hypotheses implements the frequency principle. As result of this process not only the target stimulus but also each element of the comparison set receives a judgment. This does not mean that people would be aware of all these judgments – most or even all of them might remain unconscious.

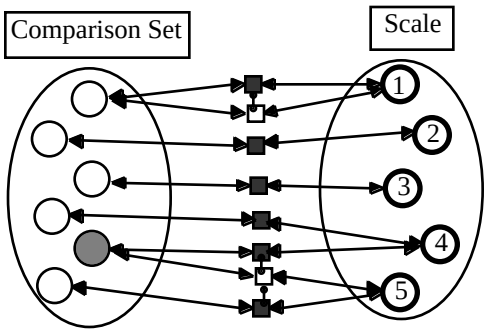


Figure 5. The process of mapping accomplished by the constraint satisfaction mechanism. The winning hypotheses are in black.

Speculative prediction. Since the activation spreads from the target stimulus (represented in a decentralized way by many agents), exemplars, similar in some respect to it (sharing some feature with the target), can be potentially activated and thus become members of the comparison set in working memory. This means that in addition to currently perceived stimuli, to recently activated exemplars, and to highly familiar (typical) exemplars, exemplars which are simply similar to the target will also participate in the comparison set. Moreover, these exemplars might be similar along the relevant (judged) dimension or along an irrelevant dimension.

Let us consider the following example. Suppose we are judging the length of line segments but the lines are colored. Let the target stimulus be a red line of certain length. In this case we may expect that there will be more red lines in the comparison set (Figure 6) – they will be activated through the RED concept which is shared with the target. On the other hand, if the target stimulus is a green line of the same length, more green lines will become part of the comparison set (Figure 7). Now, if it happens that the known red lines are longer than the known green lines, then the two target stimuli (differing only in color) will be included in different comparison sets and thus judged differently and there will be a shift in favor of the green target. Therefore the speculative prediction of JUDGEMAP will be that even such irrelevant feature of the line like its color will play a role in the judgment process. This prediction is in sharp contrast to all theories and models described in the first section, which assume that only the relevant features play a role.

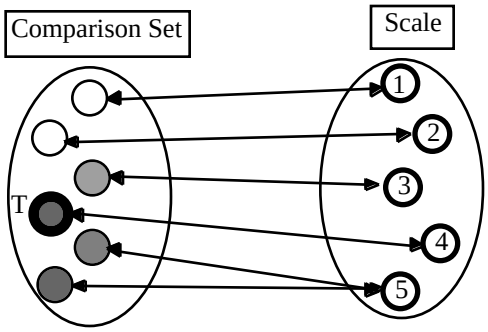


Figure 6. The target stimulus is red and therefore we expect more red exemplars in the comparison set. They happened to be larger in size and thus they compete for the upper part of the scale. In this case the target stimulus (of the same size as in Figure 7) will compete with them and will be mapped onto 4.

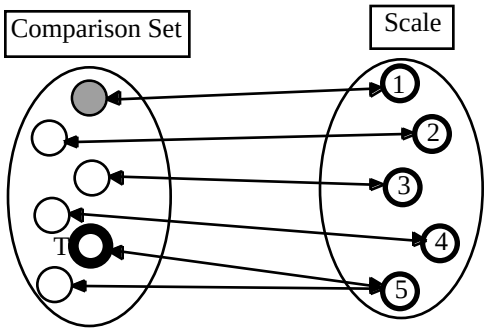


Figure 7. The target stimulus is green and therefore we expect more green exemplars in the comparison set. They happened to be smaller in size and thus they compete for the lower part of the scale. In this case the target stimulus (of the same size as in Figure 6) will compete with them and eventually will be mapped onto 5. In this way we receive an upward shift in the judgment.

Thus we will first describe a simulation experiment with JUDGEMAP that tests in practice this speculation and will also give us a rough estimation of the order of this color effect (if any). If we are successful, we will run a psychological experiment to text the model's prediction and thus verify the model.

Simulation Experiment

In this simulation experiment we use a stimulus set of 56 lines. They are all in the long-term memory of the model. The lines differ in length and color. There are 7 different sizes (from 10 units of length to 34 unit with increment of 4 units) and two different colors (red and green). Thus in each size group there are 8 lines. The frequency of the red (respectively green) lines varies across the size groups. In size group one (the shortest lines - length 10 units) there are 7 green and 1 red line, in the second shortest group (length 14 units) there are 6 green and 2 red lines, etc. In the largest group size (length of 34 units) there are 7 red lines and one green line. Thus we have positively skewed distribution of the green lines and negatively skewed distribution for the red lines.

Each line is represented by a coalition of 5 agents standing for the line itself, for its color, for its length, and for the two relations (color_of and length_of). In addition there are agents standing for the numbers from 0 to 8, but only the agents standing for 1 to 7 are instances of "scale element".

On each run of the program we connect one of these lines to the input list thus simulating the perception of the target stimulus, and connecting the agent standing for "scale_from_1_to_7" to the goal node thus simulating the instruction for rating on a 7 point scale.

We have produced 42 variations of the knowledge base of the system thus simulating 42 different participants in the experiment. The knowledge bases differ mainly in the associative and instance links among the agents, thus although all our "artificial participants" will know the same lines and the same concepts, they will activate different instances in the comparison set.

For each of these knowledge bases we have run two judgment trials: one for a red line of size 22 and one for a green line of the same size.

Simulation Results

The results from the simulations are presented in Figure 8. As we can see the mean rating of the green lines are in most cases slightly higher than the mean rating of the red lines with the same length.

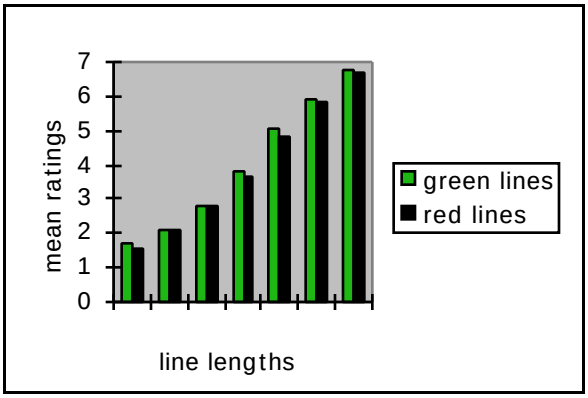


Figure 8. Simulation data. The mean rating of each line with a certain length (1-7) and color (green and red) obtained from all subjects.

Thus the mean of the mean ratings of all red categories is 4.012, while the mean of the mean ratings of all green categories is 4.065, which makes a difference of 0.053 which turns out to be almost significant tested with repeated measurements analysis ($F(1,41)=3.917$, $p=0.055$). The data show that the possible size of the color main effect is very small, but may still be significant. This prediction makes sense: on one hand it is small enough, so that we can ignore it in everyday life and this explains why our intuition says that irrelevant information does not play a role in judgment. On the other hand, the simulation predicts that the irrelevant information does play a role and shifts a bit the evaluation. This means that under specific circumstances this shift might be larger and become significant.

The experiment described below is designed to test this prediction of the model. Basically it replicates the simulation experiment with a larger number of lines.

Psychological Experiment

In this experiment human participants rate the length of red and green lines of various sizes. The interesting question is whether we will obtain a main effect of color, i.e. whether there will be a difference between the ratings of the red and green lines of the same size.

Method

Design

The experiment has a 14x2 within-subject factorial design. The independent variables are length (varying at 14 levels) and color (varying at 2 levels: green and red) of the lines. The dependent variable is the rating of the length of the lines on a 7-point-scale. The experimental question is whether there will be a main effect of color, which is supposedly an irrelevant factor in judging length.

Material

A set of 14 color lines has been presented horizontally against a gray background on a 17-inch monitor. The shortest line is 12 pixels, the longest one is 727 pixels and the increment is 55 pixels. Each particular line length has been shown eight times in red or green color. The short lines were predominantly green while the long ones were predominantly red. The color distribution within the set of all 112 lines (14 lengths x 8 times) is presented in Table 1. The frequency of the stimuli was calculated in order to receive a positively skewed distribution for the green color and a negatively skewed one for the red lines.

Table 1. Frequency of the presented stimulus lines (where 1 represents stimulus length 12 pixels, 2-67 pixels and so on).

lengths	*number of the green lines*	*number of the red lines*
1 & 2	7	1
3 & 4	6	2
5 & 6	5	3
7 & 8	4	4
9 & 10	3	5
11 & 12	2	6
13 & 14	1	7

Procedure

The participants were tested individually in front of a computer screen where all 112 stimuli were shown sequentially and in random order. They were instructed to judge the length of each line presented on the screen on a seven point scale: 1-"it is not long at all", …, 7-"it is very long". No feedback was provided to the participants and no time restrictions have been imposed on them. The whole experiment typically lasted about 15 minutes.

Participants

The participants were 18 undergraduate students (9 men and 9 women none of whom was color-blind) from the introductory classes in psychology at New Bulgarian University who participated in order to satisfy a course requirement.

Results and Discussion

We had 14x2=28 data points for each participant. The results averaged over subjects are shown in Figure 9. Each bar stands for the mean rating that a line of the corresponding size and color has received during the experiment. The repeated measurements analysis showed that the difference (0.046) between the mean judgment of the green lines (4.239) and the mean judgment of the red lines (4.193) is significant ($F(1, 17)=5.966$, $p=0.026$).

Surprisingly enough we obtained a difference (0.046) that is almost the same as the difference we obtained in the simulation (0.053). No tuning of the model was possible since we did not have the experimental data in advance.

Thus the prediction of the JUDGEMAP model has been experimentally confirmed.

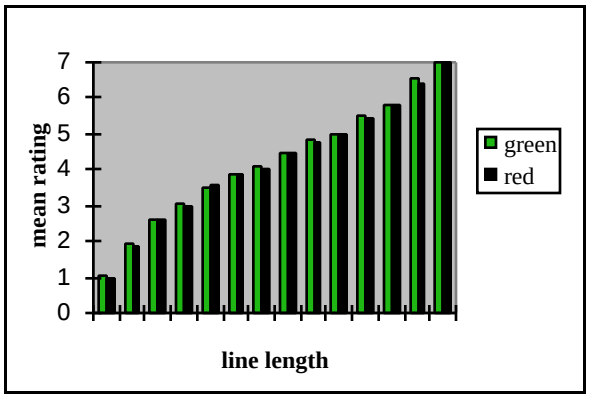


Figure 9. The mean rating of each line with a certain length (1-14) and color (green and red) obtained from all subjects.

Conclusions

The JUDGEMAP model of human judgment has been presented. This model is based on a general cognitive architecture (DUAL) and is thus integrated with the memory and analogy-making model AMBR. Moreover, this model inherits the underlying assumptions of DUAL and AMBR: human cognition is context-sensitive (Kokinov, 1994c), judgment included; human memory is constructive (Kokinov & Hirst, 2003), analogy-making is at the core of human cognition (Gentner, Holyoak & Kokinov, 2001) and its mapping mechanisms may be used in judgment.

The JUDGEMAP model is similar to the Norm theory and the ANCHOR model with respect to the constructive approach to the formation of the comparison set. However, unlike all the models described in the first section judgment in JUDGEMAP is not based on comparison of the target with some aspect of the comparison set, but rather the target stimulus is included in the comparison set and it receives a rating along with all other members of this set. This rating process is based on establishing a mapping between the comparison set and the set of scale elements which mapping preserves the order relations.

Unlike all other models JUDGEMAP does not ignore the irrelevant features of the to be judged targets, moreover these irrelevant features play a role in the construction of the comparison set (retrieving similar objects according to these irrelevant dimensions). The model makes a strange prediction that the color of the target line may play a role in the rating of its length and thus predicts a shift of the mean rating (although a small one) with the change of color. This prediction has been tested in a psychological experiment and has been confirmed.

The size of this color effect is very small, but the stimuli have been very simple and the features unremarkable. It is difficult to imagine that the green color reminds us of a particular green line. That is why we plan to repeat the experiment with more complex stimuli (human figures and clothes) and more memorable features (human faces). It is possible the size of the effect in this case to become larger.

Acknowledgments

We would like to thank the AMBR research team for their continuous support and stimulating environment as well as Stefan Mateeff and Maurice Grinberg for the valuable discussions.

References

Coombs, C. (1964). A Theory of Data. NY: Wiley.

Gentner, D., Holyoak, K. & Kokinov, B. (2001). The Analogical Mind. Cambridge, MA: MIT Press.

Helson, H. (1964) Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior. NY: Harper and Row.

Kahneman, D., Miller, D. (1986). Norm Theory: Comparing Reality to Its Alternatives. *Psychological Review*, vol. 93 (2), pp 136-153.

Kokinov, B. (1988). Associative memory-based reasoning: How to represent and retrieve cases. In T. O'Shea and V. Sgurev (Eds.), *Artificial intelligence III: Methodology, systems, applications*. Amsterdam: Elsevier.

Kokinov, B. (1994a). A hybrid model of reasoning by analogy. In K. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory: Vol. 2. Analogical connections* (pp. 247-318). Norwood, NJ: Ablex

Kokinov, B. (1994b). The DUAL cognitive architecture: A hybrid multi-agent approach. *Proceedings of the Eleventh European Conference of Artificial Intelligence*. London: John Wiley & Sons, Ltd.

Kokinov, B. (1994c). The context-sensitive cognitive architecture DUAL. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kokinov, B. (1998). Analogy is like cognition: Dynamic, emergent, and context-sensitive. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research.* Sofia, Bulgaria: NBU Press.

Kokinov, B. & Hirst, W. (2003). Constructive Memory. Sofia: NBU Press.

Kokinov, B. & Petrov, A. (2001). Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In: Gentner, D., Holyoak, K., Kokinov, B. (eds.) *The Analogical Mind: Perspectives from Cognitive Science*, Cambridge, MA: MIT Press.

Luce, D. (1959). On the Possible Psychophysical Laws. *Psychological Review*, 66(2), 81-95.

Parducci, A. (1965). Category Judgment: A Range-Frequency Model. *Psychological Review*, vol. 72 (6), pp 407-418.

Parducci, A. (1974) Contextual Effects: A Range-Frequency Analysis. In: Carterette, E., Friedman, M. (eds) *Handbook of Perception.* vol. II. Psychophysical Judgment and Measurement. NY:Academic Press, pp 127-141.

Petrov, A. & Anderson, J. (2000). ANCHOR: A Memory-Based Model of Category Rating. In: L. Gleitman & A. Joshi (eds.) Proceedings of the 22nd Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Petrov, A. & Anderson, J. (submitted). The Dynamics of Scaling: A Memory-Based Anchor Model of Category Rating and Absolute Identification.

Sarris, V., Parducci, A. (1978) Multiple Anchoring of Category Rating Scales. *Perception and Psychophysics*, vol. 27 (1), pp 35-39.

Thurstone, L. (1927). A Law of Comparative Judgments. *Psychological Review*, 34, 273-286.

Treisman, M. (1985). The Magical Number Seven and Some Other Features of Category Scaling: Properties of a Model for Absolute Judgment. Journal of Mathematical Psychology, 29, 175-230.

Treisman, M. & Williams, T. (1984). A Theory of Criterion Setting with an Application to Sequential Dependences. Psychological Review, 91(1), 68-111.

Wedell, D., Pettibone, J. (1999) *Preference and the Contextual Basis of Ideals in Judgment and Choice.* Journal of Experimental Psychology: General, vol. 128, pp 346-361.

An activation-based model of agreement errors in production and comprehension

Lars Konieczny (lars@cognition.uni-freiburg.de)
Center for Cognitive Science, IIG, Univ. Freiburg
Friedrichstr. 50, 79098 Freiburg

Sarah Schimke (sarah@cognition.uni-freiburg.de)
Center for Cognitive Science, IIG, Univ. Freiburg
Friedrichstr. 50, 79098 Freiburg

Barbara Hemforth (barbara.hemforth@lpl.univ-aix.fr)
Laboratoire Parole et Langage, UMR 6057, Univ. de Aix en Provence
29, av. Robert Schuman, 13621 Aix en Provence, France

Abstract

With this paper, we introduce the agreement production error (APE) model. APE is a model of comprehension and production performance that applies a theory of memory and cognition (ACT-R 5.0) to the task of linguistic processing embedded in a variety of psycholinguistic experimental paradigms.
With its roots in the ACT-R theory, agreement errors are modeled as a combination of symbolic processing and chunk activation dynamics. Whether a plural or a singular verb is produced depends on the accessibility of the Subject's plural marking. The activation of plural-marking chunks decays, so that it might not be found when its retrieval is attempted at the verb, resulting in a general singular error (Hemforth and Konieczny, 2003). This effect is then modulated by task and construction specific variations.

Introduction

When people speak or write they occasionally produce verbs not agreeing in number with the subject. This happens particularly often when the singular subject is followed by a plural modifier in constructions like (1; quoted form Bock & Miller, 1991).

(1) The readiness of our conventional forces are at an all-time low.

The mechanism underlying this error is attributed to the marked plural feature percolating up the tree too far (Vigliocco & Nicol, 1998). This account is substantiated by the fact that no comparable singular/plural mismatch effect for constructions with marked plural heads has been established so far.
Very recently, Haskell and MacDonald (2002) proposed the principle of proximity as an alternative explanation. They showed that in disjunctions like (2), subjects have a strong preference to match the number marking on the verb with the more local noun. In addition to distributional evidence, this was taken to indicate that the classical attraction error at least partially and at least in English is caused by number marking on a close interfering noun.

(2) a. The hat or the gloves is/are red.
b. Is/are the hat or the gloves red?

In a series of five written production experiments, Hemforth and Konieczny (2003) tested the proposed mechanisms in German.
In this paper, we follow up on this work and present a new model, APE, that accounts for the written production data in the experiments reported. The paper is organized as follows. We will start by summarizing the two most important experiments from Hemforth and Konieczny (2003). After that, the model will be outlined. Since APE is based on ACT-R 5.0, a short introduction to this theory is provided beforehand. The paper ends with a general discussion and conclusion.

Error patterns in written production

The first experiments replicated the classical results on subject-modifier-verb constructions. Two factors were varied in the first experiment: The factor "Match": matching (1,4) or mismatching (2,3) number marking on head noun and local noun, and the factors "Number of the head noun": singular (1,2) or plural (3,4) head noun.
(1)Die Farbe auf der Leinwand _____ trocken.
The color on the canvas _____ dry.
(2) Die Farbe auf den Leinwänden _____ trocken.
The color on the canvasses _____ dry.
(3) Die Farben auf der Leinwand _____ trocken.
The colors on the canvas _____ dry.
(4)Die Farben auf den Leinwänden _____ trocken.
The colors on the canvasses _____ dry..

Hemforth and Konieczny (2003) found a clear effect of the number of the head noun on the percentage of agreement errors. Neither the factor "Match" nor the

number x match interaction reached significance. However, whereas no difference in matching versus mismatching local nouns could be established for sentences with plural marked head nouns, planned comparisons showed an effect with singular marked head nouns. This result replicates the well-known modifier attraction effect (e.g. Vigliocco & Nicol, 1998, Bock & Miller, 1991).

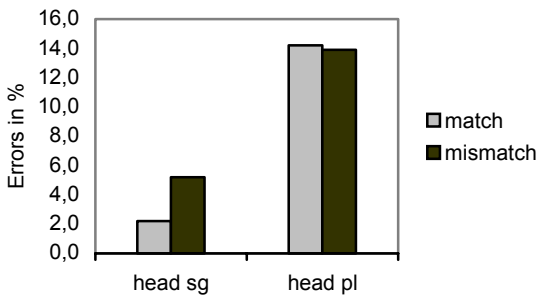


Figure 1: Agreement errors for NP-PP-V constructions

In line with earlier experiments on written production (e.g., Branigan et al., 1995; Hölscher & Hemforth, 2000), we found a considerably high number of agreement errors for plural marked head nouns, reflecting a general tendency to produce singular verbs. Nevertheless, the result for singular heads is compatible with both the feature percolation hypothesis and proximity. Hemforth and Konieczny (2003) therefore ran a series of experiments on Subject-Object-Verb (SOV) constructions. An object attraction effect for singular subjects would rule out feature percolation, because the object is not embedded within the subject.

Object attraction?

The experimental factors varied in three further experiments were "Match": matching (9,12) or mismatching (10,11) number marking on Subject NP and local object NP, and "Number of Subject": singular (9,10) or plural (11,12) Subject NP.

(9) Ich habe gehört, dass der Mann die Frau besucht _____.
I have heard that the man(masc,nom) the woman visited _____.

(10) Ich habe gehört, dass der Mann die Frauen besucht _____.
I have heard that the man(masc,nom) the women visited _____.

(11) Ich habe gehört, dass die Frauen den Mann besucht _____.
I have heard that the women the man (masc, acc) visited _____.

(12) Ich habe gehört, dass die Frauen die Männer besucht _____.
I have heard that the men the women visited _____.

In all three experiments, the number marking on the Subject had a strong effect on the number of agreement errors: more errors were produced following a plural Subject. However, less errors were produced when the local Object-NP was also plural marked.
The lack of an object attraction effect for singular subjects is consistent with the feature percolation hypothesis and contradicts proximity.

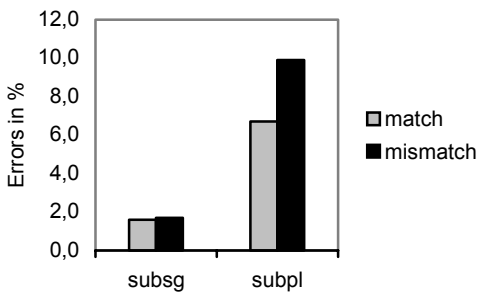


Figure 2: Agreement errors in SOV-constructions

The mismatch effect for plural subjects, however, is not predicted by feature percolation. Hemforth and Konieczny (2003) proposed *feature reactivation* as an explanation. According to that hypothesis, the plural feature of the subject (head) is subject to activation decay so that activation can be below the retrieval threshold when the verb must be produced. This mechanism would account for the general tendency to produce singular verbs. In SOV constructions, however, the subject plural feature can be reactivated by an object plural feature, because both subject and object are arguments of the verb.

APE: A hybrid model of agreement errors

The data so far suggest that both syntactic constraints and proximity affect agreement errors. For one, there are certain effects restricted to certain syntactic constructions ("feature percolation"), and second, there are effects of locality and interference best dubbed in terms of decay and reactivation. The Agreement Production Error (APE) model is built atop the ACT-R architecture, which provides us with mechanisms for *i.* declarative chunk activation and decay, embedded in a *ii.* symbolic processing architecture with *iii.* cost-dependent rule selection, and *iv.* task-specific modelling.

An informal introduction to ACT-R 5

ACT-R (Anderson and Lebiere, 1998; Anderson et al., submitted) is both a theory of cognition and a modelling framework, where scientists can build their specific models using well defined and empirically justified concepts that serve as the model's building blocks.

ACT-R distinguishes declarative from procedural knowledge. Both employ advanced sub-symbolic mechanisms.

Chunks are the elements of *declarative memory*. They bundle information in a collection of attribute value pairs, by which chunks get linked to other chunks to form networks of declarative knowledge. When chunks are created, they start out with a certain base level activation that decays over time, following the power law of forgetting. If their activation falls below a certain threshold, their chance for being retrieved by a production approaches zero.

When a chunk is retrieved from declarative memory, its base level activation is permanently pushed a bit upwards. The more often it is used, the higher it will be activated (power law of learning). In addition to base level activation, a chunk can temporarily receive additional activation spreading from associated chunks stored in the goal of the current production rule. The rationale is that chunks relevant for the problem stated in the current goal should benefit from being within the current focus of attention.

Production rules interact with declarative memory via retrieval of chunks. A retrieval request may succeed or fail, depending on whether or not a matching chunk exists, and on that chunk's activation. Among several matching chunks, the one that is activated highest has the highest probability of being retrieved. A retrieved chunk is stored in the retrieval buffer, where the next production can use it. Given a certain constellation of chunks in the available modality specific buffers (goal, retrieval, visual, audio, etc.), multiple production rules might apply to this state, but only one will be picked for firing. In ACT-R, conflict resolution in the case of multiple potential production instantiations is determined by the *utility* value of each rule in the conflict set. The *utility* is in turn a function of the past history of success and the cost associated with solving the problem by picking this rule.

If there are two productions in the conflict set, one of which has a high probability of being successful but which takes a while for getting there, and the other is quick but has only a mediocre success history, the choice will depend on the amount of time the user devotes to the problem. If the focus is on accuracy, plenty of time can be spent solving the problem so that the more successful rule will be picked. If the focus is on speed, the faster rule will be picked even though it's not unlikely that it will fail.

Basic modelling decisions

While ACT-R provides only limited means of representing and processing declarative and procedural knowledge, there are still many alternative ways in which knowledge can be modelled. We settled on the following modelling decisions before we actually started:

- Restricted use of direct storage. We do not store linguistic elements in slots of goals unless it is required for declarative reasons. Newly created chunks are released into declarative memory and retrieved when needed. Binding a chunk to a slot in the goal for procedural reasons is too strong a computational assumption, since stored chunks are excluded from activation decay.
- As a consequence of this, syntactic nodes have to be retrieved from memory in order to become integrated with other nodes. For instance, when the verb is processed, all its complements and adjuncts must be retrieved from memory to form an integrated interpretation.
- The cost of integrating a word is hence a function of the accessibility of its dependents in memory. Integration will be the easier, the more locally its dependents have been processed beforehand (cf. Gibson, 1998)
- The restricted use of direct storage has also consequences for the agreement mechanism, because NPs have to be retrieved all the time to get attached to modifiers or for incremental interpretation. The continuous retrieval of NPs influences their activation, so that some will be more accessible than others when the plural feature is to be assigned.

The sentence processing mechanism

The current model version performs the completion task as used in Hemforth and Konieczny (2003). First, a series of noun phrases are processed before the model produces a verb.

Processing starts with a goal that represents the current processing state during the assembly of the sentence elements. Each word read from the screen triggers the retrieval of a lexical element and is integrated into the currently processed phrase marker. The first NP is marked as the subject of the clause. When a modifier is processed, its host will be retrieved for attachment.

At each top level element of the sentence, a propositional interpretation is sought that matches the concepts associated with the processed phrases to a long term proposition (cf. Budiu & Anderson, 2000). If such an interpretation can be found, the concepts are hooked to that proposition.

In verb-final constructions, multiple arguments might precede the verb. In the absence of thematic

information of the verb, APE anticipates the verb by retrieving a proposition in the background knowledge that integrates the arguments processed so far. The more arguments have been read, the more likely will the interpretation match the right one when the verb is read, so that actually integrating the verb will become easier (Konieczny and Döring, 2003). In this view, anticipation is at the conceptual rather than the syntactic level.

Modelling agreement

Representation of number. Singular is the default number of nouns, whereas plural has to be assigned explicitly. A noun phrase is hence considered singular unless it is marked plural.

The plural feature is modelled as a chunk that links the plural attribute to a noun phrase, rather than as a slot in an NP chunk. As a chunk, a plural feature is subject to ACT-R's activation dynamics.

Producing number. Verb production is modelled by distinct production rules for singular and plural forms. The plural rule attempts to retrieve the plural feature of a noun phrase marked as subject of the sentence. If it succeeds, the plural form is generated from the base form. If it fails, the singular rule produces the singular base form.

The plural rule has higher utility, due to its better accuracy in actually producing the right verb when the subject is plural. The singular rule is less specific and therefore error prone, but less cost intensive.

Where syntax matters. The German production data suggest that plural attraction is construction specific: While there is a robust modifier attraction effect for singular subjects in SMV constructions, there is none in SOV. This result has been predicted by the *feature percolation hypothesis*.

Feature percolation is an encoding error by nature, in which the plural feature is erroneously assigned to the head noun instead of the embedded noun.

ACT-R, however, lacks a direct mechanism for encoding errors (i.e. the creation and release of false chunks). The only place where errors can occur in ACT-R is during retrieval of chunks.

Modelling "Feature percolation". In APE, the construction specific encoding error is modelled as a retrieval error during the search for an NP that the plural feature is to be assigned to. That is, during plural assignment, the newly created plural feature requires a root NP that has to be retrieved from memory. At the modifier in a complex-NP construction, the head NP at that point is highly activated due to the fact that is had been retrieved for modifier attachment shortly before plural assignment. Since both the head NP and the modifier NP are about equally strongly activated, there is a certain chance that the head NP, not the modifier NP is retrieved for plural marking. If that happens, the subject has inherited the plural marking from the embedded NP.

In SOV constructions, no attachment takes place between the Subject and the Object-NP. Therefore the Subject NP is not going to be retrieved before the plural assignment of the Object. Retrieving the root NP for the object plural feature is hence less error prone.

On the other hand, the subject NP will be retrieved *after* the object has been assigned to the new plural feature, because subject and object are both needed for incremental interpretation, i.e. the anticipation of a matching relationship between both. The interpretation depends upon the number of the entities to be integrated, as the examples () and () illustrate.

(8) Gestern haben die Professoren den Studenten … Yesterday have the professors$_{nom}$ the student$_{acc}$ "*Yesterday, the professors have … the student*."

(9) Gestern hat der Professor die Studenten … Yesterday has the professor$_{nom}$ the students$_{acc}$ … "*Yesterday, the professor has … the students."*

Things that multiple professors are likely to do to a single student (examined, rejected, etc.) can be different from things that a single professor is likely to do to multiple students (taught, etc.). Number is hence an important feature at the conceptual level and useful for interpretation anticipation.

Differential plural re-activation in SOV constructions. During the process of incremental interpretation of each new verb-complement, each previous concept participating in the proposition, and, importantly, its plural feature - if it exists - will be retrieved and hence its activation will be pushed a bit. That is, when the object in SOV constructions is processed, the plural marking of the subject will be re-activated. This will only happen, however, if it can be successfully retrieved when the object is interpreted. There is a slight chance that the plural feature cannot be found here because its activation has already decayed too strongly. This chance, however, will be lower, if the plural feature receives additional activation from the goal, which is the case if the object is marked for plural. The additional amount of source level activation results in a higher retrieval probability for the subject plural marking if the object is plural. Therefore, the subject plural will be reactivated more often, so that it can be retrieved better and more often when the verb is produced.

Implications

Whether a plural or a singular verb is produced depends on the accessibility of the Subject's plural marking. The activation of plural-marking chunks decays, so that it might not be found when its retrieval is attempted at the verb, resulting in a general singular error (Hemforth and Konieczny, 2003). This effect is then modulated by task and construction specific variations. The model will

come in variants for different experimental paradigms, which are nevertheless based on the same core for verb-production. The first model variant presented here performs the completion task for written production and is hence a combined sentence processing and production model. It first reads two NPs, embedded in a variety of constructions, and then produces a verb. Modifier attraction errors (cf. Bock & Miller, 1991; Vigliocco & Nicol, 1998) are due to encoding errors during plural marking (feature percolation). The plural feature of the modifier-NP sometimes gets wrongly assigned to the Subject-NP. This effect is due to the necessity of reactivating the Subject-NP in processing the modifier-NP and is therefore restricted to modifier-NPs (Hemforth & Konieczny, 2003). In SOV constructions on the other hand, at each new NP, all verb arguments get reactivated to allow incremental interpretation and verb-anticipation (Konieczny and Döring, 2003). During this process the plural feature of an Object-NP can reactivate a plural feature of the Subject, reducing the singular error in S-O-V constructions (Hemforth & Konieczny, 2003).

The model is currently extended to other types of tasks to be able to account for task-specific-differences.

Aural repetition and completion. In this task, the participants are first presented with a preamble and then have to repeat and complete it with a verb in order to form a full sentence. This type of task has been used in the majority of experiments on agreement errors (e.g. Bock and Miller, 1991, Franck, Lassi, Frauenfelder and Rizzi, 2003, Bock and Cutting, 1992, Hartsuiker, Antón-Méndez and van Zee, 2001, Vigliocco and Nicol, 1998) and is generally claimed to test "pure" production, because subjects actually utter the whole sentence and not just the verb as in a completion task. Nevertheless, there is an inevitable comprehension component in this task as well, as the preamble has to be presented to the participants in some form. Instead of accounting for only the production part, we model the entire task, including reading and memorizing the preamble, repeating and completing it. Processing the preamble entails forming declarative representations for phonological and syntactic/conceptual information as in the completion task. To repeat the preamble, the model will have to retrieve either its phonological or syntactic representation. After repetition of the preamble, the verb should than be produced in the same way as in the present model. The difference between this task and the written completion task is in the higher activation of all elements of the preamble, as it is not only read once, but than (partly) retrieved and repeated. This should have no influence on the encoding error, but should make the general singular error much less frequent, which is in fact what has been observed in the experiments above. Moreover, the model predicts a floor effect for length variations of the intervening material in this paradigm, which is in fact was has been found (e.g. Bock and Miller, 1991, Bock and Cutting, 1992).

Time pressure. In the experiments reported by Hartsuiker, Antón-Méndez and van Zee, 2001, participants had to perform the production task under time pressure. The model predicts the observed effects: Under time pressure, productions that consume too much time are more likely to be ignored in favor of faster but less accurate productions. In particular, the plural-feature will sometimes be left underspecified for its root. Under-specified plural-features will then erroneously be retrieved at the moment of the production of the verb. This would happen even with plural-features of intervening object noun phrases, hence accounting for the object-attraction-effect obtained by Hartsuiker et al. Moreover, time pressure should increase the general singular error due to a change in the utilities of the productions which determine the number of the verb: the more accurate, but also more time consuming plural rule will have a lower utility than in other paradigms. This will lead to the singular rule firing more often, which produces the singular base form without checking for the plural feature.

Dual task paradigms. Finally, the model can also be extended to dual-task-paradigms, as applied by Fayol, Largy and Lemaire, 1994, and Hemforth, Konieczny and Schimke 2003, in which participants have to perform a second working-memory consuming task in addition to the sentence completion or production task. The cognitive load created by this second task will lead to less attention, i.e. source activation, being devoted to relevant chunks during the processing of the preamble. This will make both the encoding error and the singular error more likely, because they are both due to retrieval errors that become more likely if decay is stronger or starts from a lower level. As predicted, a higher overall number of errors was found in the experiments cited above.

The influence of a specific experimental paradigm may further interact with properties of the language which is investigated. Such an interaction may eventually explain a cross-linguistic difference which has been observed in SOV-constructions: In contrast to the German results reported above, several studies conducted on French SOV-construction found a mismatch effect in the SP-condition ("object attraction", see for example Fayol , Largy, Lemaire, 1994; Franck, Lassi, Frauenfelder, und Rizzi, 2003). In French, if there is an object in preverbal position, it always has to be a clitic pronoun. As these pronouns are very short, they have to be processed in less time than a full NP. Under such conditions, the model would predict the same effect as under time pressure: the plural-feature may be left underspecified for its root and

might then be retrieved when the verb has to be produced, leading to a plural-error.

Conclusion

We have introduced APE, an activation-based model of agreement errors in production. The model emphasizes the activation dynamics of the plural feature as the major source of variability in the data. Syntax effects are accounted for via operations that some constructions require that others do not. For instance, while modifiers have to be attached to their hosts, objects are not attached to subjects. On the other hand, objects, like other arguments, participate in anticipating the verb, while modifiers to NPs do not (to same extent, at least). These construction-specific operations interact with the activation dynamics of the plural feature in systematic ways that have been demonstrated to cover a wide variety of agreement phenomena discussed in the literature. The model predicts that time and feature re-use are crucial variables in production. Unlike purely linguistic production models of agreement errors, it can account for differences in task demands and non-linguistic factors of agreement performance.

Acknowledgments

We would like to thank Rob Hartsuiker, Michel Fayol, Julie Franck, Gerhard Strube and two anonymous reviewers for their helpful comments on earlier versions of this work.

References

Anderson, J. R., Lebiere, C. (1998). The atomic components of thought. Lawrence Erlbaum Associates

Anderson, J. R., Bothell, D., Byrne, M. D., & Lebiere, C. (submitted). An integrated theory of the mind. *Psychological Review*.

Bock, K., & Miller, C.A. (1991). Broken agreement. *Cognitive Psychology, 23,* 45-93.

Bock, K. und Cutting, J.C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31,* 99-127.

Branigan, H. (1995). *Language processing and the mental representation of syntactic structure.* Unpublished doctoral dissertation, Edinburgh University.

Budiu, R. & Anderson, J. R. (2000). Integration of background knowledge in sentence processing: A unified theory of metaphor understanding, semantic illusions, and text memory. In *Proceedings of the Third International Conference on Cognitive Modeling*, pp. 50-57. Groningen, Netherlands: Universal Press.

Fayol, M., Largy, P. und Lemaire, P. (1994). Cognitive Overload and Orthographic Errors : When Cognitive Overload Enhances Subject-Verb-Agreement errors. A Study in French Written Language. In: *The quarterly Journal of Experimental Psychology, 47A (2)*, 437-464.

Franck, J., Lassi, G., Frauenfelder, U. und Rizzi, L. (submitted). Agreement and movement: a syntactic analysis of attraction.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1-76.

Hartsuiker, R., Antón-Mendéz, I., & van Zee, M. (2001). Object-attraction in subject-verb agreement construction. *Journal of Memory and Language, 45*, 546-572.

Haskell, T., & MacDonald, M. (2002). Proximity does matter. *Paper presented at the 15th annual CUNY conference on human sentence processing*, New York, March 2002.

Hemforth, B., & Konieczny, L. (2003). Proximity in agreement errors. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, August 2003, Boston, MA.

Hemforth, Konieczny, & Schimke (2003). Modifier attraction and object attraction. *Poster presented at the conference on architectures and Mechanisms of Language Processing* (AMLaP), August 2003, Glasgow.

Hölscher, C., & Hemforth, B. (2000). Subject-verb agreement in German. In B. Hemforth & L. Konieczny, *German sentence processing* (279-310). Dordrecht: Kluwer Academic publishers.

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads. Evidence from eye-tracking and SRNs. In: P. P. Slezak (ed.): *Proceedings of the 4th International Conference on Cognitive Science*, July 13-17, 2003, University of New South Wales, Sydney, Australia. 330-335

Schriefers, H., & van Kempen (1993). Syntaktische Prozesse bei der Sprachproduktion: Zur Numerus-Kongruenz zwischen Subjekt und Verb. *Sprache und Kognition, 12*, 205-216.

Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition,68,*B13-B29.

Creativity Over the Lifespan in Classical Composers: Reexamining the Equal-Odds Rule

Aaron Kozbelt (AaronK@brooklyn.cuny.edu)
Department of Psychology, Brooklyn College, CUNY
2900 Bedford Ave.
Brooklyn, NY 11210-2889 USA

Abstract

This investigation extends Simonton's (1977) seminal study of lifespan creativity to eighteen eminent classical composers' total output. Major and minor work production was positively correlated over the lifespan, consistent with Simonton's chance-configuration theory of creativity. However, contrary to Simonton's (1977) null finding, here strong, positive, linear and weaker, negative, quadratic age trends were consistently found in predicting hit ratio (proportion of masterpiece-level music to total music composed per age period). This uniformly replicated in several analyses examining alternative explanations for this pattern. Most individual composers' hit ratios increased with age; none declined. The results indicate that composers' perspicacity in evaluating ideas mostly increases with age, suggesting greater importance for evaluation and elaborative problem solving processes in creative productivity than implied by a chance-configuration account.

Introduction

How does creativity change over the lifespan? Do eminent creators learn to be more creative and to reliably intuit which ideas are worth elaborating? Or is creative productivity primarily driven by chance and by individual differences that are largely impervious to learning or to external influences? Dean Keith Simonton's (1977) seminal study of ten eminent classical composers was one of the first to address these questions in a comprehensive, methodologically sophisticated way. His results strongly argue for the second alternative: illustrious composers apparently do not learn to write a larger proportion of great music as their careers progress, and productivity is remarkably immune to external perturbations like wars or civil unrest. Simonton's (1977) analysis has become a cornerstone of his "chance-configuration" theory of lifespan creative productivity (Simonton, 1984a, 1988, 1997, 1999), the most comprehensive, elaborated, and parsimonious psychological theory of this complex phenomenon.

However, is it really the case that creativity, or at least perspicacity, does not improve, even in distinguished composers? Classical music aficionados may think of Beethoven, Haydn, or Verdi, and be hard pressed to think of any late, major works by these composers that are not critically acclaimed, even when compared to their earlier efforts. In contrast, thinking of comparably lauded early works by these composers is probably much harder (Hayes, 1989; Weisberg, 1999). Nevertheless, Simonton's (1977) quantitative analysis implies that these instances are illusory.

Understanding lifespan learning and creativity clearly has important theoretical and practical implications, and multiple approaches may prove fruitful. For instance, alternative definitions of "great" music might be used to examine how well Simonton's (1977) results generalize. Also, his nomothetic analysis did not investigate individual differences in career trajectory. Varied trends can cancel each other, accounting for the null finding. Of particular interest are two types of trajectories: 1) a completely flat agewise function, consistent with the "equal-odds rule" and a chance-configuration account, and 2) an agewise increase in hit ratio, consistent with an expertise acquisition or problem solving framework. (Of course, trajectories can also decline or exhibit curvilinear functions.)

Simonton's own Darwinian chance-configuration model is the most thoroughly researched theory of lifespan creativity (see Simonton, 1988, 1997). Its basis is the blind generation and selective retention of ideas (Campbell, 1960). Creative ideation follows a constant probability of success, or the "equal-odds" rule. Simonton (1999, pp. 188-197) has further argued that this ratio cannot be increased by any means; a creator's hit ratio stays constant with age.

This theory has profound psychological implications. First, it suggests there is little learning or improvement in perspicacity. Creators lack the ability to judge their ideas or works reliably, either as final products or works-in-progress. At a social level, it suggests that creators have little control over the ultimate fate of their products. Thus, mass-production is the best strategy for those who seek eminence. The chance-configuration theory counterintuitively predicts that the high point of a creator's career will simultaneously result in the most masterpieces and the most ephemera: writing one masterwork does not guarantee further success.

Numerous studies support this theory. High correlations (typically r = .50 to .75) are often found between the production of major and minor works over the lifespan (see Simonton, 1997). Comprehensive studies of the output of ten great composers (Simonton, 1977) and ten eminent psychologists (Simonton, 1985) also found no agewise hit ratio changes. However, Kozbelt (in press) found a very strong agewise increase (r (age, annual hit ratio) = .91, $p <$ .0001) in Mozart, a composer in Simonton's (1977) sample.

Thus, at least some composers' hit ratios may improve with age. This is consistent with an expertise acquisition or problem solving perspective (Newell, Shaw, & Simon, 1962; Weisberg, 1999). From this view, creating a musical composition is an open-ended problem, on which composers bring their musical skill to bear. Expertise acquisition studies (Ericsson, Krampe, & Tesch-Römer, 1993) suggest

individual differences in musical ability are largely attributable to deliberate practice rather than innate talent. Likewise, Hayes (1989) examined 76 composers and found that in almost all cases, at least ten years of intensive study were required before a composer wrote his first masterwork, defined by recording counts. This result seems inconsistent with Simonton's (1977) finding of no change in hit ratio over composers' entire careers, since clearly great composers are writing *some* great works in their maturity. Thus, despite potentially important structural differences between expert performance and creative performance (Simonton, 2000), agewise improvement may be a possibility, particularly if skilled performance is a central component of creativity.

As noted earlier, individual variation may overwhelm a seemingly flat aggregate agewise trajectory. Galenson (2001) has found a systematic dichotomy among painters, linking career trajectories (based on the average value of work produced at various times) with creative processes. "Finders" peak early and largely plan their work in advance, proceeding with little revision. "Seekers" peak at reliably later ages and engage in much revision as they work. A comparable dichotomy among composers was suggested by Simonton (1986). He found that the most aesthetically successful works appeared early or, better, very late in composers' careers.

The present study tests the robustness of the equal-odds rule and also examines individual composers. It adopts Simonton's (1977) basic methodology of cross-sectional time-series. However, the present investigation also seeks to vary or improve the measurement of some constructs (e.g., how "masterpieces" are defined or how works can be more sensitively weighted in the analysis). Further, it seems important to confirm that the null finding on hit ratio replicates on a larger sample of composers, since, as noted above, few studies directly and comprehensively examine this important question. Moreover, a career-long flat hit ratio seems inconsistent with the finding that composers generally write no great works in the first decade or so of their careers (Hayes, 1989; Weisberg, 1999). Unfortunately, Simonton's (1977) original data are no longer extant, due to computer technology changes (Simonton, personal communication). While catalogs of composers' works are fairly standardized, many works of questionable authenticity or dating hover at the periphery of such catalogs, so revisiting and extending this type of analysis seems worthwhile.

The four models described above permit a characterization of aggregate and individual career hit ratio trajectories. The chance-configuration theory predicts no change in hit ratio with age, a positive correlation between major-minor work production over the lifespan, and is fairly agnostic on work quality. The problem solving perspective predicts a likely increase in hit ratio and masterpiece quality over time and is fairly agnostic on major-minor work production correlations.

Method

Composers Composers were selected from three sources: one list based on eminence (Farnsworth, 1969,p. 228), one based on performance frequency (Moles, 1958/66, p. 28), and one list of aesthetic significance ratings of composers' masterpieces (Halsey, 1976). The top ten composers in the first two lists were automatically sampled. Other composers who had numerous masterworks listed by Halsey (1976) were also included. The 18 composers examined were J.S. Bach, Bartók, Beethoven, Brahms, Chopin, Debussy, Dvořák, Handel, Haydn, Mendelssohn, Mozart, Schubert, Schumann, Richard Strauss, Stravinsky, Tchaikovsky, Verdi, and Wagner. Collectively, their lives span three centuries and their works account for 54% of all classical music performed (Moles, 1958/1966, p. 28).

Works Chwiałkowski's (1996) comprehensive catalog provided lists and dates of all known works of the 18 composers (6,560 compositions total). Arrangements, revisions, and lost works were excluded from most analyses.

Age at Composition Age at composition for each work was also determined (using Chwiałkowski, 1996). When composition spanned multiple years, the median (rounded up if necessary) was used. For aggregate analyses, works were grouped into consecutive five-year age periods for each composer (cf. Simonton, 1977): 5 – 9, 10 – 14, etc. Thus, each composer contributed data only during his active composing career. Ages were put into mean-deviation form based on the entire sample of composers, yielding an *age linear* variable. Age linear scores were also squared, yielding an *age quadratic* variable to test for curvilinear effects.

Works were also pooled into one-year age windows for finer-grained analyses. A total of 714 age periods were tallied across the 18 composers. *Age linear* and *age quadratic* measures were calculated in the same way for one-year windows as for five-year windows.

Weighting Performance times were used to weight each work. Times were taken from recordings (74% of works), estimates based on scores or listings in composers' work catalogs (14%), averages based on comparable works by the same composer (8%), and estimates using Simonton's (1977) genre-based system when no other information was available (4%).

Masterpiece definitions Four criteria were used to define masterpieces. Two used Halsey's (1976) list of musical masterpieces: 1) all works listed by Halsey, 2) only works given one of the two highest ratings (scores of 1 or 2) by Halsey ('stringent Halsey'). Two used recording counts from 1) the *RED Classical Catalog* (Ford, 2003), or 2) *The Penguin guide to compact discs* (March, Greenfield, & Layton, 2001). Each composer's works were rank-ordered from most to least recorded, and cumulative performance durations were computed from the top down until the total approached 50%. Top works were designated masterpieces for each measure.

Concurrent productivity Two variables were related to overall output. One, *years in interval*, logged how many years per period involved actual composition. This ranged from 1 to 5 (periods when no music was written were not analyzed). The other, *output*, tallied the total amount of music (in minutes) each composer wrote in each five- or one-year period. To minimize overall productivity differences, individual *z*-scores for output were computed for each composer separately.

Results

Because each composer's works were exhaustively tabulated, not sampled from a larger population, inferential statistical tests are arguably inappropriate (cf. Simonton, 2000). Therefore, effect sizes and parameter estimates accompany inferential results, whose hypothetical *p*-values can be used as a heuristic for discussion.

Lifespan productivity

Lifespan productivity generally increases quickly early in a creator's career, peaks, and then declines more gradually (Simonton, 1977, 1997). When age is expressed in mean-deviation form, output usually correlates positively with *age linear* and negatively with *age quadratic*. When these effects are combined, they form the backwards, inverted J-curve characteristic of creative output over the lifespan.

To confirm this trend for the present set of composers, as in Simonton (1977), data from all composers were put into the same time-series for a multiple regression analysis. To minimize inter-composer productivity variance, *z*-scores for output across 5-year age periods were computed for each composer separately. These defined the dependent variable. *Age linear* and *age quadratic*, as well as *years in interval*, were then used to predict productivity. A total of 179 age periods were included in the analysis.

The regression was significant, F (3, 175) = 38.74, $p < .0001$, adjusted $R^2 = .39$. *Age quadratic* and *years in interval* were highly significant, respectively, ß = -.35 and .45, both $p < .0001$. The amount of variance uniquely accounted for by each variable is given by the squared semi-partial correlations, here $sr_i^2 = .09$ and .18, respectively. The regression coefficient for *age linear* was marginally significant, ß = .11, $p = .08$ ($sr_i^2 = .01$). Thus, the overall lifespan trajectory of total productivity echoes the frequently observed single-peaked function, although its present shape is closer to an inverted U than to a backwards, inverted J.

The peak age for productivity can be computed by finding the residual *output* after regressing *years in interval*, setting up a polynomial regression using *age linear* and *age quadratic* to predict residual *output*, and taking the derivative of the regression equation. When this was done, peak productivity was found to occur at 39.7 years. When untransformed productivity was analyzed, the results essentially replicated (peak = 37.1 years).

When age was statistically controlled, the results indicating an inverted U-curve replicated. In fact, peak age for productivity lies almost exactly in the middle of standardized career lengths ($z = +0.04$).

Major-minor work correlations

Recall that the chance-configuration theory predicts a positive correlation between major and minor work production over the lifespan. Here, major works were defined by inclusion in Halsey (1976). Major and minor works were positively correlated, r (177) = .36, $p < .0001$ ($r^2 = .13$), supporting the chance-configuration theory. However, only two individuals showed statistically reliable correlations: Strauss, $r = -.50$, $p = .05$, and Bach, $r = .93$, $p < .0001$. Among individual composers, the median $r^2 = .16$.

Hit ratio over the lifespan

This is the main analysis of interest. Recall that Simonton (1977) found that hit ratio was essentially flat over the entire durations of ten great composers' careers, showing no systematic change with age. However, this finding seems inconsistent with Hayes's (1989) finding that composers write no masterworks in the first decade or so of their careers. To investigate this issue with the present sample of 18 composers, a hit ratio was computed for each composer in each age period using each of the four measures (basic Halsey, stringent Halsey, *RED*, and *Penguin*).

Composers vary considerably in their overall lifetime hit ratios. When defined by all works included in Halsey (1976), these range from .18 (Handel) to .70 (Wagner). The M (SD) hit ratio among the composers was .44 (.17). Thus, to minimize variance due to inter-composer hit ratio differences, *z*-scores for hit ratios were computed separately for each composer on each measure.

Intercorrelations between hit ratios determined by the four masterpiece criteria were all highly correlated, with correlation coefficients ranging from .45 to .74. To simplify matters, an unrotated, principal components analysis was conducted and revealed two factors. Factor 1 accounted for 71.6% of the variance, with an eigenvalue of 2.87 and loadings ranging from .75 to .91. Factor 2 accounted for 16.1% of the variance, with an eigenvalue of .65 and loadings ranging from .63 to -.36. The stringent Halsey criterion, defining composers' very greatest masterworks, had the lowest loading on Factor 1 and the highest loading on Factor 2, and this pattern was exaggerated in varimax rotated and oblique analyses. This measure was thus analyzed separately. (In practice, the results reported below do not substantively differ depending on whether three or four measures are combined, or if each of the four measures is examined individually.)

Hit ratio multiple regression To test hit ratio trends over the lifespan, all composers were initially put onto the same time-series. Each composer contributed data only during his active composing career. In each age period, individual *z*-scores of hit ratios from the basic Halsey, *RED*, and *Penguin* measures were averaged, yielding the dependent measure (basic hit ratio). Individual *z*-scores for *output* in each age period, used as the dependent measure for productivity, were used here as a statistical control. *Age linear*, *age quadratic*, and *output* predicted hit ratio.

Contrary to Simonton's (1977) findings and the

predictions of the chance-configuration theory, the regression was highly significant, F (3, 175) = 40.0, p < .0001, adjusted R^2 = .40. Both *age linear* and *age quadratic* (but not *output*) significantly predicted hit ratio, *age linear* ß = .65, and *age quadratic* ß = -.43, both p < .0001 (sr_i^2 = .36 and .13, respectively). Since output was not significant, taking the derivative of the second-order polynomial regression equation yields the peak age for hit ratio: 52.6 years, almost 15 years after the peak for overall output. The basic hit ratio trajectory is shown in Figure 1.

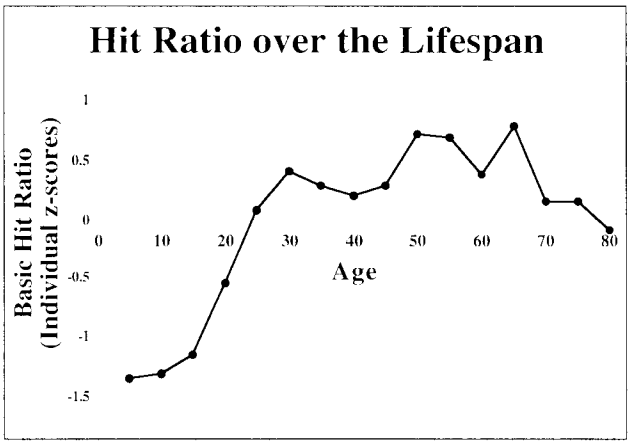


Figure 1: Average basic hit ratio over the lifespan (*z*-scores computed for each composer individually).

Subsample and individual analyses Why do these results differ so strikingly from Simonton's (1977)? Perhaps the additional eight composers affected the results. However, this is not the case: when his ten composers and the additional eight are analyzed separately (recalculating age *z*-scores for each subsample), the results essentially replicate: for Simonton's (1977) ten composers, the regression equation accounts for 37% of the variance in hit ratio, *age linear* ß = .64, p < .0001, (sr_i^2 = .38) and *age quadratic* ß = -.26, p < .05, (sr_i^2 = .04); for the remaining eight, the regression equation accounts for 41% of the variance in hit ratio, *age linear* ß = .63, age quadratic ß = -.52, both p < .0001 (sr_i^2 = .33 and .19, respectively) .

Each composer was also individually analyzed. In these analyses, one-year age windows were used and age *z*-scores were recomputed. The hit ratio of almost every individual composer was impacted by age (median adjusted-R^2 = .34). In particular, thirteen composers showed reliable (p < .05) positive *age linear* effects (median individual sr_i^2 = .22). Five showed reliable negative *age quadratic* effects (median individual sr_i^2 = .02). Only Bach and Strauss showed no age-related trends. The linear increase seems quite robust; the mid-career quadratic boost is consistently weaker.

Stricter hit ratio criterion The stricter Halsey criterion (reversed ratings of 4 or 5) was also tested. Hit ratios and individual hit ratio *z*-scores were recalculated, and the same three-predictor regression was performed. This demanding criterion reduces power, as happens whenever hit ratios approach 0 or 1. Despite this, the regression remained highly significant, F (3, 175) = 15.05, p < .0001, adjusted R^2 = .19, *age linear* ß = .49, p < .0001, (sr_i^2 = .20) and *age quadratic* ß = -.19, p < .05, (sr_i^2 = .02). *Total output* was not significant. Individual analyses, using one-year age windows, revealed ten composers with reliable (p < .05) positive *age linear* trends (median individual sr_i^2 = .12). *Age quadratic* was a weaker predictor (median individual sr_i^2 = .03). Three composers showed reliable quadratic trends, one (Wagner), negative, and two (Bartók and Schubert), positive, indicating a late surge in great masterwork production. Taking the derivative of the polynomial regression equation (after accounting for *output*) yields a peak age was 65.1 years for stringent hit ratio, notably later than that for productivity (late thirties) or the basic hit ratio measure (early to mid-fifties).

Mature works Perhaps these results, particularly the repeatedly observed *age linear* effect, are an artifact of composers' unacclaimed early works (recall Hayes's (1989) ten-year rule), which dampen hit ratio at career outset. To test this, the first two age periods of each composer (plus the third periods of Bartók and Dvořák, when both hit ratios were 0) were dropped from the analysis. Age *z*-scores were recalculated. The average of the basic Halsey, *RED*, and *Penguin* hit ratios defined hit ratio. The three-predictor regression remained significant, F (3, 140) = 7.79, p < .0001, adjusted R^2 = .13. *Age linear* positively predicted hit ratio, ß = .29, p = .001 (sr_i^2 = .07). *Age quadratic* and *output* each negatively predicted hit ratio, respectively, ß = -.30, p = .001 (sr_i^2 = .08), and ß = -.15, p = .09 (sr_i^2 = .02). The agewise effects are attenuated during compositional maturity, but the effects remain reliable.

Arrangements and revisions Not incorporated into any previous analyses were arrangements and revisions that were included (albeit minimally weighted) in Simonton's (1977) analysis. As a final test of the work measure, arrangements, revisions, and datable lost works were included in the analysis. Simonton's (1977) weighting system was employed. Hit ratios were recalculated using inclusion in Halsey (1976) as masterpiece criterion. A total of 8,057 data points were included. The number of age periods also increased, to 186. Untransformed hit ratios in each age period were used as the dependent measure, and *years in interval* was used as a control, echoing Simonton (1977). The results replicated and are comparable in strength to most previous analyses. The three-predictor regression remained significant, F (3, 182) = 31.85, p < .0001, adjusted R^2 = .33. *Age linear* positively predicted hit ratio, β = .57, p < .0001 (sr_i^2 = .27). *Age quadratic* negatively predicted hit ratio, β = -.30, p < .0001 (sr_i^2 = .06). *Output* was a marginally significant predictor, β = .13, p = .05 (sr_i^2 = .01). Eight composers showed significant positive *age linear* trends, and three showed significant *age quadratic* trends (two negative, one positive).

Thematic measure A final analysis used a measure based on musical themes or melodies rather than entire works (cf.

Simonton, 1977). Hit ratio in each age period was defined by the number of melodies included by Barlow and Morgenstern (1948, 1950) in all works listed by Halsey (1976) divided by the total number of anthologized melodies from that age period. Fewer age periods (141) were represented by this measure than the work measure, since not all periods contained anthologized themes. Age *z*-scores for the sample were recalculated. Here, *total themes* per period was used as a control. The three-predictor regression predicting hit ratio remained significant, contrary to Simonton's (1977) report: F (3, 137) = 9.64, $p < .0001$, adjusted $R^2 = .16$. *Age linear* positively predicted hit ratio, ß = .37, p < .0001, ($sr_i^2 = .12$) and *age quadratic* negatively predicted hit ratio, ß = -.26, $p = .002$, ($sr_i^2 = .06$). *Total themes* was marginally significant, ß = .15, $p = .06$, ($sr_i^2 = .02$).

Alternative analyses In addition to the analyses reported here, additional analyses ruled out further candidate alternative explanations for the inconsistency of the present results with those of Simonton (1977). Among the analyses performed were those using the raw average hit ratio for the basic *Halsey*, *RED*, and *Penguin* criteria, rather than individually z-transformed hit ratios; using an aggregate one-year age window rather than a five-year age window; computing age based on years of musical study (career age) rather than chronological age; analyzing only mature works using a one-year window dating from the onset of musical study or of musical composition; and performing all of these analyses on each two subsamples of composers: Simonton's (1977) ten and the additional eight. The same pattern of a significant, positive *age linear* trend coupled with a significant (but weaker), negative *age quadratic* trend was found in each analysis.

Discussion

Without exception, the present results contradict Simonton's (1977) null finding on agewise changes in classical composers' hit ratios. *Age linear* increases were uniformly found in analyses using two sets of masterpiece criteria, two measures of overall output (works and themes), two weighting systems, statistically controlling for individual differences in lifetime hit ratio, analyzing only composers' mature works, including arrangements and revisions and lost works, and separately examining two subsamples of the composers. Individual analyses generally showed reliable ($p \leq .05$) *age linear* increases in at least half of the sample, and no change in hit ratio in the rest. A weaker, negative *age quadratic* trend was also frequently evident, suggesting a mid-career boost in hit ratio. These findings support the problem solving and peak age perspectives and seem largely inconsistent with the chance-configuration theory. An increase in the quality of masterpieces, also significant in about half of individual composers, further supports the problem solving and peak age perspectives. The chance-configuration theory was supported, however, by a somewhat lower (though still reliable) positive correlation between the production of major and minor works.

It remains unclear why these results differ so dramatically from those reported by Simonton (1977). Numerous alternative explanations (e.g., including arrangements and revisions, using various masterpiece criteria) were tested and ruled out. More detailed comparisons are difficult to make, since Simonton (1977) reported few descriptive statistics about the analyzed compositions, and sometimes the reported statistics contradict the present data (e.g., in the number of age periods analyzed). Unfortunately, because of computer technology advances, his original data are no longer retrievable (Simonton, personal communication).

Aspects of the observed agewise changes are compatible with both the problem solving and chance-configuration perspectives. For instance, much of the improvement in hit ratio occurs in the early part of composers' careers, consistent with expectations about skill and expertise acquisition as a precondition for outstanding creative achievement (Hayes, 1989; Weisberg, 1999). However, age trends during compositional maturity accounted for only one-third to one-half as much variance as those spanning composers' entire careers, which is at least *somewhat* more consistent with the chance-configuration theory.

The slight hit ratio decline late in life is inconsistent with both theories. The problem solving perspective does not posit a mechanism to explain a drop in perspicacity late in life. The hit ratio decline among elderly composers also violates the equal-odds rule applying until the end of creators' lives. However, this decrement should be interpreted cautiously, due to a low N and high variability in hit ratios of composers in their seventies and beyond.

The pattern of peak ages is more difficult to reconcile with the chance-configuration theory. Overall productivity peaks first, in composers' late thirties, consistent with previous work (see Simonton, 1997). Hit ratio defined by the three less stringent masterpiece criteria peaks around age 53. The more stringently defined hit ratio peaks even later, at 65 years. Note also that these data are not minor statistical aberrations, as they represent parameters for the career landmarks of these 18 composers.

These results dissociate two predictions of the chance-configuration theory. Many studies, including this one, show positive major-minor work correlations. However, such correlations do not preclude an agewise improvement in hit ratio and cannot by themselves rule out alternative theoretical models. Comprehensive analyses of agewise hit ratio trends, as performed here, are also necessary.

The consistent *age linear* increase and late peaks for hit ratio and masterpiece quality suggest that at least some creators boost creativity as their careers progress. While certainly quantity of ideation provides the raw material for creative productivity, the present results suggest that some creators are perspicacious and able to consolidate gains as they elaborate ideas. Thus, goal-directed problem solving and evaluation processes, as well as accumulated expertise (Hayes, 1989; Simonton, 2000; Weisberg, 1999), play substantial roles in creative productivity. Naturally, creators' instincts are not perfect, but many composers'

intuitions about the value of their ideas appear to have been sound, and they seem to have improved these intuitions substantially with age.

The evaluative and elaborative processes that might permit an increase in hit ratio and masterpiece quality remain ill defined, but these results suggest that both cognitive and motivational variables play key roles. Average hit ratio does not increase linearly but is a more curvilinear, single-peaked function. Later in their lives, composers write less music than in midcareer, but a progressively higher proportion is masterpiece-level, and the masterpieces themselves are of higher quality. What older composers lack in energy, they appear to make up for in wisdom.

The underlying psychological processes may also be informed by further ideographic analyses. Here, over half of the 18 composers showed systematic agewise increases in hit ratios and masterpiece quality, and the rest showed few agewise changes. Further research linking career trajectory with characteristic creative processes of composers (cf. Galenson, 2001) may inform the conditions under which improvement in hit ratio or masterpiece quality may be realized.

The present data must be interpreted with caution. They represent a single (but robust) violation of the equal-odds rule. However, they clearly show that the equal-odds rule does not universally characterize creative productivity. In addition to ideographic analyses, comprehensive studies like this one are necessary to determine the extent of equal-odds rule application (or violations) in other domains. At a minimum, the results suggest that there can be important roles for elaborative and evaluative problem solving processes, as well as for learning, in creative productivity.

Acknowledgments

I thank two anonymous referees, Arthur Reber, Laraine McDonough, David Owen, and Neil Macmillan, who offered constructive conceptual and methodological advice at various stages of the project's development. Dean Keith Simonton provided valuable information on the methodology and weighting system of his 1977 study. I also thank the Brooklyn College Library, the New York University Elmer Bobst Library, the New York Public Library for the Performing Arts, as well as the Tower Records, HMV, and Virgin Megastores of New York City for access to musical scores, reference books, and recordings. Cynde McKenzie, Jodie Cohen, and Bracha Schoffman assisted with some data coding.

References

Barlow, H., & Morgenstern, S. (1948). *A dictionary of musical themes*. New York: Crown.

Barlow, H., & Morgenstern, S. (1950). *A dictionary of opera and song themes*. New York: Crown

Campbell, D. T. (1960). Blind generation and selective retention in creative thought as in other thought processes. *Psychological Review, 67*, 380-400.

Chwialkowski, J. (1996). *The Da Capo catalog of classical music compositions*. New York: Da Capo.

Ericssson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in acquisition of expert performance. *Psychological Review, 100*, 363-406.

Farnsworth, P. R. (1966). *The social psychology of music* (2nd ed.) Ames, IA: Iowa State University Press.

Ford, G. (Ed.) (2003). *The RED classical catalog*. London: RED Publishing.

Galenson, D. W. (2001). *Painting outside the lines: Patterns of creativity in modern art*. Cambridge, MA: Harvard University Press.

Halsey, R. S. (1976). *Classical music recordings for home and library*. Chicago: American Library Association.

Hayes, J. R. (1989). *The complete problem solver* (2nd ed.). Hillsdale, NJ: Erlbaum.

Kozbelt, A. (in press). Factors affecting aesthetic success and improvement in creativity: A case study of the musical genres of Mozart. *Psychology of Music*.

March, I., Greenfield, E., & Layton, R. (2001). *The Penguin guide to compact discs: The guide to excellence in recorded classical music*. New York: Penguin Books.

Moles, A. (1966). *Information theory and esthtetic perception* (J.E. Cohen, Trans.). Urbana, IL: University of Illinois Press. (Original work published 1958)

Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In H. E. Gruber, G. Terrell, & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking* (pp. 63-119). New York: Atherton Press.

Simonton, D. K. (1977). Creative productivity, age, and stress: A biographical time-series analysis of 10 classical composers. *Journal of Personality and Social Psychology, 35*, 791-804.

Simonton, D. K. (1985). Quality, quantity, and age: The careers of 10 distinguished psychologists. *International Journal of Aging and Human Development, 21*, 241-254.

Simonton, D. K. (1986). Aesthetic success in classical music: A computer analysis of 1,935 compositions. *Empirical Studies of the Arts, 4*, 1-17.

Simonton, D. K. (1988). *Scientific genius: A psychology of science*. Cambridge: Cambridge University Press.

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career landmarks and trajectories. *Psychological Review, 104*, 66-89.

Simonton, D. K. (1999). *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.

Simonton, D. K. (2000). Creative development as acquired expertise: Theoretical issues and an empirical test. *Developmental Review, 20*, 283-318.

Weisberg, R. W. (1999). Creativity and knowledge: A challenge to theories. In R.J. Sternberg (Ed.), *Handbook of creativity* (pp. 226-250). New York: Cambridge University Press.

Constructing and Revising Mental Models of a Mechanical System: The role of domain knowledge in understanding external visualizations

Sarah Kriz (kriz@psych.ucsb.edu)
Department of Psychology, University of California
Santa Barbara, CA 93106 USA

Mary Hegarty (hegarty@psych.ucsb.edu)
Department of Psychology, University of California
Santa Barbara, CA 93106 USA

Abstract

External visualizations such as diagrams and animations are frequently used to teach people about the workings of mechanical systems. The present study considers the types of mental models that can be constructed from visual-spatial (non-verbal) materials alone, and the extent to which people revise their incorrect mental models. Comparing 10 high physics knowledge participants to nine low physics knowledge participants, we assessed how these two groups constructed and revised mental models of a flushing cistern. High domain knowledge participants extracted more meaningful information from the materials, although their initial models of the system were not as accurate as expected. However, after answering comprehension questions and viewing the learning materials again, high domain knowledge participants were more likely to revise their mental models into correct representations of the system, whereas the participants with low domain knowledge continued to rely on incorrect models. The discussion of these findings focuses on how prior knowledge may contribute to understanding visual instructional materials.

External visualizations (e.g., diagrams and computer animations) are often used to inform people about how a complex system behaves. Physical systems such as machines are causally and temporally complex, and understanding these systems depends on an appreciation of the spatial relations between their components and how these change over time. The spatial and temporal aspects of mechanical movements can be illustrated directly via visual-spatial representations, while the same information presented in a verbal format might be more difficult to understand.

It seems plausible that the design of external visualizations could greatly affect one's success at extracting relevant information from the display. For example, adding accompanying text that describes aspects of phenomena presented in the visualizations may help in providing additional information that a diagram alone could not provide. Previous studies researching the integration of text and diagrams have shown that people with low domain knowledge or low spatial ability rely heavily on accompanying textual descriptions (Hegarty & Just, 1993; Kalyuga, Chandler, & Sweller, 1998). These studies also suggest that as a person becomes more familiar with a domain, the reliance on textual explanations decreases. However they have not specifically addressed which types of information are best understood from diagrams and animations by people with different amounts of background knowledge.

The purpose of the present study is to examine how understanding a mechanical system is achieved when visual materials such as diagrams and animations are presented without accompanying textual or verbal explanations. As part of a larger research objective, we analyze how purely visual materials are understood by both high and low physics knowledge participants. By examining how people with varying degrees of domain knowledge interpret visual materials, we can design future materials containing verbal descriptions that supplement informational gaps in the visual displays. We assume that these informational gaps will differ for low and high domain knowledge individuals, therefore, this study focuses on how domain knowledge contributes to the comprehension of visual materials conveying a complex mechanical system.

Constructing Mental Models

Creating a mental model of a complex system requires that a person identify the parts involved, understand their causal relationships, and relate the causal steps to the larger functions of the system. Our cognitive model of how people come to construct mental models from multimedia materials follows that outlined by Narayanan & Hegarty (2002). To summarize, there are five steps that a person must take to understand a machine from multimedia presentations. First, the system must be decomposed into individual components. Second, the learner must make representational connections to prior knowledge. Third, if verbal information is present, a person is required to make further referential connections between the visual media and the verbal explanations. Then, she must determine the causal chain of events. Finally, a dynamic mental model is constructed. In the present experiment, we are particularly interested in the second step-- how prior domain-related knowledge affects the construction and revision of mental models.

We predict that, in accordance with previous studies (Spilich et al., 1979; Chiesi, Spilich, & Voss, 1979; Lowe, 1994; 1999), high domain knowledge individuals will be more likely to construct initial mental models that incorporate high-level functional understanding, whereas people lacking domain-related knowledge will focus on the movements of the parts on a local level. Additionally, we expect the level of domain knowledge to influence the extent to which models are revised. Assuming that learning is an iterative process of understanding (Miyake, 1986), how people move from a state of understanding to a state of non-understanding may depend on their level of domain-related knowledge. As previous studies have shown (Chi, 2000; Chi et al., 1994), it is when conflicts between internal models and external information occur that people are more likely to revise their internal mental models. We propose that conflicts are more likely to be perceived by people with high domain knowledge because they are at an informational advantage for meaningfully evaluating their models.

Three Types of Mental Models

The stimulus used in this experiment was a British model of a toilet tank. While the purpose of the system is the same as an American model (i.e., to flush water into the toilet bowl), the mechanism used to accomplish this function differs vastly from its American counterpart. Thus, we assume that the participants in our experiment (American college students) did not have prior knowledge of the mechanism that they studied in this experiment.

Specifically, the main difference is the manner in which water exits the tank into the bowl. In the British model, water exits the tank through a siphon process. The siphon process begins when two disks (located at the bottom of the bell in the middle of the diagram) are pushed together (by pulling the handle) and push water up through the main siphon pipe. As water flows up the siphon pipe and down into the toilet bowl, the siphon process begins. This enables water to flow through the siphon pipe without the aid of the disks. The process ends when the water level in the tank falls below the siphon bell, and air enters the siphon pipe. This is reflected in Figure 1.

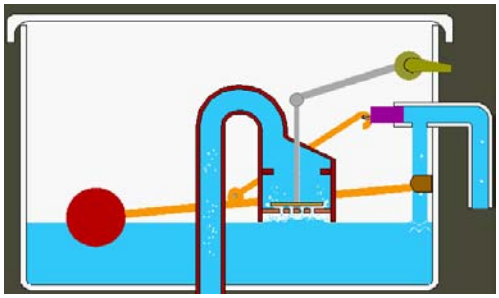


Figure 1: Air enters and stops the siphon process.

Data from this and previous experiments (e.g., Hegarty, Kriz, & Cate, 2003), indicate that people construct one of two types of mental models of the flushing cistern. The first (physically correct) model works according to the physical process of siphoning. As explained above, a siphon occurs when liquid in an enclosed system moves, via a pressure differential, from a point of high pressure to a point of lower pressure. In the case of the British toilet tank shown in Figure 1, a siphon enables water to continuously flow up and back down the large pipe in the middle in order to exit the tank. The siphon is broken when air, which is lighter than water, enters the enclosed system. This breaks the pressure differential, and the water flow out of the pipe stops.

The incorrect model of this system involves the disks as the main stopping agents. Participants with incorrect models may or may not understand the initialization of the siphon process. Their model is characterized, however, by the function of the disks. The incorrect model assumes that the water stops flowing out of the tank because the upper disk falls on the lower disk and creates a water-tight barrier, or seal. As Figure 1 illustrates, the two disks do not touch when the lower disk falls. If they did, air would not be able to enter. Therefore, the visual materials that participants view are in direct conflict with this model.

Many participants do not offer an explanation for how water starts and stops flowing from the tank to the bowl. A possible reason for this omission is that they have an informational "gap" (Chi, 2000) in their mental model. In other words, they are missing the knowledge necessary to explain the causal relationship between activity in the tank and the stopping of exiting water. However, it is quite possible that this process is, in fact, represented in their mental models but was simply not explained during the protocol. Because there was no method for empirically differentiating between these two possibilities, these cases were not considered in analyses.

Method

Participants

Nineteen adults (10 high domain knowledge, 9 low domain knowledge) volunteered for the study as paid participants. The high domain knowledge (HDK) participants were UCSB graduate students from Mechanical Engineering or Material Science, with the exception of one participant, who was the staff lab manager of the undergraduate Mechanical Engineering lab. All experts held Bachelors degrees in engineering or physics and had been studying physics and engineering for a mean of 6.4 years (range 5-8 years). It was assumed that the HDK group had knowledge of pressure differentials and siphoning, as these topics are covered in undergraduate physics and engineering courses.

The low domain knowledge (LDK) participants were UCSB graduate students from Social Sciences, Art, Humanities, or Biology, and all considered themselves physics novices. Four had taken introductory physics in high school and one had taken freshman physics in college. None had taken engineering courses.

Materials and Apparatus

Participants viewed a variety of visual displays depicting a toilet tank either in a resting state or in motion. The *labeled static diagram* showed a color picture of the toilet tank in its resting state and included labels naming the mechanical parts. The *unlabeled static diagram* was identical to the labeled version, but without the labels. The *four phases static diagram* was a series of four diagrams displayed together. Each diagram showed a different phase of the flushing process. The labeled, unlabeled, and four phases diagrams were all viewed as PowerPoint slides.

Three animations were also available to the participants. All of the animations consisted of a series of 134 bitmap images. The *computer-controlled animation* was displayed in Macromedia Flash MX and played at a rate of 6 frames per second. The participant pressed a button with the mouse in order to begin playing, but otherwise had no control over the speed or the direction of the animation. The *participant-controlled animation with arrows* was run in a Quicktime player, which allows one to control the speed and direction in which a video file plays. In both animations, arrows appeared at various points to indicate a part's direction of movement or to signal an important event. The *participant-controlled animation without arrows* was identical to the other participant-controlled animation except it did not contain arrows.

All materials were displayed on a 17" desktop monitor at 1024x768 resolution.

Procedure

Participants sat in front of the computer and were told, "You are going to view diagrams and animations that illustrate a toilet tank, but note that this is not an American model. Please view the materials and learn how the system works. You have as much time as you need to study the materials." They were then shown the six visual learning aids. The researcher briefly explained each learning aid, without mentioning the presence or absence of arrows in the animations, and demonstrated how to manipulate the controls of the Quicktime Player.

After viewing the material, the monitor was turned off and participants were given a booklet of comprehension questions. The first question asked to explain step-by-step what happens in the toilet tank after the handle is pushed. The next four questions were troubleshooting questions, in which novel breakdown scenarios were described. The participants were required to generate as many responses as possible that would account for the breakdown of the system. The final four questions were function questions that asked about the function of specific parts of the mechanical system. Participants were asked to provide written answers at their own pace.

Upon finishing the written portion of the experiment, the participants were then asked to view the visual materials again. The participant-controlled animation without arrows was displayed and participants were asked to orally report to the researcher where events began and ended. An "event" was not predefined by the researcher, and participants were allowed as much time as they needed to formulate their answers before reporting. When they were ready, participants reported what they saw as "events." This portion of the session was video taped for later analysis.

Results

Constructing Initial Mental Models

In order to assess participants' initial mental models of the system, we evaluated the first written question, in which participants described the step-by-step process of a flush. The two groups' responses showed both quantitative and qualitative differences. As Figure 2 shows, the HDK participants reported on average four more steps than the LDK participants, and this difference reached significance: $M = 16.8$ v. $M = 12.8$; $t(17) = 3.176$, $p < .05$. This difference indicates that the HDK individuals were able to extract more information from the visual materials.

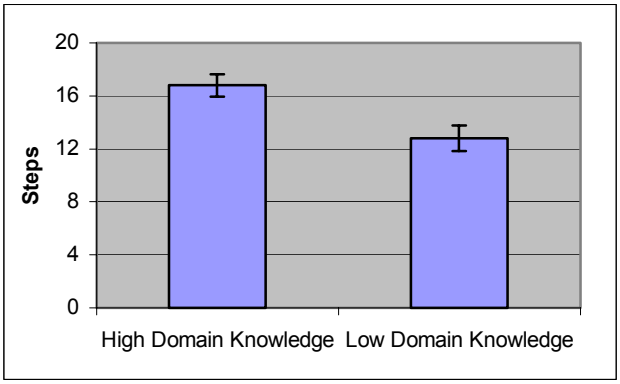


Figure 2: Mean number of steps in written responses.

Furthermore, the types of events that were mentioned by each group differed. The majority of participants (i.e., seven participants or more) in both groups reported eight common steps. However, the HDK participants tended to mention steps that the LDK students did not mention. (See Table 1.) Of note, the HDK participants tended to focus on the rising and falling of the water level, and that focus was not evident in the majority of the LDK participants' written reports.

To evaluate the accuracy of participants' initial mental models, the step-by-step written reports were evaluated for steps that reflected correct, incorrect, or unstated models of the siphon process. If participants mentioned the siphon process beginning when the disks pushed the water up the siphon pipe and ending when air entered the system, their models were considered correct. Incorrect models were those in which the disks were reported as stopping the water from leaving the tank. Finally, a mental model was coded as "not stated" if participants made no mention of the how the water stopped flowing out of the tank. As Table 2 shows, the distribution of initial model types did not differ at all between the two groups.

Table 1: Steps mentioned by at least 7 participants in their step-by-step written responses.

Steps	HDK	LDK
Push down on the handle	x	x
The upper disk moves up	x	x
The lower disk moves up	x	
Water enters siphon pipe	x	x
The upper disk moves down	x	x
The lower disk moves down	x	
Water level falls below siphon bell	x	
Water level lowers in tank	x	
Float lowers	x	x
Inlet valve opens	x	x
Water flows from inlet pipe to tank		x
Water level rises	x	
Float rises	x	x
Inlet valve closes	x	x

Data from the troubleshooting and function question responses reveal, however, that high and low knowledge participants did differ on how strongly they relied on incorrect models in later comprehension questions. Figure 3 reflects the mean number of troubleshooting and function question answers that contained the incorrect model. The data indicate that the LDK participants used incorrect mental models significantly more often than the HDK participants to account for system breakdowns and overall functions of the tank: M= 3.4 v. M=1.6; $t(17)$=2.535, p<.05.

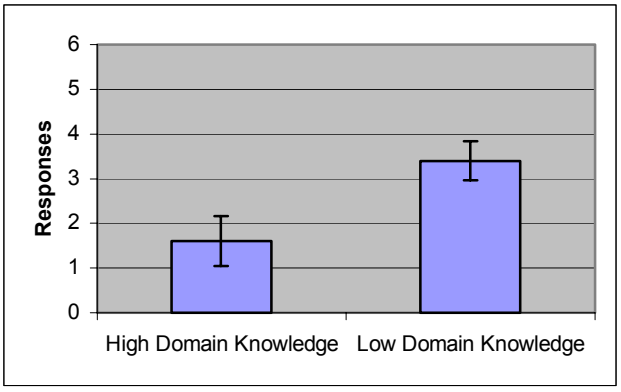


Figure 3: Mean number of troubleshooting and function responses conveying incorrect model.

Revising Mental Models

Although the written step-by-step responses and the orally presented event structure responses resulted from slightly different tasks, a paired samples t-test yielded no significant differences for either group in the number of steps mentioned between the two tasks. Thus, we were able to compare the written and oral reports in order to assess model revisions.

In their oral reports, the majority of participants in each group (at least 7) mentioned three steps that were not in their written reports. For the LDK participants, these included: (1) the lower disk moves up, (2) the lower disk falls down, and (3) water stops flowing into the siphon pipe. The three events that the majority of HDK participants mentioned in their oral reports but not in their written response were: (1) water flows into the toilet bowl, (2) the siphoning process begins, and (3) the siphoning process is broken. The striking difference between the two groups is in the perceptibility of the added steps. The three steps commonly added by the LDK participants are directly perceptible from the animation, which they were allowed to view while giving their reports. On the other hand, the HDK group added steps that were not directly perceptible from the animation, but instead involved higher-level processes and functions.

The number of participants in each group who changed from one model to another is shown in Table 2. As is evidenced in this table, the majority of the HDK participants moved from an incorrect or unstated model to a correct model. Whereas two HDK participants began the study with a physically correct model, eight of the ten finished the study with this model. That is, they orally reported that the siphon phenomenon, not the disks, ends the outflow process. Contrary to this, the LDK participants show no clear pattern of model revision. Many of their final models remain incorrect. In sum, the steps reported in the written response compared to the oral response clearly showed signs of model revision in the HDK sample, whereas no pattern was evident in the LDK data.

Discussion

Initial Models

The results of this study indicate that HDK participants were able to extract more information about the flushing cistern system from the visual materials provided. They not only reported more initial steps than the LDK individuals, but were also better at integrating higher-level causal changes, such the rising and falling of the water level, into their initial reports. As predicted, and following previous findings on reasoning with external visualizations (Lowe, 1994; 1999), the LDK participants' step-by-step reports revolved around small mechanistic movements and did not indicate a functional understanding of the system.

A comparison of HDK and LDK participants' written protocols revealed that the types of models constructed initially were similarly distributed across the two groups. This result was rather surprising, given that we expected the engineers to fully comprehend the siphon process upon viewing the materials. Additionally, we expected reading the label "siphon pipe" to prime this siphon schema. Our findings conflict with previous accounts of experts solving physics problems (Chi, Feltovich, & Glaser, 1981), as well as lay beliefs that people trained in a certain domain are able

Table 2: Summary of Initial and Final Models.

HDK	Initial Model	Final Model	LDK	Initial Model	Final Model
E01	Correct	Correct	N02	Incorrect	Not Stated
E02	Incorrect	Incorrect	N03	Incorrect	Incorrect
E03	Not Stated	Correct	N04	Not Stated	Incorrect
E04	Not Stated	Correct	N05	Correct	Not Stated
E05	Correct	Correct	N06	Incorrect	Incorrect
E06	Incorrect	Incorrect	N07	Incorrect	Incorrect
E07	Incorrect	Integrated[1]	N08	Incorrect	Incorrect
E08	Not Stated	Correct	N09	Not Stated	Not Stated
E09	Incorrect	Correct	N10	Not Stated	Incorrect
E10	Incorrect	Correct			

to understand domain-related phenomena quickly and easily. As evidenced by their later oral reports, the engineers were able to spontaneously report the siphon process, indicating that they had the relevant domain-knowledge, yet they did not grasp the process in their initial viewing of the materials.

Here we offer a possible explanation for why many of the HDK participants did not incorporate the siphon process into their initial mental models. In the initial viewing phase of the experiment, participants were trying to integrate their prior domain-related knowledge with the external visualizations in order to create a cohesive causal model of the system. Because the HDK individuals have a larger body of prior knowledge than the LDK participants, they have more explanations in competition. Two sources of knowledge that may contribute to the understanding or misunderstanding of how water stops exiting the tank are: (1) domain-specific knowledge about a siphon process and pressure differential, as described previously, and (2) domain-general knowledge of "damming." From experience with the real world, we know that flowing liquid can be stopped by solid objects. Integrating the damming principle with the external visualizations leads to the incorrect model of the disks blocking the outgoing water. This integration of domain-general information seems to be how LDK individuals reason about the flushing cistern. Although both explanations were available to the HDK participants, the domain-specific explanation may not have been adequately cued by the visual materials. Thus, the HDK relied on their domain-general knowledge until they had reason to switch to using domain-specific understanding. This explanation is purely speculative, and further studies exploring these issues need to be conducted.

Model Revision

While the distribution of both groups' initial model types were found to be relatively similar, analyses of their final oral reports revealed that the model revision process differed across the groups. Although both groups did, on average, add three steps that were not present in the majority of the initial models, the steps differed qualitatively across the groups. The LDK participants seemed to perform "model addition" after answering the comprehension questions and reviewing the visual materials. They *added* visually salient information that was left out of their initial models. However, the LDK participants did not *change* their model to the correct model. Moreover, the sealing of the disks continued to be the dominant view of how the water stopped exiting the tank, even though this explanation was in direct conflict with what was shown in the visual materials. These findings are consistent with previous findings that LDK individuals do not integrate functional information into their mental models (Spilich et al., 1979; Chiesi, Spilich, & Voss, 1979; Lowe, 1994; 1999) and that they tend to stay at a kinematic/behavioral level of explanation, even after spending additional time with the learning materials (Hale & Barsalou, 1995).

The HDK participants in this study can be described as truly revising their mental models. Rather than simply adding perceptually salient steps to their final models, the majority of HDK participants changed their models to include the beginning siphon process and the correct explanation for the ending of the siphon process. The HDK participants did not tend to rely as heavily on their initial incorrect model while answering the troubleshooting and function questions. This indicates that the HDK group was more flexible in generating other responses to the questions, possibly because they had more prior knowledge available to them.

There are many possible explanations of how the revision process occurred. Because a variety of activities took place between the initial written step-by-step explanations and the final oral reports, we can only speculate on the possible causes of revision. One possibility is that the troubleshooting and function questions lead the HDK group to internal conflicts within their models. Troubleshooting

[1] The participant incorporated both the correct and incorrect models into his final model.

questions can be used to induce causal knowledge of a system (Hale & Barsalou, 1995), and can also be used to judge deep comprehension (Graesser & Olde, 2003). While providing responses to these questions, HDK participants might have become aware of conflicts between their initial mental model and the possible explanations for the breakdown scenarios presented in the troubleshooting questions. Alternatively, viewing the visual materials a second time may have contributed to model revision. As the HDK participants were able to compare their mental models to the information presented in the visual tools, they may have realized inconsistencies in their models.

The LDK group, on the other hand, did not seem to experience conflicts between their models and their troubleshooting and function responses, nor between their models and the external representations. Although their models conflicted with what was shown in the learning materials, none of the participants explicitly identified a conflict. These participants may have refrained from revising their mental models even after re-viewing the materials because they began to rely on their mental models as perceptual evidence (Rozenblit & Keil, 2002). The HDK group, on the other hand, seemed to be more sensitive to conflicts between their mental models and information that did not match.

Conclusion

This study demonstrates the limitations of visual materials such as diagrams and animations for communicating about how machines work. Although the animations showed how the parts of the mechanism move when it is in operation, LDK individuals were unable to construct an accurate mental model from the visual materials, and tended to construct erroneous mental models that were in fact inconsistent with what they viewed. Most HDK individuals were able to construct the correct mental model eventually, but this took some time and occurred only after engaging in other activities such as answering troubleshooting and function questions.

The results of this study suggest that materials designed for low domain knowledge participants must explicitly describe the siphon process (e.g., through language), while materials targeting learners with adequate domain knowledge may need to merely induce these learners to access the relevant domain information they already possess. Thus, examining the mental models constructed from visual materials alone can provide insights into the design of instructional materials for individuals with different amounts of background knowledge, and suggest when and how visual-spatial instruction materials should be supplemented by verbal instruction.

Acknowledgments

This research is supported by the Office of Naval Research grant number N00014-03-1-0119.

References

Chi, M. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology: Vol 5. Educational Design and Cognitive Science* (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum.

Chi, M. de Leeuw, N., Chiu, M., LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chiesi, H., Spilich, G., Voss, J. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 257-273.

Graesser, A. & Olde, B. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95*(3), 524-536.

Hale, C. & Barsalou, L. (1995). Explanation content and construction during system learning and troubleshooting. *Journal of the Learning Sciences, 4*(4), 385-436.

Hegarty, M. & Just, M. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition and Instruction, 21*(4), 325-360.

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40*(1), 1-17.

Lowe, R. (1994). Selectivity in diagrams: Reading beyond the lines. *Educational Psychology, 14*(4), 467-491.

Lowe, R. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education, 14*(2), 225-244.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science, 10*, 151-177.

Narayanan, N. H. & Hegarty, M. (2002). Multimedia design for communication of dynamic information. *International Journal of Human-Computer Studies, 57*(4), 279-315.

Rozenblit, L. & Kiel, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521-562.

Spilich, G., Vesonder, G., Chiesi, H., Voss, J. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 275-290.

Causal Structure in Conditional Reasoning

Tevye R. Krynski (tevye@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139

Abstract

Causal reasoning has been shown to underlie many aspects of everyday judgment and decision-making. We explore the role of causal structure in conditional reasoning, hypothesizing that people often interpret conditional statements as assertions about causal structure. We argue that responses on the Wason selection task reflect the selection of evidence expected to maximally reduce uncertainty over candidate causal structures. We present a model in which people's selections depend on their interpretation of which causal relationship is asserted by a given conditional statement.

Introduction

Consider the following statement: "If a pot falls in the kitchen, then you will hear a clang". Is this statement true? Not if something breaks its fall, like a pillow. Now consider the statement: "If a clang is heard then a pot has fallen in the kitchen." Is this statement true? Not if something else can cause a clang, such as falling silverware. The first statement is not always true because there are conditions that can disable the mechanism by which falling pots cause clangs to be heard. The second statement is not always true because there are alternate causes of clangs other than falling pots. As this example illustrates, causal knowledge often underlies how people reason about conditional statements.

Recent research has shown that causal reasoning permeates many aspects of cognition, including associative learning (Waldmann, 2000; Glymour & Cheng, 1998), category learning (Rehder, 2003; Ahn, 1999), and judgment under uncertainty (Krynski & Tenenbaum, 2003). In this paper we analyze the role of causal structure in conditional reasoning (Over & Jessop, 1998), and argue that people's responses on the Wason selection task reflect sophisticated abilities to induce causal structure.

An important open question in causal reasoning is how people's background knowledge interacts with observations when inferring causal structure. Causal domain knowledge places important constraints on which cause-effect relationships exist and how the effects depend functionally on the causes (Pearl, 2000; Krynski & Tenenbaum, 2003; Ahn, Kalish, Medin, & Gelman, 1995). This effectively specifies a hypothesis space of candidate causal structures, which we model using causal Bayes nets (Pearl, 2000). Observational evidence can then be used to determine which causal structure is most likely. We propose that this interplay of causal domain knowledge and observational evidence underlies people's judgments on the Wason selection task.

The Wason selection task presents subjects with a conditional statement of the form "if p then q", and asks subjects to choose evidence to determine whether the statement is true. Prior accounts of people's responses on the selection task have emphasized logical reasoning (Wason, 1966; Ahn & Graham, 1999), probabilistic reasoning (Oaksford & Chater, 1994), or social reasoning (Cosmides, 1989), as well as others. In contrast, we argue that the selection task often engages causal reasoning: for conditional statements in which p and q are causally related, people choose cards that will be most useful to determine which of several candidate causal structures is correct for a given situation.

We have developed a model that extends Oaksford & Chater's (1994) probabilistic information gain framework to handle causal hypotheses. The information gain framework of O&C proposes that in the Wason selection task, people seek to reduce their uncertainty among hypotheses about the relationship between the antecedent (p) and the consequent (q) in a conditional statement of the form "if p then q". The model of O&C (1994) proposes that these hypotheses are assertions about conditional dependencies (e.g., q depends on p, q is independent of p, etc.), whereas we propose that these hypotheses are assertions about causal structure (e.g., p causes q, p does not cause q, etc.).

Our causal framework enables us to explain some previously puzzling results from the literature, as well as compelling intuitions that are not predicted by other approaches. We also address an important open question with both logical and probabilistic accounts: they leave unspecified how people interpret conditionals to determine which hypothesis is being asserted. We propose that the interpretation of conditionals often depends on causal domain knowledge, which imposes constraints on candidate causal structures, as well as pragmatic considerations.

Why interpret conditionals causally?

In contrast to O&C's proposal that conditional statements assert a conditional dependency, we propose that people interpret conditional statements in which p and q are causally related as assertions about causal structure. The underlying reason for this is that conditional dependencies are often a symptom of some underlying causal relationship. "If p then q" states that there is some dependency between p and q, which in turn implies there is some mechanism by which p and q are related; i.e., p causally influences q, q causally influences p, or they have some common cause. The term "causally influences" does not necessarily mean

"directly causes"; causal influence can be generative, inhibitive, enabling, permissive, or otherwise.

Examples of the prevalence of causal interpretations come from statements in which the logical interpretation and the causal interpretation are at odds; in these cases, the causal interpretation tends to take precedence. Some conditionals are logically false but seem true because they are causally true. For example, "If you spin around then you will get dizzy" seems true enough, although it's possible to spin around without getting dizzy, therefore it's logically false. Other conditionals are logically true but seem false because they are causally false. For example, "If you drink coffee during the day then you will fall asleep at night" sounds false because it seems to be saying that coffee causes you to fall asleep, but it is logically true (assuming you eventually fall asleep every night). These examples suggest that it is often, but not always, more natural to interpret conditionals as causal assertions, rather than logical implications.

Causal Structure Induction

We adopt the following Bayesian framework: given conditional statement "if p then q", reasoners consider a total hypothesis space T of candidate causal structures relating p and q. The conditional statement is interpreted to be asserting that a specific causal relationship holds between p and q. T then partitions into a subspace of structures S consistent with the statement, and its complement, T-S, inconsistent with the statement. Testing the conditional amounts to testing whether the true structure, s^*, is in S or T-S. The probability that the conditional is true is the probability that s^* is in S: $P(s^* \in S) = P(S) = \sum_{s \in S} P(s)$.

Initial degrees of belief in these hypotheses are represented as prior probabilities, and those structures that do not satisfy the constraints of causal domain knowledge are not considered. For example, people know that falling pots can cause noise, but noise cannot cause pots to fall, hence no structures with noise causing falling pots will be included in T. In this case of the conditional "if a pot falls, then it makes a noise", T could be the set of all causal structures consistent with domain knowledge in which falling pots exist, and S could be the subset of structures in T in which falling pots are a cause of noise.

Data can help determine how likely the conditional statement is to be true. Using Bayesian belief updating,

$$P(S \mid d) = \sum_{s \in S} P(s \mid d) = \sum_{s \in S} \frac{P(s)P(d \mid s)}{P(d)}$$

According to the information gain (IG) approach (O&C, 1994), when determining whether a particular conditional statement is true, the most informative data are those that are expected to maximize information gain, I_g:

$$I_g(S \mid D) = \sum_H P(H) \log \frac{1}{P(H)} - \sum_H P(H \mid d) \log \frac{1}{P(H \mid d)}$$

However, O&C (1996) propose that when $P(S)$ is not 0.5, a better measure is the distance between the probability distributions of the new and old beliefs, as measured by Kullback-Leibler distance, (we will use this, and call it I_{KL}):

$$I_{KL}(S \mid d) = P(S \mid d) \log\left(\frac{P(S \mid d)}{P(S)}\right) + P(T-S \mid d) \log\left(\frac{P(T-S \mid d)}{P(T-S)}\right)$$

In the case of the Wason selection task, $I_{KL}(S|d)$ is the amount of information gained from turning over cards. The selection task can be used to test two claims: (1) people often interpret conditional statements in which p and q are causally related as assertions that a particular causal relationship holds, and (2) people select information with the goal of maximally reducing uncertainty in that assertion.

Applying the IG approach to the selection task

The Wason selection task and its variants present people with a conditional statement of the form "if p then q", where p and q can be any propositions. Cards are then presented which represent trials; one side specifies whether p was true on the trial, while the other side specifies whether q was true. Subjects are presented with four cards, having each of the four possible sides (p, q, $\neg p$, $\neg q$) facing up. The specific task instructions vary depending on the experimenter's intent, but they generally instruct participants to select only those cards necessary to turn over in order to determine whether or not the given conditional statement is true.

Consider the information gained from turning over a single card with v on the visible side and finding u on unseen side: (v,u take on values in $\{p, q, \neg p, \neg q\}$, subject to the constraints of the selection task):

$$I_{KL}(S \mid d) = I_{KL}(S \mid v,u) = P(S \mid v,u) \log\left(\frac{P(S \mid v,u)}{P(S)}\right) + P(T-S \mid v,u) \log\left(\frac{P(T-S \mid v,u)}{P(T-S)}\right)$$

$$P(S \mid v,u) = \sum_{s \in S} P(s \mid v,u) = \sum_{s \in S} \frac{P(s \mid v)P(u \mid v,s)}{P(u \mid v)}$$

Since it is generally obvious that the cards in the Wason selection task were not randomly sampled, but rather one card of each possible side (p, q, $\neg p$, or $\neg q$) was presented, no information can be gained from learning that the visible side of the card is v, thus $P(s \mid v) = P(s)$.

One more step is necessary for predicting card selection: summing over all possible values of the unseen side of the card to obtain the expected information gain from turning the card with v on the visible side, $EI_g(S,v)$:

$$EI_g(S,v) = \sum_u I_{KL}(S \mid v,u) P(u \mid v)$$

The IG approach proposes that subjects select cards in the Wason selection task as a function of expected information gain, with selection favoring cards with higher expected information gain.

Applying the IG approach to causal hypotheses

The Bayesian framework presented thus far is similar to O&C (1996), except that it treats conditional statements as asserting the validity of a set of hypotheses rather than a single hypothesis. We now turn to the major differences between our account and that of O&C:

(1) The hypotheses in our framework are assertions about causal structure rather than conditional dependency; hence, a causal framework predicts different values for information gain than do O&C.
(2) We propose that mapping conditional statements onto hypotheses about causal structure is inherently ambiguous and depends on pragmatic considerations.

Here we will discuss the implications of (1), leaving the implications of (2) for the next section.

The information gained from turning over card *v* and finding *u* on the other side depends on the hypotheses under consideration; in particular, $D_{KL}(S \mid v,u)$ depends on $P(v \mid u,h)$ for every $h \in T$, which in turn depends on the content of each hypothesis *h*. In the O&C (1994) approach, *H* is the hypothesis that *q* depends deterministically on *p*, while $\neg H$ is the hypothesis that *p* and *q* are independent, and these are the only two hypotheses considered. Thus,

$$P(q \mid p,H)=1; P(q \mid \neg p,H)=b; P(p \mid q,H)=a/b; P(p \mid \neg q,H)=0$$
$$P(q \mid p,\neg H)=P(q \mid \neg p,\neg H)=b; P(p \mid q,\neg H)=P(p \mid \neg q,\neg H)=a$$

where the parameters $a=P(p)$ and $b=P(q \mid \neg p)$ are the same for *H* and $\neg H$. Other possible hypotheses are proposed by O&C but not developed, specifically those in which *q* depends probabilistically on *p*, such that $P(q \mid p,H)<1$.

In our approach, the hypothesis space *T* consists of causal structures. The conditional statement asserts that a particular causal relationship holds between *p* and *q*, thus the true causal structure is in the set *S* of structures for which this relationship holds ($S \subset T$). For a given causal structure, *h*, $P(u \mid v,h)$ can be derived using the formalism of causal Bayes nets (Pearl, 2000). For the subsequent presentation we will work with a simple causal structure that provides a reasonable approximation to many of the causal structures asserted by common conditional statements. In this structure, a cause (*C*) generates an effect (*E*), but there are conditions (*D*) that can disable the mechanism, and there are alternative causes (*A*) of the effect (see Figure 1). *D* represents all disabling conditions aggregated together, and *A* represents all alternative causes aggregated together. The arrow coming from *D* in Figure 1 indicates that the presence of *D* blocks the causal path from *C* to *E*. This structure is the causal model behind Cheng's power-pc theory (Cheng, 1997) (where $P(\neg D)$ is equal to the causal power of *C* to generate *E*); the model can also be expressed as a noisy-or Bayes net (Glymour & Cheng, 1998). For this simplified structure, the total hypothesis space *T* contains all structures with one or more of the links shown in Figure 1 (subject to the constraint that the link from *D* cannot exist without the link from *C* to *E*).

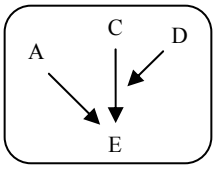


Figure 1: Noisy-or causal model

As an example of the model of Figure 1, consider a dropped pot (*C*) causing a clang (*E*). This can be disabled by various things (*D*), such as someone catching the pot or a pillow breaking the pot's fall. There are also alternate causes (*A*) of clangs, such as falling silverware.

Next we will use the semantics of the noisy-or Bayes net to derive $P(u \mid v,h)$, for the case where *h* is the hypothesis that the model of Figure 1 holds. This derivation works for all cases in which the cards in the selection task contain *C* on one side and *E* on the other, as is the case in our example "if a pot is dropped then a sound is heard" (here *p* is *C* ("a pot is dropped") and *q* is *E* ("a sound is heard"), hence *v, u* take on values in {*C, E,* $\neg C$, $\neg E$}):

$$P(E \mid C,h)=P(\neg D \vee A \mid h)=P(\neg D)+P(D)P(A)$$
$$P(E \mid \neg C,h)=P(A)$$
$$P(C \mid E,h)=\frac{P(C)P(E \mid C,h)}{P(E \mid h)}=\frac{P(C)\big(P(\neg D)+P(D)P(A)\big)}{P(A)+P(C)P(\neg D)P(\neg A)}$$
$$P(C \mid \neg E,h)=\frac{P(C)P(\neg E \mid C,h)}{P(\neg E \mid h)}=\frac{P(C)P(D)P(\neg A)}{P(\neg A)\big(P(\neg C)+P(C)P(D)\big)}$$

where the prior probabilities of *C* and *A*, *P*(*C*) and *P*(*A*), correspond to *a* and *b* in the O&C model, and the prior probability of *D*, *P*(*D*), is 1-*P*(*E*|*C*). (*P*(*D*) is taken to be zero in O&C's model for the case where *p* is *C* and *q* is *E*.) We do not require that the parameter values be the same across hypotheses, eliminating some objections to O&C's model. One could, for example, interpret a statement to be asserting that alternate causes are rare, hence *S* is all structures with P(*A*)<0.1. For simplicity, however, we will discuss only those interpretations in which a structural claim is being made (such as, a link exists from *A* to *E*); for these cases, the parameters will be the same across hypotheses.

Interpreting conditionals as causal assertions

Conditional statements are inherently ambiguous. Those for which *p* could be a cause of *q* we will call "forward" conditionals. They generally assert that *p causes q*, but the exact causal structure being asserted depends on pragmatics. For example, the statement "if a pot is dropped then it makes a clang" could have several different meanings, as demonstrated by the following hypothetical exchanges:

(1) A: "What sound will be made if I drop this pot?"
B: "If a pot is dropped then it makes a clang."
Meaning: dropped pots cause clangs
Causal Assertion: ***dropped pots*** *can cause* ***clangs***
Hypothesis Space: all structures in which ***dropped pot*** *is the cause and a* ***sound*** *is the effect*

(2) A: "I think a pot just fell."
B: "That's impossible; I didn't hear a clang. If a pot falls then it makes a clang."
Meaning: dropped pots always cause clangs
Causal Assertion: no D exists to block the path from ***dropped pots*** *to* ***clangs***
Hypothesis Space: all structures in which ***dropped pot*** *is a cause of* ***clangs***

In contrast, conditionals for which *q* could causally influence *p* we call "reverse" conditionals. They generally assert that *q is the only cause of p*, but again the exact causal structure being asserted depends on pragmatics. For example, the statement "if you hear a clang then a pot was dropped" could have several different meanings, as

demonstrated by the following hypothetical exchanges:

(1) A: "What are those sounds coming from the kitchen?"
B: "Those are items being dropped. For instance, if you hear a clang then a pot was dropped."
Meaning: falling pots are the primary cause of clangs, but not necessarily the only possible cause.
Causal Assertion: ***dropped pots*** *can cause* ***clangs***
Hypothesis Space: all structures in which ***dropped pot*** *is the cause and a* ***sound*** *is the effect.*

(2) A: "I heard a clang. What do you think happened?"
B: "It must have been a dropped pot. If you hear a clang then a pot was dropped."
Meaning: the only cause of a clang is a dropped pot.
Causal Assertion: no alternative cause A exists that can cause ***clangs****.*
Hypothesis Space: all structures in which ***dropped pot*** *is the cause and* ***clang*** *is the effect.*

Predicting card selection

The key point of distinction between our model and that of O&C is in predicting information gain, because EI_g is a simple function of information gain. $I_{KL}(S|v,u)$ depends on the particular set of causal structures in the hypothesis space T, as well as the set of asserted hypotheses S, and the parameters $P(C)$, $P(A)$, and $P(D)$. S in turn depends on pragmatic considerations. In Figures 2 and 3 we give qualitative predictions for S, T, and I_{KL} for common types of conditionals, which will motivate future experiments.

Some generalizations are worth noting: (1) $EI_g(p)$ is often high. (2) for rare p and q, when structures with no C to E link have sufficient priors (e.g., a uniform prior), $EI_g(S,q)$ is often high; (3) if the conditional assumes the C to E link exists, then structures with no C to E link will have low priors and $EI_g(S,q)$ will be low. In general, with pragmatic considerations, the model predicts selection of p and q cards if the conditional *asserts* that C causes E, and selection of p, $\neg q$ if the conditional *assumes* that C causes E.

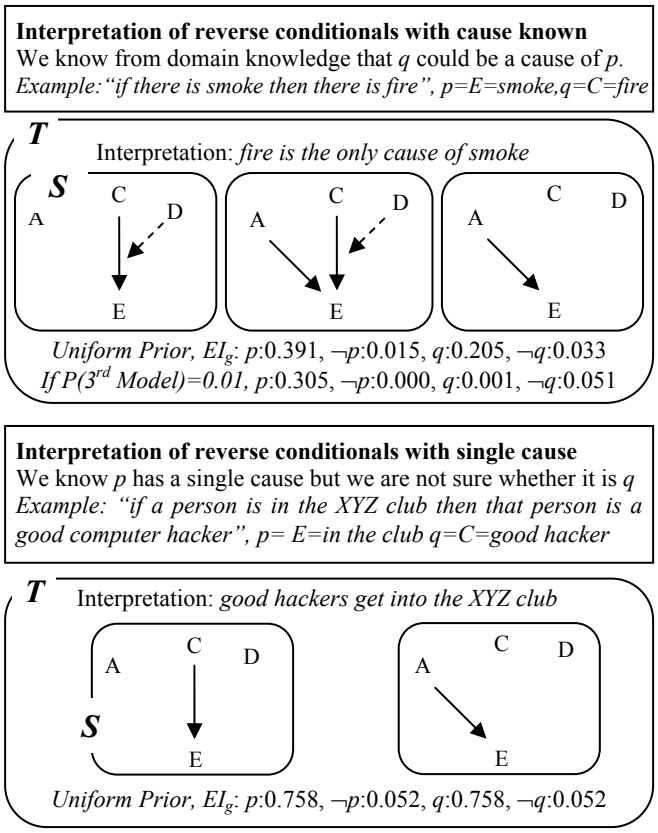


Figure 2: Predictions for reverse conditionals
Dotted arrows indicate a mixture of two hypotheses, one in which the arrow is present, and one in which the arrow is absent
All EI_g values assume $P(C)=P(A)=P(D)=0.1$

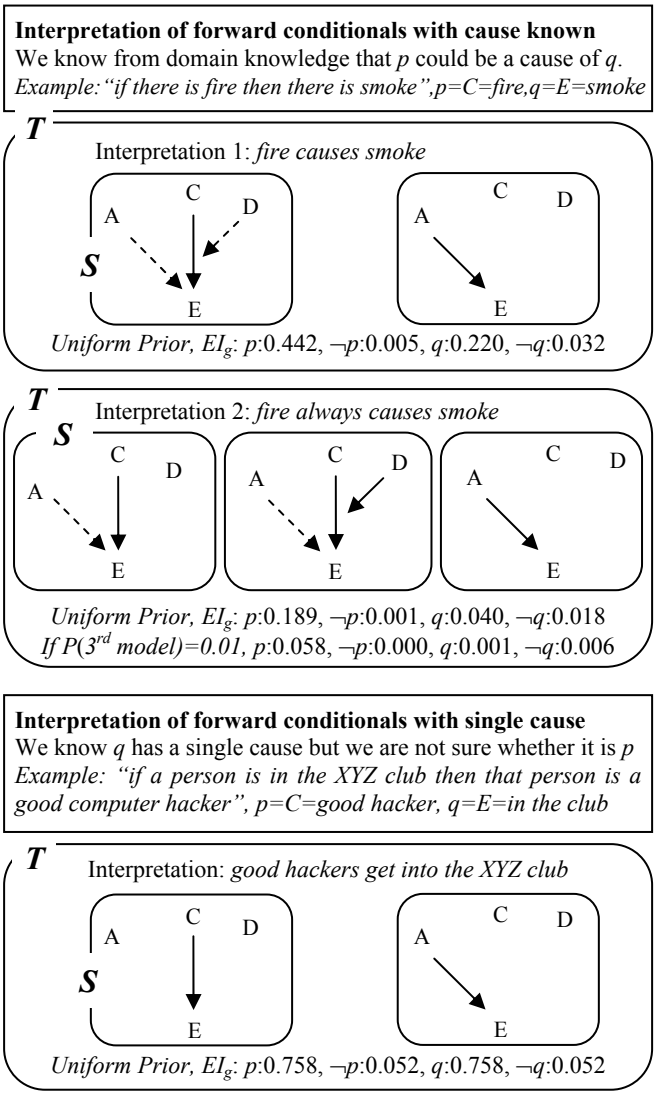


Figure 3: Predictions for forward conditionals.
Dotted arrows indicate a mixture of two hypotheses, one in which the arrow is present, and one in which the arrow is absent.
All EI_g values assume $P(C)=P(A)=P(D)=0.1$

Relation to previous analyses and phenomena

In this section we discuss how our approach accounts for previous phenomena on the selection task. We group these phenomena within a discussion of previous approaches, while highlighting the distinctive aspects of our approach.

Information-gain (Oaksford & Chater, 1994)

In many of their publications, O&C analyze a simple comparison in which H asserts complete dependency and $\neg H$ asserts a complete independency between p and q. In our model, this is identical to the assertion that T corresponds to all structures in which $p=C$, $q=E$, and D does not exist, while S corresponds to the subset of those structures in which a link exists from C to E.

With the richer hypothesis space of causal models, the information gain framework predicts some previous results on the selection task that are not predicted by O&C (see next section). O&C predict that the p and q cards should be chosen when both p and q are rare and that the p and $\neg q$ cards should be chosen when p or q is common. This is predicated on the assumptions that $I_g(p,\neg q)$ is 1, $I_g(p,q)$ is high for rare p and q, and $I_g(\neg p,q)$ and $I_g(\neg p, \neg q)$ are zero. Our analysis suggests that these assumptions are only valid if structures in T with links from C to E have sufficient priors. If these structures have very low priors, the $\neg q$ card should be more informative than the q card because $I_{KL}(S|p,q)$ and $I_{KL}(S|\neg p,q)$ will be low, hence $EI_g(S,q)$ will be low. For example, suppose one asserts that "if a clang is heard then a pot was dropped". It is reasonable to assume that dropped pots cause clangs, hence a low prior should be placed on any structure with no link from dropped pots to clangs. Thus, finding a dropped pot that clanged (p,q) will not be very informative, despite p and q being rare, but finding a dropped fork that produced a clang ($p,\neg q$) will be very informative, hence p and $\neg q$ should be chosen.

Almor & Sloman (1996) provide evidence that appears to contradict O&C (1994), in which p and q are rare, yet people choose the p and $\neg q$ cards. Some examples of their conditionals are "*if a product gets a prestigious prize then it must have a distinctive quality*", and "*if a product breaks then it must have been used under abnormal conditions*". O&C (1996) claim these results can be accounted for using their utility-theoretic analysis of deontic tasks. However, these statements would only be deontic if they were rules that people have to follow, but it is not apparent that this is the case. According to our analysis, the conditional is reversed (q causes p), leading subjects to interpret the statement as asserting the absence of alternate causes (A) while taking for granted the link from C to E, thus assigning low probability to structures without this link. For example, since there are other possible causes of a product breaking, subjects choose the $\neg q$ card (no abnormal usage) to see if p occurred (the product broke for some other reason), but there is no need to see if abnormal usage causes breakage.

A further point of differentiation is that our causal framework predicts the $\neg p$ card should be chosen in certain cases (when $I_{KL}(\neg p,q)$ and $P(q|\neg p)$ are both high).

Social Contracts and Precautions (Cosmides, 1989)

People have been found to provide high levels of logically correct responses to Wason selection tasks about social contracts (Cosmides, 1989; Fiddick, Cosmides, & Tooby, 2000). A debate has emerged over whether this is evidence for a specialized social reasoning engine. Social contracts do indeed seem to be special, but do people reason about them differently than other tasks? In our framework, what makes them special is that they all have a consistent causal structure, in which people follow rules to ensure that C produces E reliably. For example, in the social contract "if you pay $10 then you get a watch", the rules compel the seller to give the buyer a watch (E) once $10 is paid ($C$). When the link from C to E is assumed to exist, as is the case in most social contract tasks, one should assign a prior of zero to any hypothesis in T with this link missing. Since all the hypotheses with non-zero prior then contain links from C to E, our causal analysis predicts that only $I_{KL}(p,\neg q)$ is high, and hence only the p and $\neg q$ cards should be selected.

Precaution tasks (Fiddick et. al, 2000) have essentially the same structure as social contracts: they assume that the precaution is in force (i.e., the link from C to E exists), and ask subjects to determine whether the rule is being followed by everyone. Our analysis suggests that if instead the rule itself is questioned (i.e., is the rule in force?), people will interpret S as asserting that there is a link from C to E; since this link questioned, one should assign a non-zero prior to the structures in T in which this link is missing, making the q card useful (if p and q are rare) because $I_{KL}(p,q)$ is high.

The results of Fiddick et. al (2000) show that this is exactly what people do. Fiddick et. al (2000) published precaution experiments that show people choose q more than $\neg q$ in "standard" versions of precaution studies such as "*if you go hunting then you wear [orange] jackets to avoid being shot*". In the "standard" version, subjects are instructed to see if it is true that the jackets are for hunting, whereas in the "precaution" version they are instructed to see if any people are endangering themselves. This result confirms that when testing whether a social contract or precaution is in force, people will test the assertion that a link exists from C to E, and hence will choose the p and q cards (provided p and q are rare).

O&C (1994) propose a utility-theoretic account for how people make choices in social reasoning tasks. This is appropriate for tasks in which the participant is told that catching rule violators is important (i.e., has high utility). If, however, the participant is being asked simply to determine whether or not the rule is being violated, the assignment of utility to this information is not warranted. We avoid the difficulty of assigning utilities to information by using expected information gain as the sole basis on which to select cards. A causal analysis predicts the selection of p and $\neg q$ responses for any task in which structures without C to E links are given low priors, which should be the case in all social reasoning tasks that assume the rule is in force and ask subjects to detect violators.

Perspective Shifting

Perspective shifts (interpreting "if p then q" as "if q then p") have been explained as the result of adopting different perspectives on a rule – the enforcer vs. actor. We propose that perspective shifts occur when three conditions are met: (1) C is a known cause of E, (2) it is not obvious whether D exists, and (3) it is not obvious whether A exists. This sets

up the hypotheses in Figure 4. For example, "if you pay $10 then you get a watch" can be shifted to "if you got a watch then you paid $10" because our domain knowledge tells us that no disabling conditions exist (the buyer must get the watch once $10 are paid), and no alternate causes exist (the buyer cannot get the watch without paying).

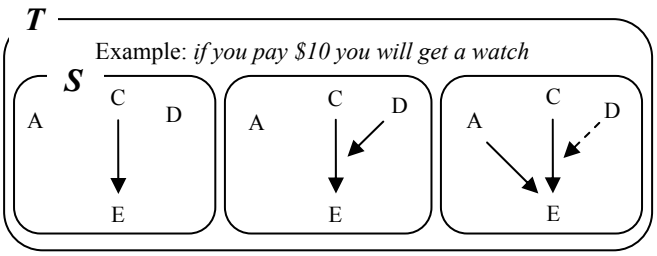


Figure 4: Perspective shifting hypotheses

On this view, perspective shifting occurs not just on deontic tasks, but in any situation in which the above three conditions are met. For example, "if water boils then it is over 100°C" could easily be interpreted to imply that "if water is over 100°C then it will boil". Perspective shifting can therefore occur, even in non-deontic situations, when the asserted structure does not contain *D* or *A* links.

Necessity and Sufficiency (Ahn & Graham, 1999)

Ahn & Graham (1999) show that most people choose the normative response if it is clear that the statement asserts either that *p* is a sufficient condition for *q* or that *q* is a necessary condition for *p*, or both. Asserting that *p* is a sufficient condition for *q* corresponds to asserting that *D* does not exist. For example, asserting that *flipping the switch* is sufficient for the *lights turning on* corresponds to asserting that nothing (*D*) can disable the switch. In contrast, asserting that *q* is a necessary condition for *p* corresponds to asserting that *A* does not exist in the causal model. For example, asserting that *flipping the switch* is necessary for *lights turning on* corresponds to asserting that nothing else (*A*) could turn on the lights. Both of these cases assume that the link from *C* to *E* exists, hence as before, *p* and $\neg q$ are the most informative cards (or *q* and $\neg p$ when a conditional with "must" is reversed to say "may"), which follows Ahn & Graham's predictions. Ahn & Graham (1999) also discuss cases in which *p* is asserted to be both necessary and sufficient for *q*, in which cases subjects choose all 4 cards. This corresponds to asserting that neither *A* nor *D* exist, and $I_{KL}(S|\neg p,q)$, $I_{KL}(S|p,\neg q)$ are both high.

An open question in Ahn & Graham's (1999) theory is how people know whether *p* is necessary or sufficient for *q* in cases when it is not explicitly stated. A cause can be necessary or sufficient for an effect, but it does not make sense to say that an effect is necessary or sufficient for a cause (e.g., a clang could not be necessary or sufficient for a pot to drop, because dropped pots precede clangs). Because of this causal asymmetry, some amount of causal reasoning must precede determination of necessity and sufficiency relationships. Furthermore, determining necessity or sufficiency can be done using just causal knowledge, as *p* is necessary for *q* if *A* does not exist, and *p* is sufficient for *q* if *D* does not exist.

Conclusion

Causal reasoning underlies many of our intuitive judgments in everyday life, and the results we present here demonstrate that causal structure plays an important role in a domain of reasoning previously thought to be governed by logic and probability. Our approach predicts a number of effects on the selection task that do not follow naturally from previous approaches. If used appropriately, the selection task is an excellent tool for testing people's abilities to gather evidence and become more informed about their world. Since knowing the causal structure of the world is of great value for making predictions in every life, it is perhaps not surprising that the cards people naturally select tend to be those that maximize the amount of knowledge that can be obtained about causal structure from a single observation.

References

Ahn, W., & Graham, L. M. (1999). The impact of necessity and sufficiency on information choices in the Wason four-card selection task. *Psychological Science*, 10, 237-242

Ahn, W. (1999). Effect of Causal Structure on Category Construction. *Memory & Cognition*, 27, 1008-1023

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, 54, 299-352.

Almor, A. & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review*, 103(2): 374-380

Cosmides, L. 1989. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition,* 31, 187-276

Fiddick, L., Cosmides, L., Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition*, 77, 1-79

Glymour, C. & Cheng, P. (1998). Causal mechanism and probability: a normative approach. In *Rational models of cognition,* Oaksford, M. & Chater, N., Oxford University Press.

Krynski, T.R., Tenenbaum, J. B.. (2003). The Role of Causal Models in Reasoning Under Uncertainty. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Oaksford, M. and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review* 101, 608-631

Oaksford, M. & Chater, N. (1996). Rational Explanation of the Selection Task. *Psychological Review*, 103 (2), 381-391

Over, D. & Jessop, A. (1998). Rational analysis of causal conditionals and the selection task. In *Rational models of cognition,* Oaksford, M. & Chater, N., Oxford University Press.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press

Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science,* 27, 709-748

Wason, P.C. (1996). Reasoning. In B.M. Foss (Ed.), *New Horizons in Psychology*, (pp. 135-151). Harmondsworth, Middlesex, England: Penguin.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 53-76.

Acknowledgements We thank Brigid Dwyer, Sarah Newton, and Suzanne Luther for their enthusiastic assistance. JBT was supported by the Paul E. Newton chair. TRK was supported by a graduate fellowship from the National Science Foundation.

On the Representation of Physical Quantities in Natural Language Text

Sven E. Kuehne (skuehne@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
1890 Maple Avenue, Ste. 300
Evanston, IL 60201, USA

Abstract

In this paper we investigate the forms in which quantity information can appear in written natural language. Our focus is on physical quantities found in descriptions of physical processes, such as expansion, movement, or transfer. Using Qualitative Process Theory as our underlying formalism, we show how information extracted from natural language text corresponds to the five constituents of physical quantities. The results of this analysis can be used for the creation of interpretation rules and extraction patterns in NL systems.

Introduction

Ordinary people know a lot about the physical world around them. They know that water will eventually boil if you heat it on a stove, that a ball placed at the top of a steep ramp will roll down, and that a cup will overflow if you continue to pour coffee in it. When people talk and write about such phenomena in everyday language, references to continuous properties are usually part of these descriptions. From simple utterances like *"The coffee is hot"* to a more complicated comparison like *"The velocity of gas molecules is higher than the velocity of molecules in a liquid."* being able to identify and extract the information about physical quantities is essential to understand these sentences. Using Qualitative Process Theory (Forbus, 1984) as the underlying formalism, we investigate the forms in which continuous properties can appear in written natural language. Our focus is on *physical* quantities found in descriptions of *physical* processes, such as expansion, movement, or transfer.[1]

The way in which continuous parameters and processes are described in natural language is not accidental. Since Qualitative Process Theory is a formalism of how people reason about the physical world, the basic ideas of the theory should be reflected in the language that people use to communicate their understanding of physical phenomena. This paper shows that the natural language descriptions of physical processes contain abundant information about the constituents of physical quantities. Moreover, the results of this study can be used in a variety of applications, such as grammatical rules of a parser or in the design of information extraction algorithms.[2]

Physical quantities

In Qualitative Process Theory, all physical changes in *continuous properties* are caused by physical processes. The identification of continuous parameters is therefore an essential step in the extraction of information about physical processes from natural language text. In an earlier analysis (Kuehne & Forbus, 2002) we presented a scheme for the extraction process that uses FrameNet-compatible representations (Baker, Fillmore, & Lowe, 1998; Fillmore, Wooters, & Baker, 2001) to capture information about physical processes. The examples presented draw from the same corpus material (Buckley, 1979; Maton et al., 1994; Moran & Morgan, 1994) used in our previous analysis. Our goal here is to show how information about continuous parameters can appear in natural language, and the ways in which this information corresponds to the following five constituents of physical quantities:

- The *Entity* is a uniquely named object or an instance of a process associated with the quantity. For example, the word 'brick' in the noun phrase 'the temperature of the brick' denotes an entity.[3]
- The *Quantity Type* specifies the kind of parameter. The word 'temperature' in the noun phrase 'the temperature of the brick' is a reference to a quantity type.
- The *Value* specifies the numerical or symbolic value of the property. The number '3' in '3 liters of water' or

[1] The findings of this analysis are applicable to other types of quantities as well. The framework of QP theory determines just determines kind of information we are interested in, i.e. constituents of a physical quantity. Abstract and conceptual quantities are often referred to metaphorically by words with a physical basis and require a different semantic interpretation. 'The price is hot.' is does not have anything to do with temperature, unlike 'The water is hot.' However, the techniques for the extraction of information about such quantities are essentially the same.

[2] Although we use the results of the analysis for exactly these purposes, the findings are presented in a general way and not limited to any particular type of grammar or pattern language.

[3] The noun 'brick' actually refers a particular individual, maybe 'brick32', not the collection of all bricks.

the adjective 'hot' in 'the hot ground' are values associated with a quantity.
- The *Unit* specifies the physical units of the property. Example: The word 'kilograms' in '3 kilograms of lead.' Units usually appear in combination with a numerical value or with a quantifier.
- The *Sign of the Derivative* specifies how the parameter is changing. In the sentence "The temperature is increasing." the sign of the derivative is expressed by the word 'increasing', which indicates that the parameter is changing in a positive direction.

Only the first two of these five constituents are required to identify a physical quantity. The quantity type together with the entity are sufficient to talk about quantities like 'the temperature of a brick' or the 'the flow rate of heat'. Values, units, and information about changes are optional and often not explicitly stated.

Entities and quantities types

We begin with a look at the forms commonly used in natural language descriptions to express information about the two required constituents of physical quantities, the entity and the associated quantity type. The remaining three constituents, i.e. values, units, and changes, will be discussed in the subsequent sections.

Explicitly referenced quantities

Natural language text can refer to physical quantities either directly or indirectly, depending on whether the type of the quantity is explicitly mentioned in the sentence. *Explicit references* can be found in nouns, verbs, and adjectives that are morphologically related to quantity types.

Nouns

The quantity type can be explicitly mentioned as a noun, together with one or more entities that it is associated with.

(1) VOLUME flows from the *can* to the *ground*.
(2) The TEMPERATURE of the *brick* is rising.

Sentence 1 contains information about two physical quantities, the volume of some substance in the can and on the ground. The quantity type 'volume' is associated with both locations, i.e. the 'can' and the 'ground'. In (2) the quantity type 'temperature' is associated with a single entity.

The quantity type can also appear as the head of a compound noun. The remaining constituents of the compound noun can be treated as information about a specialization of the quantity type. For example, in (3) the quantity type 'radiation heat' is a specialization of 'heat'; in (4) 'heat energy' is a type of 'energy'.

(3) RADIATION HEAT flows from the *heater*.
(4) The HEAT ENERGY of the *water* increases.

Verbs

Verbs can refer to physical events as well as to quantity types associated with these events.[4] The verb in sentence 5 appears as a direct reference to the quantity type 'length'. Sentence 6 is slightly more complicated, because it allows two different interpretations. The obvious interpretation is to treat the verb as an explicit reference to a quantity, as it is in (5). In this case, the quantity type 'heat' is tied to both entities, the stove as the source of the heat flow and the kettle as the destination of the heat flow.

(5) The press LENGTHENS the *iron beam*.
(6) The stove HEATS the *kettle*.

Alternatively, (6) can be interpreted as an increase in temperature of the kettle caused by the stove. Even though the quantity type 'temperature' is not mentioned in the sentence, we might infer that heating the kettle also increases the temperature of the kettle. This is an inference that most readers of such a descriptions will readily draw, and it coincides with the kind of conclusions that are supported by QP Theory.

Adjectives

Certain adjectives can refer to quantity types directly, if the adjective is morphologically related to a quantity type. For example, in (7) the adjective 'denser' refers to the quantity type 'density'. The quantity type in this sentence is associated with both entities, i.e. 'iron' and 'wood'.

(7) *Iron* is DENSER than *wood*.

Implicitly referenced quantities

While the quantity types in explicitly referenced quantities are usually easy to identify and extract, *implicit references* to quantities are more difficult to figure out. Implicitly referenced quantities do not mention a quantity type. Instead, the reader has to use the contextual information provided by the sentence as well as available background knowledge. The following examples show how nouns, verbs, adjectives, and adverbs can determine a quantity that is not explicitly mentioned in a sentence.

Verbs

A quantity type can be implicitly referenced by a verb that describes a physical process, e.g. movement, expansion, or transfer. The sentence in which the verb occurs usually

[4] Events such as the increase or decrease of a parameter, e.g. the temperature of a brick, can be involved in an instance of a physical process. For one linguistic perspective on actions, processes, and events, see (Parsons, 1990).

provides additional contextual information for the interpretation of the implicitly referenced quantity.

(8) As the temperature rises, the *liquid* EXPANDS.

The verb 'expand' in (8) indicates that something is changing in different physical dimensions, i.e. in length, area, or volume. For the three-dimensional entity 'liquid' the appropriate quantity type is therefore 'volume'. The verb also includes implicit information about a positive change in the quantity, i.e. an increase in volume of the liquid, which we will address later.

Adjectives

The quantity type can be implicitly referenced by certain adjectives. For example, the quantity type described by the adjective 'hot' in (9) is 'temperature'. The comparative also encodes the ordinal relationship between the quantities associated with the two entities, i.e. the fact that the temperature of the stone is greater than the temperature of the water. Similarly, the quantity type expressed by 'lighter' in (10) is 'weight'.

(9) The *stone* is HOTTER than the *water*.
(10) The *upper air masses* are LIGHTER than the *lower air masses*.

For a correct interpretation the relationship between the adjective and the associated quantity type has to be known. The fact that the adjective 'hot' is associated with 'temperature' is a fact learned by a human reader and is typically provided as background knowledge in NL systems.

Verb/Adverb combination

Quantity types can also be determined by combining verbs and adverbs. The quantity type referenced in (11) is the rate of movement, or 'velocity'. The adverb alone is not sufficient to determine the quantity type. Although 'faster' is generally associated with velocity, it just qualifies the rate of change, i.e. that something is happening in less time. There are cases in which the quantity type referenced by 'faster' is not velocity. For example, 'expanding faster' in (12) refers to the rate of expansion.

(11) The *gas molecules* are MOVING FASTER than *molecules in a solid*.
(12) *Liquid A* is EXPANDING FASTER than *liquid B*.

All these cases have one thing in common: the referenced quantity is a rate, most likely associated with a process referenced by the verb ('movement', 'expansion', 'decay').

Noun/Verb combination

This type of implicitly referenced quantity uses a noun/verb combination to refer to the rate of change of a quantity.

(13) The less heat is supplied, the slower the *temperature* RISES.

The quantity type in (13) is not 'temperature' but the rate of change in temperature, resulting from a change in the amount of heat. The combination of 'rises' and 'temperature' determines the quantity type, while the combination of the verb 'rises' and the adverb 'slower' gives the direction of change.

Noun/Adjective combination

The quantity type is only implicitly referenced by a combination of a noun and an adjective.

(14) The BIGGER the *surface* [is], the more heat is absorbed.

The quantity type in (14) is the size of the surface (not the surface itself) associated with an unnamed participant or the size of a participant 'surface'. The adjective 'bigger' refers to the quantity type 'size' (or 'area'). Since 'big' can also refer to the quantity type 'volume', the dimensionality of the entity determines the appropriate quantity type in this case.

Representation of values in physical quantities

Knowing the type of a quantity and the entity it is associated with enables us to talk and reason about it. A simple noun phrase such as 'the depth of the water' contains enough information recognize it as a physical quantity, even without having any information about a particular value the quantity might have, the unit of that value, or the direction in which the quantity is changing. The following two sections examine how values and units of quantities appear in natural language text, and how changes in quantities can be identified.

There are three common types of references to values and units that can be found in natural language text: in the context of comparisons, as symbolic labels, and as quantitative information. We will discuss values and units together because units usually appear in combination with values.[5]

[5] Units can appear separately from values in definitional statements, like "Length is measured in Meters." or, even more explicit, "The unit of power is the Watt."

Comparison

Values in the context of a *comparison* appear in sentences like "The brick is warmer than the plate." The comparison orders the quantities, i.e. the temperature of the brick is greater than the temperature of the plate. However, it does not contain exact information about the possible values of the quantities. Even though the comparative 'warmer' might refer to a specific range of temperature, the exact values cannot be known or even guessed from the information provided by the sentence. The brick might be red hot, while the plate is frosted with ice.

It is impossible to determine how far the values associated with the two compared quantities are apart from each other. The only information that can be extracted from this sentence is the fact that the value of one quantity is greater than the other. With several of these comparisons along the same dimension, it is possible to identify the potential ranges of the values for particular quantities. For example, the temperature of a coffee is greater than the temperature of an ice cube, and it is lower than the temperature at the tip of a lit cigarette.

Symbolic labels

Values can also take the form of a *symbolic label* associated with an entity, e.g. "The brick is hot." Even though the exact temperature of the brick is unknown, the adjective 'hot' suggests a certain temperature range. The range might be different depending on the context of the sentence. In refrigeration 'hot' might be in a very different range of temperatures than in the context of metallurgy.

Nouns that are associated with the adjective can impose restrictions on the range of the value in certain cases. For example, (Bierwisch, 1967) compares two simple sentences, "*The room is tall*." and "*The space is tall"*. In the first sentence the noun 'room' might restrict the average range of values for the height to those for a typical room, e.g. between 8 and 10 feet. Without further information, this kind of assumption is more difficult to make for the second sentence. Is the space a small compartment or a crawl space? Or is it the inside of a cathedral? The range of typical values would be quite different for these two cases.

Adjectives that represent a value are generally quantity-specific, i.e. they refer to a particular type of quantity as in the sentence "The brick is hot." Alternatively, a quantity-neutral form could be used to express the same fact, e.g. "The temperature of the brick is high."[6]

While adjectives and adverbs such as 'hot' or 'slow' generally refer to a range of values along a dimension, natural language also uses symbolic labels to refer to concrete values, i.e. particular points along a dimension. The noun phrase 'boiling point of water' usually refers to the point where liquid water turns into steam and the value of approximately 212 degrees Fahrenheit. The noun phrase provides a label for this particular point.[7]

The structure for labels that describe limit points is not arbitrary. Usually the head of the noun phrase refers to a point on a scale (e.g. 'point', 'barrier'), while the noun modifier is associated with a process, a dimension, or a quantity type (i.e. 'boiling', 'sound'). These two parts are mandatory components of the label. Determining the quantity type and the dimension is difficult in many cases, e.g. we have to know that 'boiling point' is associated with 'temperature' and that 'sound barrier' actually refers to the speed of sound or velocity. Additionally, the label can take an optional complement phrase that restricts the compound noun. For example, the complement phrase 'of water' restricts the interpretation of boiling point to a particular substance. The key idea here is that the underlying mechanisms for handling limit points are essentially the same as those for symbolic references to intervals on a particular dimension.

Concrete numeric values and units

The most explicit form in which values can appear is *quantitative information*, i.e. by using concrete numeric information and units. For example, in (15) the quantity type ('temperature') is explicitly stated, together with detailed information about the numeric value ('120') and the unit ('degrees Fahrenheit').

(15) The temperature of the brick is *120 degrees Fahrenheit*.

Sentences that contain concrete numeric values and units typically do not use quantity-specific adjectives or adverbs instead of explicit references to the quantity type.

(16) *The water is 80 degrees Celsius *hot*.
(17) The water has 80 degrees Celsius.

Sentence 16 should be considered anomalous, because the adjective 'hot' provides at best redundant information in the form of a symbolic value. Units can refer indirectly to the quantity that they are associated with, as shown in (17). The association between units and quantity types (degrees Celsius as a unit for temperature) is a learned fact and has to be encoded as background knowledge.

Representations of changes in physical quantities

The values of physical quantities cannot always be treated as static information, because they can change while

6 The Cyc knowledge base (Lenat & Guha, 1989) handles values in a similar way. For example, the value `#$Hot` is the result of `#$HighAmountFn` of `#$Temperature`.

7 Note that the compound noun 'boiling point' would be an underspecified symbolic label because different substances have different boiling points. Other labels such as 'sound barrier' may not need the additional complement.

physical processes are active. The sign of the derivative indicates whether a quantity is changing and in which direction.[8]

The most obvious choice to express changes in the physical world is the use of verbs. For example, if water is flowing from one container into another, there are several ways of expressing the change of the amount of water in each container.

(18) The amount of water in container A is decreasing, while the amount of water in container B is increasing.
(19) Water flows from container A to container B.

Although (18) and (19) might be describe the same scenario, they are not equivalent. For example, (19) only implies a decrease of the amount of water in location A. It does not state this information explicitly. On the other hand, (18) implies a flow, without actually mentioning it. These distinctions are important for a semantic interpretation process, because the information that is directly available from the sentences is different.

Verbs with direct references to a quantity change

Verbs can directly refer to a change in a quantity and its direction, i.e. whether the quantity is increasing or decreasing, when the verb alone contains all the information about the change and the direction and we can therefore distinguish between verbs for positive and negative changes in quantities. For example, *gain, increase,* and *add* are verbs for positive changes, while *lose, decrease,* and *leak* are associated with negative changes.[9] Some verbs belonging to this class also allow prepositional phrase as a complement, which is restricted to the particular direction of change indicated by the verb itself (e.g. 'add to' vs. *'add from').

(20) The *brick* LOSES *heat* to the *room*.
(21) The *temperature of the water* is INCREASING.
(22) The *brick* GIVES OFF *heat*.

Some otherwise 'neutral' verbs can also fall into this class if they use specific particles to indicate a change in a quantity, as in (22).[10]

Verbs with directional prepositional phrases

Verbs associated with Transfer and Motion event do not contain a direct reference to changes in quantity. For example, verbs like *flow* or *move* indicate a transfer of something between two physical or conceptual locations, but they do not contain information about the actual direction of the change. Instead, this information is provided by directional prepositional phrases attached to the verb. The description of the transfer can be complete when both the source and the destination are identified by prepositional phrases, as in (23), or partial when only one of the directional prepositional phrases is attached, as in (24) and (25).

(23) Heat is *transferred* FROM inside the house TO the outdoors.
(24) Energy is *moved* TO a new location.
(25) The *fan* moves heat away FROM the processor.

Verbs in combination with quantity-specific adverbs

Quantity-specific adverbs can determine the change in a quantity in conjunction with a verb. Analogous to verbs with direct reference to a quantity change, the combination of verbs and quantity-specific adverbs can be associated with a decrease in a quantity, as in (26) or with an increase, as in (27).

Similar to the interpretation of the quantity type from verb/adverb combinations, there are cases in which the same adverb can refer to an increase (or a decrease) of a particular quantity type, depending on the verb with which it is used. For example, in the context of (27), the adverb 'faster' would indicate a positive change in the velocity of the molecules, while in (28) it will indicate an increase in the rate at which a substance dissolves.

(26) The *glass* is COOLING FASTER.
(27) The *molecules* are MOVING FASTER.
(28) The *substance* DISSOLVES FASTER.

Nouns with direct references to change

Nouns provide another way of describing changes in physical quantities. They can be divided into similar classes as verbs, i.e. nouns with direct references to a change in a quantity, and nouns that use directional prepositional phrases.

Nouns can directly refer to a change in a quantity, and analogous to verbs they can be divided into nouns that refer to positive, as in (29), and negative changes, as in (30).

(29) The INCREASE in *temperature* is significant.
(30) The DECREASE in *pressure* caused a failure.

Nouns with directional prepositional phrases

Similar to verbs of the Transfer and Motion domain, the corresponding nouns will also need directional prepositional phrases to describe changes in a quantity.

[8] Information about changes in quantities can support other aspects of QP theory, e.g. in determining relationships between continuous parameters such as direct and indirect influences.
[9] Another distinction could be made between verbs that can only used with extensive quantities. For example, heat can be *added*, while temperature cannot.
[10] The particle has to agree with the complement structure of verb. For example, the verb phrase *'gives in' cannot take 'heat' as its argument.

Again, the information about the transfer can be complete, as in (31) or partial as in (32).

(31) The *flow* of oxygen FROM the tank TO the capsule is blocked.
(32) The *transfer* of heat TO the kettle has been completed.

Discussion

Parts of our current research are concerned with the design of a controlled language for describing physical phenomena. One important aspect in the development of such a language is the goal to reduce possible syntactic and semantic ambiguity. The identification of patterns used for references to continuous parameters in natural language is an essential part of the semantic interpretation process, which must include the detection of directly referenced quantities as well as indirect references.

Research on the lexical semantics of adjectives has tried to establish taxonomies for the different semantic categories of adjectives (see Raskin & Nirenburg (1995) for an overview). Several of these taxonomies focus on the class of adjectives that we are most interested in for extracting information about physical quantities, i.e. qualitative (scalar, gradable) adjectives (Dixon, 1991; Frawley, 1992). From our perspective, using the semantics of Qualitative Process Theory, the taxonomies suggested by Dixon and Frawley are inconsistent. The breakup of types and subtypes appears to be arbitrary, because several of the types of quantities can be collapsed into a single type. In Dixon's taxonomy the adjectives of the 'speed' and 'physical property' types are separated from those classified as 'dimension'. Similarly, 'age' and 'value' are listed as separate types.

Many quantity-specific adjectives and adverbs form opposing pairs for the same quantity type along a single dimension. For example, 'tall' is the opposite of 'short' for the quantity type 'height', and 'wide' the opposite of 'narrow' for the quantity type 'width' (see Bierwisch (1967, 1989) and Kennedy (2001) for a detailed analysis of polar adjectives). For certain quantity types we can identify not just a single opposing pair but a set of quantity-specific adjectives. For the quantity type 'temperature' we can find adjectives such as 'warm', 'cool', 'tepid', and variations such as 'lukewarm' as references besides just 'hot' and 'cold'. It is an interesting question to speculate why this variety of quantity-specific adjectives exists for some quantity types but not for others. Frequent use or familiarity with the concept 'temperature' cannot explain this fact alone.

Understanding the connections between Qualitative Process Theory and natural language is important for understanding the general cognitive plausibility of qualitative models. It will also give us greater insight into how results from qualitative reasoning can be communicated back to human users in an intuitive way – by using natural language.

Acknowledgements

I would like to thank Ken Forbus and Dedre Gentner for insightful comments on this paper, as well as Praveen Paritosh and Chris Kennedy for interesting discussions on the topic. This research was supported by the Artificial Intelligence program of the Office of Naval Research.

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). *The Berkeley FrameNet Project.* In: Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 98), Montreal, Canada.

Bierwisch, M. (1967). Some semantic universals of German adjectivals. *Foundations of Language, 3*, 1-36.

Bierwisch, M. (1989). The Semantics of Gradation. In M. Bierwisch & E. Lang (Eds.), *Dimensional Adjectives* (pp. 71-261). Berlin, Germany: Springer-Verlag.

Buckley, S. (1979). *From Sun Up to Sun Down*. New York: McGraw-Hill.

Dixon, R. M. W. (1991). *A New Approach to English Grammar, on Semantic Principles*. Oxford, England: Clarendon Press.

Fillmore, C. J., Wooters, C., & Baker, C. F. (2001). *Building a Large Lexical Databank Which Provides Deep Semantics.* In: Proceedings of the Pacific Asian Conference on Language, Information, and Computation, Hong Kong, China.

Forbus, K. D. (1984). Qualitative Process Theory. *Artificial Intelligence, 24*, 85-168.

Frawley, W. (1992). *Linguistic Semantics*. Hillsdale, NJ: Erlbaum.

Kennedy, C. (2001). Polar Opposition and the Ontology of 'Degrees'. *Linguistics and Philosophy, 24*(1), 33-70.

Kuehne, S. E., & Forbus, K. D. (2002). *Qualitative Physics as a component in natural language semantics: A progress report.* In: Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society, George Mason University, Fairfax, VA.

Lenat, D. B., & Guha, R. V. (1989). *Building large knowledge-based systems : representation and inference in the Cyc project*. Reading, MA: Addison-Wesley.

Maton, A., Hopkins, J., Johnson, S., LaHart, D., McLaughlin, C. W., Warner, M. Q., & Wright, J. D. (1994). *Heat Energy* (annotated teacher's ed.). Englewood Cliffs, NJ: Prentice Hall.

Moran, J. M., & Morgan, M. D. (1994). *Meteorology - The Atmosphere and the Science of Weather* (4th ed.). New York, NY: Macmillan College Publishing.

Parsons, T. (1990). *Events in the Semantics of English*. Cambridge, MA: MIT Press.

Raskin, V., & Nirenburg, S. (1995). *Lexical Semantics of Adjectives: A Microtheory of Adjectival Meaning* (Technical Report MCCS-95-288). Las Cruces, NM: New Mexico State University.

Learning Relational Categories by Comparison of Paired Examples

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, P.O. Box 6000
Binghamton, NY 13902 USA

Olga Boukrina (oboukri1@binghamton.edu)
Department of Psychology, P.O. Box 6000
Binghamton, NY 13902 USA

Abstract

Our central question is whether comparison of co-presented instances promotes category learning. We report results of four experiments testing acquisition of relational categories under conditions of Comparison learning versus traditional Single item learning. In order to control for frequency of exposure, the Single group received twice as many learning trials. Experiment 1 showed more accurate single-item classification at test for both old and new items by the Comparison group relative to the Single group. Experiment 2 used only within-category pairs in the Comparison condition (rather than both types of pairs), but no accuracy advantage was found. Experiment 3 repeated this design using a reduced training set and showed a learning effect of comparison and a marginal advantage in transfer to new items. In Experiment 4, a novel paradigm revealed further evidence of a facilitative effect for within-category comparison. The power of comparison to promote learning and transfer is discussed in terms of mechanisms of encoding and knowledge change.

Introduction

The present research addresses the effect of comparing co-presented instances during classification learning. Nearly all theorists propose that categorizing an instance involves some type of comparison between an instance and stored category representations. A further role for comparison in category learning is between presented instances and remembered instances. Sequential effects may occur if items presented in immediate or near succession are brought into temporal juxtaposition (e.g., Elio & Anderson, 1981). Learners may also experience remindings of previously encountered instances which can guide further processing (Spalding & Ross, 1994; Ross, Perkins & Tenpenny, 1990). Of considerable interest to our project, Ross and Spalding (2000) report that reminding-driven comparisons during category learning mediate attribution of abstract features to individual instances. We investigate the effect of comparison of instances presented together within a classification learning trial with the core prediction of better learning and transfer of relational categories.

This prediction is motivated by several sources including the rich literature supporting the structural alignment account of analogy and similarity (Gentner & Markman, 1997). Perhaps the most directly related evidence is the finding that 4-year-old children extend a label according to category match more frequently than by perceptual match when the label had been applied to two examples (Gentner & Namy, 1999). After only a single labeled example, children did not favor the category-based extension. Gentner and Namy conclude that a structural alignment process (invited by the common linguistic label) yielded a deeper, more conceptual encoding.

In light of recent findings that classification learning influences similarity, Boroditsky (in press) collected similarity ratings of pairs of object drawings from participants who had first listed either similarities or differences between the items. For both familiar and novel stimuli, items were rated more similar by participants who made comparisons than by those who did not. The effect depended critically on the items being similar – suggesting that comparison drew out a richer realization of the commonalities between alignable objects.

How might comparison work to mediate learning and representation? In the structural alignment framework, comparison influence encoding by: 1) highlighting common relations or alignable differences between examples; 2) projecting candidate inferences from one example to another; 3) promoting abstraction of shared structure as the basis for a generic knowledge structure; and 4) fostering re-representation that alters or re-organizes representational elements in one or both cases (Gentner & Wolff, 2000).

To illustrate, imagine a pair of cases for which a particular relation is encoded in the learner's mental representation of each instance. The process of comparing the representations would render this relation salient and promote abstraction and transfer (e.g., Loewenstein, Thompson & Gentner, 1999). Now, consider a common relation that is differently encoded in each case. Re-representation is posited as a means of aligning non-identical relational structures when there is semantic overlap (Gentner & Kurtz, in preparation) or a computational opportunity (Yan, Forbus, & Gentner, 2003). Next, consider a relation that has only been encoded in the representation of one of two instances. If there is sufficient surrounding structure in common, then a candidate inference would be projected from the more fully elaborated case to the sparser one (Gentner, 1983; Markman, 1997).

Finally, consider a common relation that is not encoded in the mental representations of either case. The mechanisms

listed above depend on the presence of relational information encoded in item representations. In the current project, we study learning in a novel domain and the underlying relation defining each category is far from self-evident to the uninitiated. In fact, short of resorting to exemplar memorization, the learning task is best characterized as trying to discover each relation. We posit a role for comparison in the discovery of relational content.

The notion of *manifest* versus *latent* representational content is of use here (Clement, Mawby, & Giles, 1994). While an individual may have somewhere in their idea of 'dog' the knowledge that dogs often feast on foodstuffs fallen to the floor during a family meal, this relational content is probably not routinely activated in a context-independent manner (Barsalou, 1982). Therefore, an analogy between a dog and vacuum cleaner might initially fall flat for someone who does not have the 'cleaning-up-of-table-scraps' aspect of their 'dog' concept activated. However, a thorough comparison of dog and vacuum cleaner could well activate latent matching content. Such *resurfacing* occupies a place between novel inference and highlighting, but, like the others, it relies on available relational content.

What is needed is a mechanism for articulating relational content over presumably unstructured initial inputs. Such processing is likely be of critical importance in any type of routine formation of structured mental representations since constraints are needed on which of the vast range of possible relations among objects, scenes, and situations in everyday experience should be explicitly encoded. As a number of theorists have put forth, language may be of particular use with regard to this problem.

We propose that comparison provides potential for a kind of *side-to-side* (as contrasted with top-down or bottom-up) interpretation process that promotes relational construal. The best evidence we can draw upon is the phenomenon of analogical bootstrapping in which intensive comparison of two partially understood depictions of a simple physics principle (heat flow) led participants to a deeper, more relationally-rich construal (Kurtz, Miao, & Gentner, 2001).

It is not clear how such analogical insight occurs, but here are two speculations. The first is consistent with the notion of progressive alignment (Kotovsky & Gentner, 1996) and states that observed commonalities at the level of lower-order representational elements (i.e., attributes, objects, first-order relations) may serve as entry points from which a familiar higher-order relation can be invoked. An example of instantiating a richer representation would be going from: *is-high(square)* and *is-low(circle)* to: *is-above (square, circle)*. The second speculation is that relations do not need to be built up so much as they need to be picked out of a crowd. The idea here is that many relations hold for any given case; too many to routinely articulate and encode. When given an opportunity to compare cases, the potential arises to find a manageable intersection of the relations.

In order to explore the power of comparison in knowledge change, our experimental question is as follows: can comparison promote the acquisition of novel categories defined by non-obvious relations? Relational categories have been treated theoretically (Gentner & Kurtz, in press; Markman & Stilwell, 2001) and have begun to receive empirical attention (Kurtz & Gentner, 2001; Rehder & Ross, 2001).

In a study using an early version of the present paradigm, Kurtz & Gentner (1998) found that participants reached a learning criteria for classification accuracy more quickly with trials consisting of within-category pairs than with single-instance trials. However, this can be attributed either to comparison or to more frequent exposure to training items, i.e., two instances per trial versus one. This creates a difficult circumstance for the researcher since fully convincing evidence for a comparison effect in learning (with frequency of exposure controlled) requires obtaining reliably higher classification accuracy on the basis of half the number of trials. This is the challenge we pursue.

An additional purpose of this project is the advancement of greater naturalism in the study of categorization. The dominant paradigm is a two-way classification task with instances that are clearly dimensionalized sets of perceptual or verbal features. Our stimuli are line drawings depicting a set of realistically varying "rock arrangements" having no clear reduction into a compositional set of underlying dimension values. Learners are asked to acquire three different categories to avoid two limitations inherent in binary classification: 1) a perfect success rate can be achieved based on an ability to identify examples of only one category and; 2) task demands encourage hypothesis-testing for a boundary over positively-defined concepts.

Experiment 1

One major concern in designing the first study was ensuring that the Comparison condition actually elicited comparison. An act of comparison can be shallow or intensive, and this difference can be a causal factor (Kurtz, et al., 2001). A failure to observe a comparison effect might be due to a failure by participants to compare. Classifying a within-category pair can easily be done with consideration of only one of the instances. Therefore, instead of all same-category pairs, we designed the Comparison condition to use an equal mix of within- and between-category pairs. In this mixed-pairs version, the status of any given pair is not known to the learner. Since the two instances may or may not belong to the same category, we collect two separate classification judgments on each learning trial. Accordingly, the participant must give direct consideration to each member of the pair. It is implicit in the task that the participant must consider whether or not to guess the same category for the two instances in each trial. However, classifying each instance could still be done largely independently despite taking place in a common task space.

For this reason, we used an orienting task at the beginning of each learning trial to encourage comparison. Participants were asked to consider the role played by one of the rocks relative to the rest of the arrangement and then to look for a

corresponding rock in the other instance. The orienting task for Single learners was to consider the role of one of the rocks in the arrangement. In both conditions this unenforced orienting task (no response was collected) was followed by a question that did require a response: whether or not the participant found the orienting task helpful. This was to discourage participants from ignoring the orienting task.

Method

Participants A total of 100 undergraduate students at Binghamton University received a course credit.

Materials A set of 36 images of rock arrangements was created on the computer. Rocks in each arrangement varied in color, shape, and size. A subset of 24 images were designated as the training instances and the remaining 12 images constituted the transfer set. The rock arrangements were evenly distributed across three categories given the names: "Tolar," "Besod," and "Makif." The category Tolar was defined by the presence of two stacked rocks similar in color and shape. Besod was defined by the presence of one rock supported by two others. Makif was defined by monotonically decreasing height from left to right. Care was taken that each instance conformed to exactly one of the relational categories. For the Comparison condition, a fixed set of pairings was established with an equal number of within-category and between-category pairs.

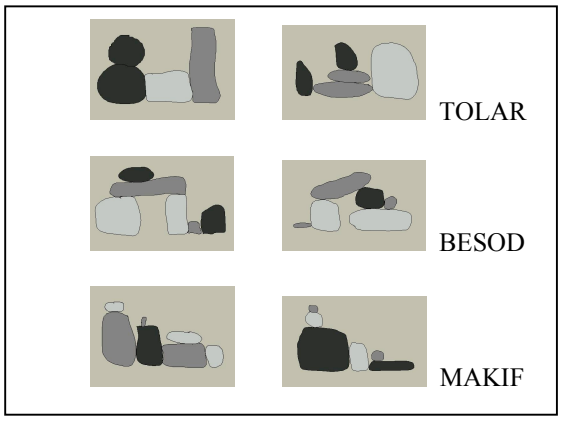


Figure 1: Sample Rock Arrangement Stimuli used in Experiments 1-3 Shown in Same-Category Pairs

Procedure Each participant was randomly assigned to one of the two conditions. Before the learning phase, participants read a set of instructions including a cover story about different rock arrangements created by the "Ladua" culture. Ss were instructed to try to learn to tell which rock arrangements belonged to which of the three types.

In the Single condition (n=50) an attempt was made to minimize the potential for temporal comparison by using a pseudo-random order in which each trial showed an instance from a different category than that of the previous trial. On each of the 48 learning trials participants were shown a single instance from the training set on the computer screen along with the orienting task: "Study the example, then focus on a single rock and consider the role it plays in the arrangement." Participants gave a forced-choice response regarding the helpfulness of the orienting task and were then asked to classify the rock arrangement into one of the three categories. After their choice was entered, corrective feedback was provided for a fixed interval of 3s. The stimulus image remained on screen for the entire trial.

In the learning phase of the Comparison condition (n=50), instances were presented two-at-a-time for 24 trials. On each trial, the presentation of the two instances on either the left or right side of the screen was randomized. The orienting task instructions for each trial were: "Study the examples, then focus on a single rock in one of the examples and consider the role it plays in that arrangement. Try to decide which rock plays a corresponding role in the other example." Participants made helpfulness judgments as in the Single condition. Ss were then asked to classify one of the instances followed by the other. Whether the left or right instance was queried first was alternated by trial. After the second response, corrective feedback for each of the responses was presented simultaneously for a total of 6s.

The learning phase in both conditions was followed by a common testing phase. Participants were presented with 24 old and 12 new items in random order and asked to classify each in a single-instance trial without feedback. Additional dependent measures were subsequently collected, but space limitations prevent their inclusion in this report.

Results and Discussion

The learning data reveal that it was not easy for most participants to acquire the relational categories in the allotted number of trials. We note that a set of pilot data showed that performance did not increase notably with twice the training. It is, however, important to remember that chance is 33.3% percent on a three-way classification, so the accuracy data reflects considerably more learning than it would appear at first glance. First we describe the learning data though we did not conduct statistical tests since the critical comparison between conditions is performance in the test phase when all participants respond to the same type of trial (single instance). Early (first quarter) classification accuracy shows that Comparison learners (M = .44, SD = .18) got off to a slow start compared to the Single group (M = .52, SD = .19), but they caught up by the final quarter: Single (M = .70, SD = .22) and Comparison (M = .69, SD = .25).

In the test phase, Comparison (M = .75, SD = .22) was significantly better than Single (M = .65, SD = .23) on old instances, $F(1, 98)= 4.73$, MSe = .234, $p < .05$. In addition, a transfer effect was found with Comparison (M = .72, SD = .22) significantly more accurate than Single (M = .59, SD = .27) in classification of novel instances, $F(1,98)= 7.29$, MSe = .444, $p < .05$. In sum, while Comparison learning presented a similar level of challenge during acquisition, a

reliable comparison advantage was found at test compared to Single learners receiving equal exposure.

Experiment 2

The goal of the second study was to determine whether a comparison effect would be found using only within-category pairs and a single categorization response per trial. The structural alignment view predicts a greater likelihood of comparison-driven effects on learning and encoding given the opportunity to compare alignable examples. However, as discussed, it is difficult to pin down the extent to which participants invoke a comparison process when making a joint classification response.

Method

Participants A total of 95 undergraduate students at Binghamton University received a course credit.

Materials The materials were the same as in Experiment 1. The assignment of pairs for the Comparison condition was accomplished by random generation of within-category pairings for each participant.

Procedure Each participant was randomly assigned to one of the two conditions. The Single condition (n=46) was conducted as in Experiment 1. The Comparison condition (n=49) followed the procedure of Experiment 1 except that participants were trained only on within-category pairs. Unlike Experiment 1, participants made a joint classification choice in response to both of the instances on each trial. Corrective feedback for the one response was shown for 3s.

Results and Discussion

A much different result was obtained relative to the findings of Experiment 1. Comparison learners with only within-category pairs showed good performance in the learning phase (M = .63, SD = .16) as compared to Single learners after an equal number of trials (M = .53, SD = .14), but not after an equal number of exposures (M = .60, SD = .14). No significant differences were found in test performance on old items (Comparison: M = .66, SD = .23 and Single: M = .67, SD = .21; $p > .8$) or transfer to new items (Comparison: M = .61, SD = .23 and Single: M = .63, SD = .24; $p > .6$). The evidence suggests that mixed pairs offer a more productive learning context than exclusively within-category pairs. This could be a benefit derived from evaluating whether or not co-presented pairs are from the same category. It could be due to useful contrastive evaluation of different-category items. However, it is worth noting on this point that no reliable difference was observed between learning accuracy on within-category and different-category trials in the Comparison condition of Exp. 1 ($p > .3$). Therefore, we are inclined to consider additional explanations. One possibility is that the joint classification task failed to fully encourage comparative evaluation of both instances in the trials. A final and somewhat compelling possible culprit is the 3s window for evaluating feedback as opposed to the 6s window for the dual-feedback in the Comparison condition of Exp. 1. In the current study, Comparison learners actually had half the overall amount of time to study images with their correct labels then was provided to Single learners.

Experiment 3

Given the lack of comparison advantage in Experiment 2, we considered the question of which is better: many different within-category comparisons over a large training set or repeated within-category comparisons over a small training set? It has been shown that larger category size promotes better transfer to novel examples when the instances in the training set are sufficiently variable (Homa & Vosburgh, 1976). Our hypothesis was that in the case of relational categories, repeated comparisons of within-category pairs in a smaller set would actually be more likely to promote an advantage of comparison in transfer accuracy. If Comparison learners in Exp. 2 underachieved due to a failure to fully compare and/or insufficient feedback time, repeated training on fewer examples might prove more conducive to comparison-driven learning.

Method

Participants A total of 87 undergraduate students at Binghamton University received a course credit.

Materials The materials were the same as in Experiment 1 except that the total number of examples in the training set was reduced to 12. Each category was represented by four, rather than eight, instances. All possible within-category pairings appeared once (determining 18 of the 24 trials in the Comparison condition). The remaining 6 trials were randomly determined for each participant including exactly one exposure of each training item.

Procedure The procedure was the same as in Experiment 2, though the same number of learning trials with fewer items in the training set yielded more exposures to each instance in the Single (n=42) and Comparison (n=45) conditions.

Results and Discussion

Comparison learners (M = .79, SD = .16) showed excellent overall accuracy on learning trials with the small category size relative to Single learners (M = .68, SD = .16). Although we have emphasized test performance rather than learning accuracy, Comparison learners were reliably more accurate across learning trials, $F(1, 85)= 10.39$, MSe = .25, $p < .005$) with equal frequency of exposures. The Comparison group (M = .68, SD = .16) also performed better on transfer items, $F(1, 85)= 3.91$, MSe = .18, p = .051, though the significance here was marginal. In performance on old items at test, a trend was found ($F(1, 85)= 2.54$, MSe = .09, p = .11) favoring the Comparison condition (M = .83, SD = .16) over Single (M = .76, SD = .21). These results provide yet another turnaround—the previous failure to find an advantage of comparison using within-category pairs is

overturned in the case of repeated comparisons with a small set of training items. The advantage is not limited to overlearning of the items in the small set since the results at test including transfer to new items favor Comparison. Our interpretation is that repeated comparison opportunities among increasingly familiar instances better allows the fruits of comparison to be borne out.

Experiment 4

We developed an additional paradigm to evaluate comparison of instances in category learning. The key difference is that in all conditions the orienting task is dropped and each learning trial begins with a single-instance classification judgment. In the Single condition, feedback is provided and the trial is done. In the Comparison condition, a within-category context item appears next to the target item, and the participant is asked for a second time to classify the initially presented target. Once the learner has made their second response, feedback is then provided based on the final response. Therefore, the only difference between conditions is that learners in the Comparison group are asked to repeat their classification choice in light of the availability of a context item from the same category for their consideration. We believe this is a naturally motivating and "unforced" version of comparison. A further advantage of this design is that since only the single target item is classified in each condition, we are better able to evaluate the impact of comparison during the learning phase.

Method

Participants A total of 50 undergraduate students at Binghamton University received a course credit.

Materials A full-sized stimulus set was used as in Experiments 1-2, but some alterations were made to the set. It was decided that the Tolar category was of a somewhat different character than the other categories since only two rocks in the entire arrangement participated in the relation of "same shape and color of two stacked rocks." In the other categories, the relation was more globally realized in the overall arrangement. A new relational definition and item set for the Tolar category was developed in terms of a symmetrical outline for each arrangement across the vertical axis. In addition, instances of the Makif and Besod categories were fine-tuned to ensure that the relation was globally realized in each rock arrangement. For example, the relation "one rock supported by two" would not be localized in a set of small rocks off to the side. These modifications were expected to make the learning task somewhat easier and the results more interpretable. Pairings for the Comparison condition were assigned such that all 28 possible same-category pairs in each category occurred once during learning and the remaining 12 learning trials were repetitions equated for instance exposure.

Procedure Each participant was randomly assigned to one of three conditions. Prior to the start of the learning phase participants read the instruction set. However, in order to help limit cases in which a learner embarked on a counterproductive approach, the instruction set was given the following addition: "Each of the three types is based on a distinct way of arranging rocks. Please note: It is not a small detail or a feature of one single rock. It is something about the way in which the group of rocks are arranged."

In the Single condition (n=20), participants completed 96 classification learning trials of single instances of the new set of rock arrangements in pseudo-random order. Feedback was given after each trial with study time self-paced rather than a fixed window. In the Comparison condition (n=17), each trial began exactly like a Single condition trial. However, participants did not receive feedback on their response. Instead they were shown another within-category instance from the training set as a context item. Participants were asked for a second time to classify the initial target item (this was reinforced by presenting the question under the target, not the context item). An accompanying instruction encouraged Ss to compare the target to the additional example from the same category. Participants were instructed to feel free to change their initial answers or not. Ss received feedback on their second response with self-paced study time.

A third condition called Identical (n=13) was conducted just as the Comparison condition except that the additional context item was a repeat of the target item—resulting in two identical images shown side-by-side. This task (responding twice to the same stimulus) was justified in the instructions as something Ss might find helpful.

In all conditions, participants went on to a test phase like that used in Experiments 1-3.

Results and Discussion

Means and standard deviations for the learning phase accuracy are shown in Table 1. One-way ANOVA showed a main effect of learning condition on classification accuracy in the first quarter, ($F(2, 47)= 5.22$, $MSe = .16$, $p < .05$), in the last quarter, ($F(2, 47)=3.19$, $MSe = .12$, $p < .05$), and in the overall performance, ($F(2, 47)= 3.77$, $MSe = .11$, $p < .05$). The first quarter difference was driven by the Identical group and most likely reflects participants adjusting to the somewhat odd repeated query. Planned comparisons showed that last-quarter accuracy (the final 24 trials) was significantly higher in Comparison versus Single, $t(35)= 2.22$, $p < .05$), as well as Comparison versus Identical, $t(28)= 2.27$, $p < .05$.

In the test phase there was a marginal main effect of learning condition on accuracy, $F(2, 47)= 3.167$, $MSe = .142$, $p = .051$. Planned comparisons showed that Comparison learners performed significantly better on old items ($M= .93$, $SD= .15$) than Single learners ($M = .82$, $SD = .18$), $t(35)= 2.06$, $p < .05$. Performance on new items was also better for the Comparison condition ($M = .87$, $SD = .18$) than the Single condition ($M = .70$, $SD = .20$), $t(35)= 2.71$, $p < .05$. Trends (presumably due to small sample size) were found in favor of Comparison over Identical for old items ($p = .08$) and new items ($p = .12$). No difference was found in

accuracy between the Single condition and Identical condition. We see in these results good evidence for better learning with the opportunity to compare to a within-category context item versus conditions with no additional comparison or a kind of item self-comparison that serves as a full control for exposure (equal number of classification responses; equal number and duration of item exposures).

Table 1. Classification Accuracy in Learning.

	Mean	SD
First quarter		
Single	.69	.18
Comparison	.71	.17
Identical	.52	.16
Last quarter		
Single	.81	.18
Comparison	.94	.15
Identical	.77	.24
Overall		
Single	.77	.16
Comparison	.84	.16
Identical	.67	.18

General Discussion

We conclude that comparison of instances during category learning is not necessarily of great impact, but when task constraints emerge that engage the learner to apply the machinery of comparison, superior performance in learning relational categories is achieved. These findings are most naturally understood in terms of learning to construct richer, more sophisticated encodings of category instances. While this is a difficult process, it is made easier by comparison.

Acknowledgements

We wish to thank Dedre Gentner for her important role in the work leading up to this research. We thank the members of the Learning and Representation in Cognition (LaRC) Laboratory including Aliza Nelson for constructing stimuli.

References

Barsalou, L.W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition, 10*(1), 82-93.

Boroditsky, L. (in press). Comparison and the development of knowledge. *Cognition*.

Clement, C., Mawby, R., & Giles, D. (1994). The effects of manifest relational similarity on analog retrieval. *Journal of Memory & Language, 33*(3), 396-420.

Elio, R. & Anderson, J.R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 397-417.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gentner, D. & Kurtz, K.J. (in press). Learning and using relation categories. In Ahn, W.K., Goldstone, R.L., Love, B.C., Markman, A.B., & Wolff, P.W. *Categorization inside and outside the lab.* Washington, DC: American Psychological Association.

Gentner, D. & Kurtz, K.J. (in prep). Relational and object matches in on-line evaluation of sentence analogies.

Gentner, D. & Markman, A.B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*, 45-56.

Gentner, D & Namy, L. (2000). Comparison in the development of categories. *Cognitive Development, 14,* 487-513.

Gentner, D. & Wolff, P. (2000). Metaphor and knowledge change. In E. Dietrich, & A. Markman, Cognitive Dynamics: *Conceptual Change in Humans and Machines* (pp. 295-342). NJ: Lawrence Erlbaum Associates.

Homa, D. &Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning & Memory, 2*(3), 322-330.

Ross, B., Perkins, S., & Tenpenny, P. (1990). Reminding-based category learning. *Cognitive Psychology, 22*, 460-492.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797-2822.

Kurtz, K.J. & Gentner, D. (1998). Category learning and comparison in the evolution of similarity structure. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 1236.

Kurtz, K.J. & Gentner, D. (2001). Kinds of kinds: Sources of category coherence. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* 522-527.

Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10*(4), 417-446.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*, 586-597.

Markman, A.B. (1997). Constraints on analogical inference. *Cognitive Science, 21*, 373-418.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Intelligence, 13*, 329-358.

Rehder, B. & Ross, B. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition. 27*(5), 1261-1275.

Spalding, T. & Ross, B. (1994). Comparison-based learning: Effects of comparing instances during category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*(6), 1251-1263.

Spalding, T. & Ross, B. (2000). Concept learning and feature interpretation. *Memory & Cognition, 28*, 439-451.

Yan, J., Forbus, K., Gentner, D. (2003). A theory of rerepresentation in analogical matching. *Proceedings of Twenty-Fifth Annual Meeting of Cognitive Science Society.*

Converging on a New Role for Analogy in Problem Solving and Retrieval

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, P.O. Box 6000
Binghamton, NY 13902 USA

Jeffrey Loewenstein (jeffrey.loewenstein@columbia.edu)
Columbia Business School, 3022 Broadway
New York, NY 10023 USA

Abstract

A novel approach to generating retrieval and transfer of structured knowledge is presented. We investigate the effect of comparing two analogous unsolved problems at test as opposed to comparing two solved analogous stories during initial study. We found that both procedures facilitate transfer relative to a standard baseline group studying one solved story and then attempting to solve a new analogous problem. In two studies we demonstrate that: 1) comparing two unsolved problems at test promotes analogical problem solving at least as effectively as comparing two fully solved problems during study; and 2) comparing two unsolved problems is helpful even when no source story is made available for retrieval.

Introduction

There is a wealth of cognitive science research about how people learn from examples and use them to solve new problems (Reeves & Weisberg, 1994). We also know that people are unlikely to spontaneously compare examples that seem different on the surface even though such comparison can provide learning and transfer advantages (Gick & Holyoak, 1983; Kurtz, Miao & Gentner, 2001). Retrieving analogous matches is therefore both important and demonstrably difficult. Research on retrieval shows that people have an easier time accessing examples on the basis of surface features than structural match (Catrambone, 2002; Gentner, Rattermann & Forbus, 1993; Ross, 1987). It is not that structural matches, particularly partial matches, are impossible or even rare, just that surface matches tend to predominate among novices, whereas experts seem able to exhibit structure matches more reliably (Dunbar, 2003; Novick, 1988).

We know that comparing examples can lead people to focus on common systems of relations which can in turn facilitate knowledge transfer (e.g., Loewenstein, Thompson & Gentner, 1999). The conventional wisdom in the field is that upon encountering a test problem, people are able to retrieve the earlier analogous cases, or a schema abstracted from those cases, to generate potential solutions to the problem (Gick & Holyoak, 1983). The implication is that the similarity function used in memory retrieval can link a current case to a prior case if the prior case is represented well in long term memory. The specific nature of such a superior representation is a challenging question for the field, but we take as a starting point the idea that a good representation is one that accurately encodes pertinent systems of relational structure and does so with sufficient generality to support transfer. This generality can be considered in terms of domain generality of encoded relational content (Clement, Mawby & Giles, 1994), uniformity of representational elements (Forbus, Gentner, & Law, 1995), or filtering out of mismatching irrelevant case details (Hummel & Holyoak, 1997).

We are currently intrigued by a new role for analogy in memory retrieval (see also, Loewenstein, Gentner, & Thompson, 2004). Our question is whether the benefits of this kind of representational "improvement" to the analogical source might also be observed with respect to the target (probe). Is a structural reminding more likely with a target that is better encoded? Theories of memory retrieval rely on a similarity function between the probe and stored items. Such similarity functions are symmetric. Since the empirical data suggest that only one side (i.e., the source) needs to be well-encoded to encourage a match, then it is plausible that a relevant, but regularly encoded source might become more retrievable on the basis of applying a probe with a superior encoding. In addition to being a theoretical possibility, there is a phenomenon, admittedly rare, of recalling an example with the sense of having a new understanding about it as a result of something we have just learned. The current line of thinking could explain such occurrences. Furthermore, it suggests a mechanism by which reflection upon a newly learned principle or abstraction could be a prod to retrieve prior examples, reinterpret them, and integrate the new knowledge with the old. Drawing analogies might then not only be a source of changes in knowledge from this point forward, but could also be a means for reorganizing the knowledge we already have and retrieving further analogous matches.

To reiterate, one of the seminal findings in the analogy literature is that problem solvers are more successful in retrieving an available solution strategy when they have previously made use of comparison to improve the encoding of the source analogs (Gick & Holyoak, 1983). We adapt this highly influential paradigm to ask the following question: Can comparison of target problems be used to facilitate analogical retrieval? There is considerable reason for skepticism. First, the advantage of source comparison is thought to rely on storing a generalized version of the solution principle, but in the case of target comparison the solution is not part of the compared cases—only the two problem statements are available. Secondly, the traditional

account suggests that structural reminding depends on having a well-represented source in memory—it may well follow that structural reminding is largely a dead end without a well-encoded source. Third, it is easy to imagine that having two problems to solve rather than just one could divide attention and processing resources in a detrimental manner. Finally, there is an extensive tradition of failed attempts to improve analogical problem-solving performance. Even so, if comparison at test can improve the encoding of targets such that retrieving structural matches is facilitated, this would have significant theoretical and applied ramifications. In the following two studies, we explore comparison-improved representation at the point of actual problem solving in hopes of gaining new insights into learning, retrieval, and transfer.

Experiment 1

In the current studies we use classic materials to study a novel set of questions about analogical problem solving: Duncker's (1945) tumor problem and its associated materials generated by Holyoak and colleagues (Gick & Holyoak, 1980, *inter alia*). In Experiment 1, we use these materials to examine whether retrieval is a two-way street. That is, if comparing two examples at study facilitates transfer at test (as shown by Gick & Holyoak, 1983), then can comparing two examples at test facilitate retrieval from study? In addition to the comparison being on-line rather than during initial study, the other key difference is that compared target problems do not include the solution.

We include two conditions to replicate prior data: a baseline group receiving one solved story at study and one problem at test (which will presumably yield little transfer) and a group comparing two solved stories at study and then receiving one problem at test (which will presumably show transfer). The key question is what will result in a new condition with one solved story at study and a comparison of two unsolved problems at test. Will participants who compare two test problems show greater success than participants in the baseline condition? Furthermore, to address the question of whether success hinges on transfer via retrieval of the source story, we include a group asked to compare and solve two problems without having first seen a solved story at study. Participants in this group are the only ones to receive no exposure at all to the relevant solution strategy.

Method

Participants A total of 293 undergraduate students at Binghamton University participated in partial fulfillment of a course requirement. Participants were randomly assigned to one of four conditions: Baseline, Source Comparison, Target Comparison, or Just Targets.

Materials The target case in all conditions was the well-known Radiation problem developed by Duncker (1945) and further studied by Gick & Holyoak (1980, 1983). The source and comparison cases were analogs based on the convergence principle used by Gick & Holyoak (1983). The source case was "The General" set in a military context and the comparison case was "Red Adair" set in a firefighting context. The comparison case was given with solution included during the study phase in the Source Comparison condition. We used the same Red Adair problem for comparison at test without the last lines that give the convergence solution in the Target Comparison and Just Targets conditions.

Procedure All phases of the experiment were conducted using paper packets for the presentation of instructions and materials as well as for the collection of responses. Separate packets were created for study and test phases. Participants did not receive the test packet until they completed and handed in their study packet (if a study packet was required by their condition).

In the Baseline condition (1:1, meaning participants were given 1 source story and 1 target problem), participants were instructed at the beginning of Part I to read the story (General) carefully and to gain sufficient familiarity that they could retell the story in their own words. Toward the bottom of the page, participants were asked: "What critical insight allowed the problem in the story to be solved?" In Part II, participants were asked to read the problem (Radiation) and to "use the space at the bottom of the page to explain how the problem can be solved."

In the Source Comparison condition (2:1), participants were instructed in Part I to carefully read two stories (General and Red Adair). The two stories were shown on the same page with General appearing first. At the top of the second page were two tasks to encourage better encoding. As in the control condition, participants were asked to gain sufficient familiarity that they could retell the stories in their own words. In addition, participants were asked to "Consider the parallels between the two stories" and complete a task in which five elements of Column A (General) had to be matched with elements of Column B (Red Adair). Each element had exactly one appropriate match. The columns were prepared in a jumbled order so that no correctly corresponding elements were directly across from one another. In Part II, participants were asked to solve the Radiation problem. The exact same procedure was used as in the Baseline condition.

In the Target Comparison condition (1:2), participants were presented with the General story using the exact same procedure as in the Baseline condition. In Part II, these participants were given two problems to solve (Radiation and Red Adair) The first page of the packet gave the following instructions: "What approach would you take to solve both of the following problems? After reading the problems carefully, please complete the matching task and then explain your proposed solutions in the space provided. Here's an important hint: The same strategy can be used to solve both problems."

Below the instructions were the two problems: Radiation followed by Red Adair. On the second page was a matching task between the Radiation and Red Adair problems

constructed in the same manner as above. On the last page, participants were again given the hint that "The same type of solution can be used" and asked to: "Please write down how these two problems can be solved."

In the Just Targets condition (0:2), participants were given Part II of the Target Comparison condition only. That is, they were asked to solve the Radiation and Red Adair problems without any prior exposure to the General story and its convergence solution. There was one additional procedural difference. The same-solution hint was provided, but in this case it was given only at the point when participants were actually asked to produce their solutions (rather than mentioning the hint twice).

An important point to clarify here is that this hint is distinct in type from the well-known use of a hint in the Gick & Holyoak studies. In that prior work the hint was to use the initial story as a basis for solving the target problem. That hint removed the fundamental obstacle in analogical problem solving: achieving a spontaneous structural reminding. The manipulation was of critical theoretical importance since it revealed that the Radiation problem was widely solved once participants accessed the source analog. In our current work, the hint has nothing to do with retrieval; instead it enforces mutual consideration of the two target problems.

Scoring A rater blind to condition scored each response for success in solving the Radiation problem in terms of the convergence solution. Responses were scored as correct if they captured the key principle of a multiplicity of low-intensity rays acting in concert on the tumor. The rater marked any responses they considered questionable for discussion with another rater. The agreed-upon scoring was then recorded. In occasional cases in which more than one solution was proposed, participants were given credit for the correct answer if it was produced.

Results

As expected, we replicated prior data showing that people who compared two source stories showed a transfer advantage relative to a baseline group who only read one source story (as shown in Table 1, 38% vs. 13% generated convergence solutions), $\chi^2(1, N=146) = 12.12$, $p < .01$. The important new result is that the group comparing at test, rather than study, performed as well or better than all other groups. The Target Comparison group performed better than the Baseline group (51% vs. 13%), $\chi^2(1, N=142) = 24.06$, $p < .001$. There was also a trend toward better performance by the Target Comparison group than the Source Comparison group (51% vs. 38%), $\chi^2(1, N=142) = 2.62$, $p = .11$. Critically, the Target Comparison group also performed better than the Just Targets (25%) group, $\chi^2(1, N=147) = 10.58$, $p < .005$. This suggests that participants who compared target problems were drawing upon the story from study since this was the major difference between the Target Comparison and Just Targets conditions. Finally, the Just Targets participants were marginally more likely to derive the convergence solution than were Baseline participants, $\chi^2(1, N=145) = 3.62$, $p = .06$. This suggests that comparing two unsolved target problems facilitated reaching the correct solution as compared to the Baseline condition of single cases at study and test.

Table 1: Proportion of convergence solutions by condition

Condition	N	Proportion generating convergence solution
Baseline (1:1)	70	.13
Source Comparison (2:1)	76	.38
Target Comparison (1:2)	72	.51
Just Targets (0:2)	75	.25

Discussion

We were able to replicate the well-known finding that comparing two examples at study yielded transfer benefits at test relative to a control group reading just one story at study. The intriguing result is that comparing two problems at test resulted in higher performance than the control group. Perhaps even more surprising, the Target Comparison group performed slightly, but not significantly, better than the comparison at study group. Confronted by one hard problem, these results suggest that a reasonable course of action would be to seek another problem with the same underlying structure!

Prior research suggests that abstracting the convergence schema was critical for success on the Radiation problem. Yet it is highly unlikely that the Target Comparison group abstracted the convergence schema from one example (otherwise the control group should have done well too). Our interpretation is that comparing analogous problems can lead to better representations of one or both problems. Such an encoding is likely to serve as a more effective retrieval cue for analogical problem solving. Due to having better representations of the problems via comparison, participants recalled the initial source story and borrowed its convergence solution. That is, retrieving prior examples on the basis of structure might be feasible if the probe is sufficiently well encoded, just as the comparison at study condition suggests that retrieval is feasible if the stored item is sufficiently well encoded.

The lower level of performance by the group who compared test problems, but did not receive a story at study, provides support for the claim that retrieval was a factor. Further tests are needed however to determine whether the single versus repeated hint played any role in this finding. The marginal advantage obtained in the Just Targets (0:2) condition over the Baseline condition indicates potential, not only for problem comparison as a means to achieve analogical retrieval, but also as a means to generate problem insight right then and there via analogical encoding.

In sum, we found that performance on a difficult problem can be greatly facilitated by an on-line technique. It is not necessary to construct improved representations at the time of encoding because one can do the necessary work through

comparison at test. Furthermore, such comparison is between problems, not between solved stories. The power of this comparison is not based on highlighting the convergence principle, but arises from comparison of two problem scenarios both amenable to a convergence solution.

Experiment 2

A second study was designed to replicate our basic finding and to further test whether drawing comparisons was an important component of the Target Comparison group's strong performance in Experiment 1. In this study we contrast the Target Comparison condition with a condition also receiving one study story and two test problems, but not guided with a hint to seek one solution for both problems. This Separate Targets condition still includes a matching task and the task to write down how "these problems can be solved," but the specific suggestion to work toward a single solution strategy is removed. If the Target Comparison group outperforms the Separate Targets group, this would serve as an indication that the depth of comparison of the problems is critical, just as comparing study problems is critical (Catrambone & Holyoak, 1989; Loewenstein, et al., 1999; Kurtz, et al., 2001). Additionally, in Experiment 1, the Source Comparison group tended to perform less well than the Target Comparison group, so a Source Comparison condition was included to test for a reliable difference.

Method

Participants A total of 224 undergraduate students at Binghamton University participated in partial fulfillment of a course requirement. Participants were randomly assigned to one of three conditions: Source Comparison, Target Comparison, or Separate Targets.

Materials, Procedure and Scoring The same source and target cases, the same use of paper packets, and the same scoring procedures were used as in Experiment 1. The Source Comparison (2:1) and Target Comparison (1:2) conditions were conducted using the same experimental and scoring procedure as in Experiment 1. The Separate Targets (1:2 without hint) condition followed the Target Comparison condition exactly with the exception that the initial hint and hint repetition were excluded from the text of the instructions.

Results

The main focus of this study was the contrast between the Target Comparison and Separate Targets conditions. People who received two problems, but no hint to compare them generated the convergence solution infrequently (16%, see Table 2). As in Experiment 1, the Target Comparison group frequently generated convergence solutions (40%), and did so reliably more often than did participants in the Separate Targets condition, $\chi^2(1, N=147) = 10.77$, $p < .005$. Thus an explicit instruction to compare and generate a common solution was critical to the effectiveness of the Target Comparison manipulation.

There was little difference between the Source Comparison (35%) and Target Comparison (40%) groups in this study, $\chi^2 < 1$. The previous study suggested there might be a difference between the two conditions, and the ordering of the means was consistent, but the difference in this study was minimal.

Table 2: Proportion of convergence solutions by condition

Condition	N	Proportion generating convergence solution
Source Comparison (2:1)	77	.35
Target Comparison (1:2)	72	.40
Separate Targets (1:2) without hint	75	.16

Discussion

This study replicated the effectiveness of comparing two target problems. It also confirmed an important boundary condition, namely that comparing the target problems toward drawing out a common solution was important. Merely receiving two target problems with minimal encouragement to assess their parallels was not effective. Indeed, solving two target problems separately led to comparable performance as solving one target problem (i.e., the baseline condition) in Experiment 1.

General Discussion

With these two studies, we provide grounds for a new emphasis, if not a new interpretation, of analogical retrieval and transfer. The usual assumption is that comparing examples facilitates generation of a representation of the common schema that clarifies the key structure and is less cluttered by unrelated contextual details than the original examples. It is clear that without drawing a comparison, people are unlikely to represent the structure in such a way that it can be retrieved and used to solve a new problem—an effect we replicated in Experiment 1. The current results open up the possibility that the benefit of comparison at study may be due to: 1) improving the encoding of the examples rather than creating a new general knowledge structure; or 2) allowing people to form better encodings of subsequent cases using a more sophisticated or general representational vocabulary.

The current studies were aimed at addressing this issue by turning it around: what if people study just one example (so they are unlikely to form any particularly clear or uncluttered representation), but they compare examples at test and then profit from having read the earlier single case. The results of our two experiments are consistent with people being able to retrieve single stored cases in just this fashion. We showed a distinct transfer advantage for a group that was: (1) specifically encouraged to compare two unsolved test problems and (2) had previously studied a single case. One may not need to store cases in a

particularly good fashion if one can later construct a superior retrieval probe.

There are multiple implications if a comparison today can facilitate retrieving a case learned yesterday. First, with respect to models of the retrieval process, it suggests a constraint on the similarity process that matches stored items to probes: it may well have to be symmetric. Second, it suggests a mechanism by which people can reorganize their knowledge. One may not have to "learn it right the first time" if, after appreciating a new abstraction, one is able to retrieve and perhaps re-represent a prior matching example. This supplies a concrete mechanism for gradual conceptual change in both development and the acquisition of expertise. Third, this implies that educators, particularly those who teach adults, can look to integrate people's prior experiences in their formal acquisition of domain expertise.

A second point arising from these studies is that drawing comparisons can facilitate learning in a new way. Typically, people draw comparisons to understand a principle or solution in a more general way. In our studies, people used comparison to generate better formulations of the problem at hand, not a better understanding of provided solutions. There are at least three reasons as to why this should facilitate problem solving. First, comparing two problems with the knowledge that that they have a common solution type means that idiosyncratic information can be ignored. Second, potentially misleading example-specific solution types can be ruled out. Third, it may allow people to formulate a more abstract or general version of the problem at hand. As Polya (1945) suggested, despite it seeming counterintuitive, sometimes a more general problem is easier to solve than a more specific problem. There may be interesting and important applications of this use of comparison both in education and in discovery.

We find these studies an intriguing first step. We are pursuing several related issues that might influence our interpretation of these studies. The Just Targets condition was given a weaker hint to compare than the Target Comparison condition, and as the Separate Targets condition showed: hints are important. We are running a new study that examines equal encouragement to draw comparisons. We are also interested in whether the Target Comparison condition benefits from one problem being easier to solve than the other (in which case its solution would be tested on the second problem) or whether it is the development of a more general version of the problem that is driving participants' success.

In conclusion, drawing comparisons may facilitate learning and transfer in multiple ways. It may enhance recalling prior experiences as much as generating knowledge that is likely to be later transferred. It may enhance clarifying a problem formulation as much as deriving generalizations from solved problems.

Acknowledgments

Special thanks to Jessica Federman and Aliza Nelson. We also thank Janeen O'Connor, Olga Boukrina, and the rest of the Learning and Representation in Cognition (LaRC) Laboratory at Binghamton University.

References

Catrambone, R. (2002). The effects of surface and structural feature matches on the access of story analogs. *Journal of Experimental Psychology: Learning, Memory and Cognition, 28*(2), 318-334.

Clement, C.A., Mawby, R., & Giles, D.E. (1994). The effects of manifest relational similarity on analog retrieval. *Journal of Memory and Language, 33*, 396-420.

Dunbar, K. (2003). The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought*. (pp. 313-334) Cambridge, MA: MIT Press.

Duncker, K. (1945). On problem-solving (L. S. Lees, Trans.). *Psychological Monographs, 58*(5), Whole No. 270.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*(2), 141-205.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability and inferential soundness. *Cognitive Psychology, 25*, 524-575.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*(3), 427-466.

Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10*(4), 417-446.

Loewenstein, J., Gentner, D., & Thompson, L. (2004). Analogical Encoding: Facilitating Knowledge Transfer and Integration. Working paper, Columbia University.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*(4), 586-597.

Novick, L. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 510-520.

Polya, G. (1945). *How to solve it*. Princeton, NJ: Princeton University Press.

Reeves, L. M., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115* (3), 381-400.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory & Cognition, 13*(4), 629-639.

Pronouns Predict Verb Meanings in Child-Directed Speech

Aarre Laakso (alaakso@indiana.edu)
Linda Smith (smith4@indiana.edu)
Department of Psychology, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

Do statistical regularities between pronouns and verbs help children learn verb meanings? This question is addressed by an analysis of child directed speech. The results show that there are statistical regularities in the co-occurrences of pronouns and verbs that could be used to cue verbs that describe physical motion, psychological states, and features such as transitivity. The learnability of these regularities is demonstrated in a simulation study.

Introduction

It is well known that learning the meanings of verbs is a difficult task for young children. It is also well known that pronouns make up a substantial proportion of the nouns that children hear. The distributional relations between pronouns and verbs thus could play a role in early verb learning.

There are several reasons why verb meanings are difficult for children to learn. Whereas parents label objects (relatively) often, they rarely label events or relations. There are no observable referents for many verbs, such as psychological state verbs like *look*, *think*, *want*, *believe*, and *know*. Even verbs that refer to observable actions present ambiguities—for example, when does the opening of a door begin? Further, the aspect of an action that is relevant is ambiguous, and could be, for example, the manner or the path. Finally, verb meaning often depends on taking a particular perspective on a scene; consider the difference between "giving" and "receiving." In brief, meaning maps between verbs and the world are not transparent. Accordingly, many have suggested that word-word relations are particularly important to learning verbs (see, for example Gleitman, 1990; Gleitman & Gillette, 1995). Here we examine how statistical relations between pronouns and verbs in parental speech might help children learn the meanings of the verbs.

Pronouns are, by far, the most common syntactic subjects and objects in adult speech to children. Most syntactic subjects in spontaneous spoken adult discourse in general are pronouns (Chafe, 1994), and English-speaking mothers often begin with a high-frequency pronoun when speaking to their children, with *you* and *I* occurring most frequently (Valian, 1991). The sheer frequency of pronouns suggests that pronouns—and their statistical co-occurrences with verbs—may be developmentally very powerful.

Consistent with this idea, Childers & Tomasello (2001) suggested that children acquire lexically specific frames such as "I do it" as a way into learning syntactic frames. Cameron-Faulkner, Lieven, & Tomasello (2003) also observed that parents use the inanimate pronoun *it* far more frequently as the subject of an intransitive sentence than of a transitive one. As Cameron-Faulkner et al. note, this suggests that intransitive sentences are used more often than transitives for talking about inanimate objects. It also suggests that the use of the inanimate pronoun might serve as a cue to some semantic aspects of the verb.

Pronouns may also help learners partition verbs that express psychological attitudes toward events and states of affairs into two rough categories. Verbs that express deontic status, such as goals, purposes or intentions (*try to*), volitions or desires (*want to*), and compulsions (*have to*) tend to take infinitival complements, whereas verbs that express epistemic status, such as perceptions (*see that*), beliefs (*think that*), and knowledge (*know that*) tend to take sentential (propositional) complements (Tomasello, 2003). In the ecology of early childhood, parents tend to be the ones who *know* whereas children tend to be the ones who *need*. All this suggests the potential value of examining the distributional relations among pronouns and verbs in language to young children.

Experiment 1

The first experiment consisted of a corpus analysis to demonstrate patterns of co-occurrence between pronouns and verbs in the child's input.

Method

Parental utterances from the CHILDES database (MacWhinney, 2000) were coded for syntactic categories, then subjected to clustering and statistical analysis. The target children represented in the transcripts were aged approximately 1;4 – 6;1.

Materials The following corpora were used: Bates, Bliss, Bloom 1970, Brown, Clark, Cornell, Demetras, Gleason, Hall, Higginson, Kuczaj, MacWhinney, Morisset, New England, Post, Sachs, Suppes, Tardiff, Valian, Van Houten, Van Kleeck and Warren-Leubecker.[1]

Coding was performed using a custom web application that randomly selected transcripts, assigned them to coders, collected coding input, and stored it in a MySQL database. The application occasionally assigned the same transcript to all coders, in order to measure reliability. Five undergraduate coders were trained on the coding task and

[1] The full references for each corpus may be found in (MacWhinney, 2000).

the use of the system. Cluster analysis and other statistical analyses were performed using MATLAB and R.

Procedure Each coder was presented, in sequence, with each main tier line of each transcript she was assigned, together with several lines of context; the entire transcript was also available for viewing by clicking a link on the coding page. For each line, she indicated (a) whether the speaker was a parent, target child, or other; (b) whether the addressee was a parent, target child, or other; (c) the syntactic frames of up to 3 clauses in the utterance; (d) for each syntactic frame, up to 3 subjects, auxiliaries, verbs, direct objects, indirect objects and obliques. Nouns appearing in prepositional phrases were coded as obliques (with the exception of recipients indicated with "to", which were coded as indirect objects). Object complements were indicated by coding the direct object of the matrix verb as "<clause>" and coding the constituents of the complement clause as the next clause associated with the utterance. This was intended both to simplify the coding scheme and to avoid attributing too much grammatical knowledge to the child—we do not assume that the child can convert an utterance into an accurate parse tree, only that she can identify verbs and the nouns that surround them.

The syntactic frames were: no verb, question, passive, copula, intransitive, transitive and ditransitive. The *no verb* frame included clauses—such as "Yes" or "OK"—with no main verb. The *question* frame included any clause using a question word—such as "Where did you go?"—or having inverted word order—such as "Did you go to the bank?"—but not merely a question mark—such as "You went to the bank?" The *passive* frame included clauses in the passive voice, such as "John was hit by the ball." The *copula* frame included clauses with a copula (including *be*, *seem* and *become*) as the main verb, such as "John is angry." The *intransitive* frame included clauses with no direct object, such as "John ran." The *transitive* frame included clauses with a direct object (or an object complement) but no indirect object, such as "John hit the ball." The *ditransitive* frame included clauses with an indirect object, such as "John gave Mary a kiss."

In total, 59,977 utterances were coded from 123 transcripts. *All* of the coders coded 7 of those transcripts for the purpose of measuring reliability. Average inter-coder reliability (measured for each coder as the percentage of items coded exactly the same way they were coded by another coder) was 86.1%.[2]

We only considered parental child-directed speech (PCDS), defined as utterances where the speaker was a parent and the addressee was a target child. A total of 24,286 PCDS utterances were coded, for a total of 28,733 clauses. More than a quarter (28.36%) of the PCDS clauses contained no verb at all; these were excluded from further analysis. Clauses that were questions (16.86%), passives (0.02%), and copulas (11.86%) were also excluded from further analysis. The analysis was conducted using only clauses that were intransitives (17.24% of total PCDS clauses), transitives (24.36%) or ditransitives (1.48%), a total of 12,377 clauses.

We formed 2 matrices from these clauses: a verbs-by-subjects matrix and a verbs-by-objects matrix. The verbs-by-subjects matrix contained only verbs used with an overt subject; its size was 621 verbs by 317 nouns (subjects). The verbs-by-objects matrix contained only verbs used with a direct object; its size was 524 verbs by 907 nouns (objects). Each cell of each matrix contained the proportion of times that verb was used with that noun (as subject or object) in a coded clause.

For the purposes of exploratory data analysis, we then performed 4 cluster analyses. First, we took the 50 nouns most commonly used as objects and clustered them according to their proximity in verb space, i.e., the space formed by considering each verb as a dimension. Each noun was placed along each dimension according to the proportion of times it was used with the corresponding verb. Hence, a noun never used as the object of a particular verb would be at 0, and a noun always used as the object of a particular verb would be at 1. Second, we clustered the 50 most common subject-nouns in verb space. Third, we took the 50 verbs most commonly used with objects and clustered them according to their proximity in noun space (defined analogously to verb space). Finally, we clustered the 50 most common verbs-with-subjects in noun space.

Results

We cannot show the cluster diagrams here due to space limits. We summarize the main regularities observed.

The observed distribution of nouns in the corpus is consistent with Zipf's law — the numerical frequency of words decays roughly as an inverse power of their rank frequency. Moreover, the most frequent subjects and objects in the corpus are pronouns, as shown in Figures 1 and 2.

Our quantitative analysis of co-occurrence relationships is based on the log likelihood ratio as described by Dunning (1993) and recommended by Manning & Schutze (1999). Suppose we have observed N clauses with m subject-(or object-) types and n verb-types. Let

$$S = s_1, s_2, \cdots, s_m \text{ and } V = v_1, v_2, \cdots v_n$$

represent the subjects and the verbs respectively. Furthermore, let

$$K_S = k_{s_1}, k_{s_2}, \cdots, k_{s_m} \text{ and } K_V = k_{v_1}, k_{v_2}, \cdots, k_{v_n}$$

represent the observed frequencies of the subjects and verbs respectively and

$$K_{SV} = k_{s_1 v_1}, k_{s_1 v_2}, \cdots, k_{s_1 v_n}, k_{s_2 v_1}, \cdots, k_{s_m v_n}$$

[2] Cohen's kappa coefficient (Cohen, 1960), which adjusts for chance agreement, is only applicable for two raters. We know of no chance-corrected multiple rater agreement measures that are widely used in the language acquisition literature. However, given the number of variables, the number of levels of each variable (3 speakers, 3 addressees, 7 frames, and 6 syntactic relations), and the number of coders (5) in the present study, the probability of chance agreement is low.

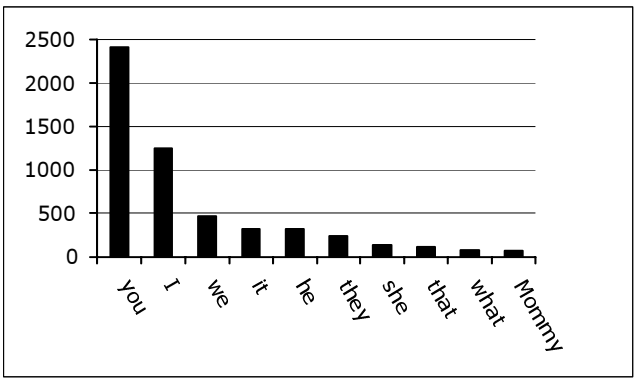


Figure 1: The 10 most frequent subjects in PCDS by their number of occurrences.

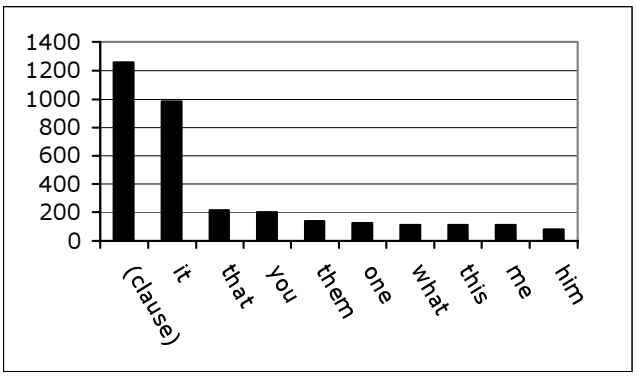


Figure 2: The 10 most frequent objects in PCDS by their number of occurrences.

represent the observed frequencies of subject-verb co-occurrences. Then the test statistic is the quantity:

$$-2\log\lambda = 2[\log L(p_1, k_{s_i v_j}, k_{v_j}) + \log L(p_2, k_{s_i} - k_{s_i v_j}, N - k_{v_j}) - \log L(p_0, k_{s_i v_j}, k_{v_j}) - \log L(p_0, k_{s_i} - k_{s_i v_j}, N - k_{v_j})]$$

Where the components are defined as follows:

$$L(p,k,n) = p^k(1-p)^{n-k} = k\log p + (n-k)\log(1-p)$$

$$p_0 = p(s_i) = \frac{k_{s_i}}{N}$$

$$p_1 = p(s_i \mid v_j) = \frac{k_{s_i v_j}}{k_{v_j}}$$

$$p_2 = p(s_i \mid \neg v_j) = \frac{k_{s_i} - k_{s_i v_j}}{N - k_{v_j}}$$

The test statistic -2logλ is χ^2 distributed with 1 degree of freedom. Intuitively, it represents how much more likely it is that s_i and v_j go together than should be expected purely by chance. It has also been demonstrated that this statistic identifies natural collocations in text corpora.

As expected, the inanimate pronoun "it" was more likely as the object of verbs of physical motion than as the object of psychological attitude verbs, whereas complement clauses were more likely to occur with psychological attitude verbs than with verbs of physical motion. As shown in Table 1, "it" tended to occur with physical motion verbs far more often than would be predicted by chance, and clauses occurred with most physical motion verbs, if at all, only about as much as would be predicted by chance. The verb "put" is an exception to this general rule, occurring with a clause more often than would be predicted by chance. As shown in Table 2, complement clauses tended to occur with psychological attitude verbs more often than would be predicted by chance, whereas "it" only occurred more often than would be predicted by chance with two of five psychological attitude verbs. The exceptions were *want* (uses such as "Oh, I want it now") and *know* ("No, that's wrong and you know it"). In any case, as shown in Figure 3, it is somewhat more likely that a physical motion verb will occur with "it" than with a complement clause, and substantially more likely that a psychological attitude verb will occur with a complement clause than with "it".

Table 1: The log likelihood ratio for uses of object "it" or a clause with physical motion verbs. * indicates p<0.01; — indicates no co-occurrences.

	"it"	(clause)
put	102.79*	70.70*
turn	72.58*	—
throw	39.55*	6.14
hold	32.17*	—
push	24.87*	3.02

Table 2: The log likelihood ratio for uses of object "it" or a clause with psychological attitude verbs. * indicates p<0.01; — indicates no co-occurrences.

	"it"	(clause)
think	—	399.13*
want	12.00*	283.28*
know	69.53*	134.44*
remember	—	37.22*
mean	0.91	15.81*

Table 3: The log likelihood ratio for uses of subject "I" or "you" with epistemic verbs. * indicates p<0.01; — indicates no co-occurrences.

	"I"	"You"
think	605.01*	24.7*
know	200.05*	108.17*
guess	60.00*	—

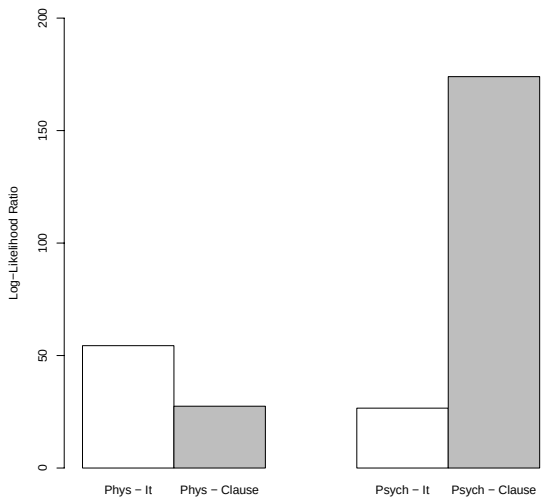


Figure 3: Mean log-likelihood ratio of the object pronoun "it" (white bars) versus a clause (gray bars) given a physical motion verb (left) or a psychological attitude verb (right).

We also found that "I" is more likely to be the subject of epistemic verbs, whereas "you" is more likely to be the subject of deontic verbs. As shown in Table 3, "I" occurred with epistemic verbs far more often than would be predicted by chance. The subject "you" also occurred more often with "think" and "know" than would be predicted by chance, but with a much lower likelihood.

As shown in Table 4, "you" tended to occur with deontic verbs far more often than chance would predict. The subject "I" was no more likely than chance would predict to appear with the verbs "like" and "need" and was only slightly more likely than chance to occur with the verb "want". In any case, as demonstrated in Figure 4, it is substantially more likely that the subject "I" will appear with an epistemic verb than it is that the subject "you" will appear with an epistemic verb. It is also somewhat more likely that "you" will appear with a deontic verb than "I" will appear with a deontic verb.

Table 4: The log likelihood ratio for uses of subject "I" or "you" with deontic verbs. * indicates p<0.01.

	"I"	"You"
want	6.72*	116.97*
like	0.03	74.24*
need	2.69	15.26*

We conclude by noting that there are many other significant co-occurrences in the corpus, some of which involve triadic correlations between specific verbs, specific nouns, and pronouns. For example the objects "book" and "story" are more likely to appear with the verb "read" than would be predicted by chance (LLR=131.51, 128.39). Both the object "book" and the object "this" are likely to appear with the phrasal verb "look at" (LLR=67.28, 88.01). Similarly, not only is "it" likely to appear as the object of "turn" (as discussed above), but so is "page" (LLR=81.89). Likewise for "play," which makes not only the objects "ball", "blocks", "game", and "house" more likely, but also the objects "this" and "it". These are potentially important on several fronts. The child may learn an association between pronouns such as "this" and "it" and inanimate objects, like books and pages. The pronouns "this" and "it" may then be used to help the child understand the meanings of other verbs that take inanimate objects as their objects. Conversely, the verb "tell" selects strongly for the pronouns "us" and "me" as well as for "Mommy" and "Daddy". Hence, the child may learn that verbs taking "us" and "we" as objects have to do with communicating with or directing attention toward other people.

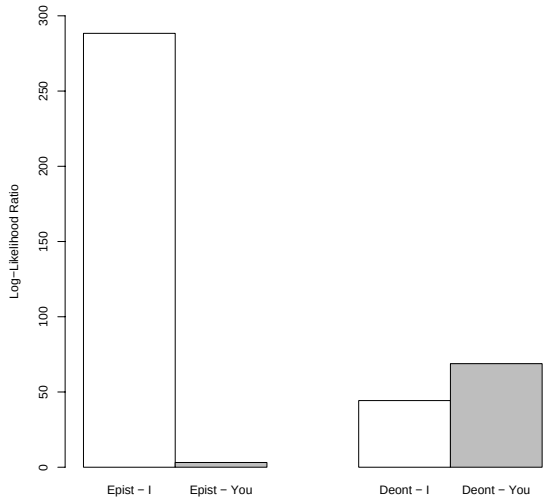


Figure 4: Mean log-likelihood ratio of the subject pronoun "I" (white bars) versus "you" (gray bars) given an epistemic verb (left) or a deontic verb (right).

Discussion

Although pronouns are "light" in their meaning, their referents determinable only from context, they may nonetheless be potent forces on early lexical learning by identifying some kinds of verb meanings as more likely than others. The results of Experiment 1 show that there are statistical regularities in the co-occurrences of pronouns and verbs that the child could use to discriminate verb meanings. Verbs that describe physical motion or transfer are likely to be followed by "it," whereas verbs attributing psychological state are likely to be followed by a relatively complex complement clause. Verbs having to do with thinking or knowing are likely to occur with subject "I," whereas verbs having to do with wanting or needing are likely to occur with subject "you." This regularity most likely reflects the ecology of parents and children—parents "know" and children "want"—but it could nonetheless be useful in distinguishing these two classes of meanings. The results thus far show that there are potentially usable regularities in the statistical relations between pronouns and verbs.

Experiment 2

To demonstrate that the regularities in pronoun-verb co-occurrences in parental speech to children can actually be exploited by a statistical learner, we trained a connectionist network to auto-associate subject-verb-object "sentences" from the input, then tested it on individual verbs and pronouns. We predict that the network should be able to learn the statistical regularities demonstrated in Experiment 1, specifically: (1) physical transfer verbs are likely to have "it" as an object, whereas psychological verbs are likely to take a complement clause, and (2) epistemic verbs are likely to have "I" as a subject, whereas deontic verbs are likely to have "you" as a subject.

Method

Data The network training data consisted of the subject, verb, and object of all coded utterances that contained the 50 most common subjects, verbs and objects. There were 5,835 such utterances. The inputs used a localist coding wherein there was exactly one input unit out of 50 activated for each subject, and likewise for each verb and each object. Absent and omitted arguments were counted among the 50, so, for example, the utterance "John runs" would have 3 units activated even though it only has 2 words—the third unit being the "no object" unit. With 50 units each for subject, verb and object, there were a total of 150 input units to the network. Active input units had a value of 1, and inactive input units had a value of 0.

Network Architecture The network consisted of a two-layer 150-8-150 unit autoassociator with a logistic activation function at the hidden layer and three separate softmax activation functions (one each for the subject, verb and object) at the output layer—see Figure 5. Using the softmax activation function, which ensures that all the outputs in the bank sum to 1, together with the cross-entropy error measure allows us to interpret the network outputs as probabilities (Bishop, 1995). The network was trained by the resilient backpropagation algorithm (Riedmiller & Braun, 1993) to map its inputs back onto its outputs. It is well known that this sort of network performs nonlinear dimensionality reduction at its hidden layers, extracting statistical regularities from the input data.

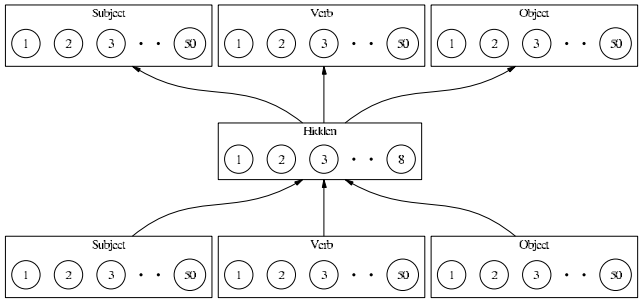


Figure 5: Network architecture

Training The data was randomly assigned to two groups: 90% was used for training the network, while 10% was reserved for validation. Starting from different random initial weights, 10 networks were trained until the cross-entropy on the validation set reached a minimum for each of them. Training stopped after approximately 150 epochs of training, on average. At that point, the networks were achieving about 81% accuracy on correctly identifying subjects, verbs and objects from the training set.

Testing After training, the networks were tested with incomplete inputs corresponding to isolated verbs and pronouns. For example, to see what a network had learned about *it* as a subject, it was tested with a single input unit activated—the one corresponding to *it* as subject. The other inputs were set to 0. Output unit activations were recorded and averaged over all 10 networks.

Results

To test the hypothesis that the network learns that psychological attitude verbs are more likely than physical motion verbs to take a clause as an object, we tested the networks with the frames "I ___ (clause)" and "You ___ (clause)" using psychological and physical verbs. The psychological verbs were "think," "want," "know," and "remember." The verb "mean" was not among the top 50 verbs used in the corpus and therefore was not used in the network training experiments. The physical verbs were "put," "turn," "throw" and "hold." The verb "push" was not among the top 50 verbs used in the corpus and therefore was not used in the network training experiments. As shown in Figure 6, the networks activated the psychological verbs more strongly at the output ($\underline{M}$ = 0.047, $\underline{SD}$ = 0.152) than the physical verbs ($\underline{M}$ = 0.002, $\underline{SD}$ = 0.014). This difference was significant, $\underline{t}(80) = -2.62$, $\underline{p} = 0.01$. Results are similar for the converse (physical verbs are significantly more activated when the object is "it") and for the epistemic / deontic distinction (epistemic verbs are significantly more activated when the subject is "I," whereas deontic verbs are significantly more activated when the subject is "you").

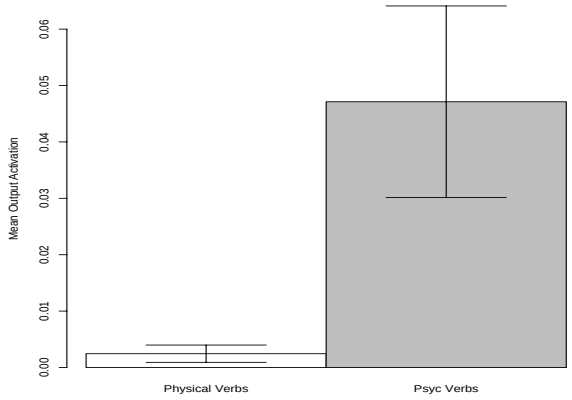


Figure 6: Network output activations for physical verbs versus psychological attitude verbs for the frames "I___(clause)" and "You____(clause)".

Conclusions

We have shown that there are statistical regularities in co-occurrences between pronouns and verbs in the speech that children hear from their parents, including regularities that distinguish between psychological and non-psychological verbs, as well as between deontic and epistemic psychological verbs. We have also shown that a simple statistical learner can exploit these regularities, not to learn the meanings of verbs per se (the network obviously does not know the meanings of the verbs), but to learn the formal associations between tokens of verbs and pronouns. How might this help the child learn the meanings of verbs? In the first place, hearing the verb framed by pronouns may help the child isolate the relevant event or action from the blooming, buzzing confusion around it; the pronouns can indicate animacy, gender, number and direction of causality. This would allow the child to focus on the relevant things. Second, if we suppose that the child has already learned one verb and its pattern of correlations with pronouns, and then hears another verb being used with the same or a similar pattern of correlations, the child may hypothesize that the meaning of the unknown verb is similar to the meaning of the known verb. For example, a child who understood "want" but not "need" might observe that "you" is usually the subject of both and conclude that "want," like "need," has to do with his desires and not, for example, a physical motion or someone else's state of mind.

Now that we have shown that the regularities exist and are learnable in principle, the next step is to show that children actually pick up on these regularities. We are planning a series of experiments with children from 24-36 months that will involve priming the children with movies showing actions corresponding to nonsense verbs in the context of various pronoun frames ("He zorps it," "It zorps," "They zorp this to her" and so on) and testing whether this influences their interpretations of these verbs, for example by having them select which of a pair of novel movies shows "zorping." This will provide a strong test of the hypothesis that children actually use pronoun distributional statistics to pick up on the meanings of novel verbs.

Future modeling experiments will attempt to capture not only the statistical relationships among verb-pronoun token co-occurrences but also their relationships with shared meanings, by associating words with featural representations of their meanings. We are also working on a mechanism for manipulating the statistical properties of the relevant conditional distributions to be used for generating network inputs in future simulations—the distribution of syntactic frames, the distribution of verbs given a syntactic frame, and the distribution of nouns (including pronouns) in each argument position given a verb. We expect that such a model could be used to test the utility of statistical associations between pronouns with verbs for a theoretical learner—simulations run with varying degrees of correlations could demonstrate not only whether but also just how much correlation is useful in principle.

Finally, work is underway to collect crosslinguistic data from Japanese and Tamil, verb-heavy languages with frequent argument dropping and case-marked pronouns referring to various levels of social status. We are especially keen to find out what sorts of cues children might be using in languages where pronouns are both rarer and "heavier" than they are in English.

Acknowledgments

This research was supported by NIMH grant number ROI MH 60200. Thanks to our coders for their hard work, to members of the Cognitive Development Laboratory at IU for useful discussions, and to three anonymous reviewers for thoughtful comments on an earlier draft.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Cameron-Faulkner, T., Lieven, E. V. M., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27*, 843-873.

Chafe, W. L. (1994). *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Childers, J. B., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology, 37*(6), 739-748.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Gleitman, L. R. (1990). The structural sources of word meaning. *Language Acquisition, 1*(1), 3-55.

Gleitman, L. R., & Gillette, J. (1995). The role of syntax in verb learning. In P. Fletcher & B. MacWhinney (Eds.), *The Handbook of Child Language* (pp. 413-427). Cambridge, MA: Blackwell.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed. Vol. 2: The Database). Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Riedmiller, M., & Braun, H. (1993). *A direct adaptive method for faster backpropagation learning: The Rprop algorithm.* Paper presented at the IEEE International Conference on Neural Networks 1993 (ICNN 93), San Francisco, CA.

Tomasello, M. (2003). *Constructing a Language*. Cambridge, MA: Harvard University Press.

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition, 40*(1-2), 21-81.

The Natural Input Memory Model

Joyca P.W. Lacroix (j.lacroix@cs.unimaas.nl)
Department of Computer Science, IKAT, Universiteit Maastricht, St. Jacobsstraat 6, 6211 LB Maastricht, The Netherlands

Jaap M.J. Murre (jaap@murre.com)
Department of Psychology, Universiteit van Amsterdam, Roeterstraat 15, 1018 WB Amsterdam, The Netherlands

Eric O. Postma (postma@cs.unimaas.nl)
H. Jaap van den Herik (herik@cs.unimaas.nl)
Department of Computer Science, IKAT, Universiteit Maastricht, St. Jacobsstraat 6, 6211 LB Maastricht, The Netherlands

Abstract

A new recognition memory model is proposed which differs from the existing memory models in that it operates on natural input. Therefore it is called the natural input memory (NIM) model. A biologically-informed perceptual pre-processing method takes local samples from a natural image and translates these into a feature-vector representation. The feature-vector representations reside in a similarity space in which perceptual similarity corresponds to proximity. By using the similarity structure of natural input, the model by-passes assumptions about distributional statistics of real-world input. Our simulations on the list-strength effect, the list-length effect, and the false memory effect support the validity of the proposed model. In particular, we conducted a face recognition simulation with the NIM model and found that it is able to replicate well-established recognition memory effects that relate to the similarity of the input.

Memory Representation

Many computational memory models represent an item by a vector of abstract features (e.g., the SAM model, Raaijmakers & Shiffrin, 1981; the REM model, Shiffrin & Steyvers, 1997, the model of differentiation, McClelland & Chappell, 1998). The feature values are usually drawn from a mathematical distribution (e.g., a geometric distribution). Since the computational models artificially generate vector representations, they do not address the contribution of the similarity structure intrinsic to natural data. However, we believe that the similarity structure contains important information. Therefore, we propose a memory model that operates on natural data and represents the similarity structure of these data.

The similarity structure of natural data can be represented in any type of space that fulfills the compactness criterion (Arkadev & Braverman, 1966). This criterion is fulfilled when similar objects in the real world are close in their representations. Several researchers developed so called 'similarity spaces', in which representations of similar items are in close proximity of each other (e.g., Nosofsky, 1986; Steyvers, Shiffrin, & Nelson, in press). An analysis of human similarity judgments or of free association data often forms the basis of a similarity space. However, we propose to derive the similarity space from the natural data by employing a biologically-informed transformation.

In the next section, a new recognition memory model that operates on natural images is introduced and described. We call this model the natural input memory (NIM) model. We will conduct a face recognition simulation with the NIM model and will evaluate its ability to replicate findings from recognition-memory studies. Finally our main conclusion will be given.

The NIM Model

The NIM model encompasses the following two stages.

1. A perceptual pre-processing stage that translates a natural image into a number of feature vectors.
2. A memory stage comprising two processes:
   (a) a storage process that simply stores feature vectors;
   (b) a recognition process that compares feature vectors of the image to be recognized with previously stored feature vectors.

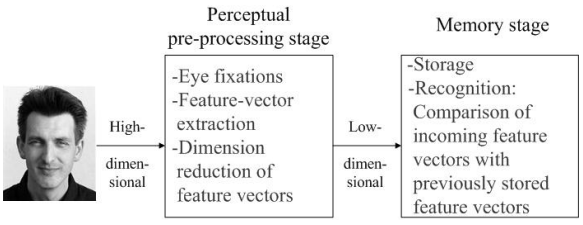


Figure 1: The natural input memory (NIM) model.

Figure 1 presents a schematic diagram of the NIM model. The face image is an example of a natural image. The two boxes correspond to the perceptual pre-processing stage and the memory stage.

The Perceptual Pre-Processing Stage

In this section, we first provide some background on the sources of biological inspiration and on the computational considerations. Then, we discuss some relevant implementation details.

Biological Inspiration and Computational Considerations
The human visual system is our main source of biological inspiration. The eye sequentially fixates on those parts of a visual scene that are most informative for recognition (e.g., Yarbus, 1967). Early visual processing in the brain leads to the activation of millions of optic nerve cells (Palmer, 1999). The nerve-cell activations may be conceived as a high dimensional vector. The high dimensionality enables the representation of a large amount of information without suffering from interference (Rao & Ballard, 1995), but it also hampers the memory performance, as the number of examples

that is necessary for a reliable generalization performance grows exponentially with the number of dimensions. This phenomenon is known as the 'curse of dimensionality' (Bellman, 1961; Edelman & Intrator, 1997). In coping with the curse of dimensionality, subsequent stages in the visual system are assumed to reduce the dimensionality of the high-dimensional input (e.g., Hubel, 1988; Tenenbaum, Silva, & Langford, 2000). The assumption is supported by findings of Edelman and Intrator (1997), who showed that the human visual system is able to retrieve the intrinsic low-dimensional structure of the high-dimensional visual input.

In the NIM model, dimension reduction of high-dimensional natural input is achieved in two sequential steps: (1) a biologically-informed feature-vector extraction (Freeman & Adelson, 1991) followed by (2) a principal component analysis (Pearson, 1901). The feature-vector extraction method employed by the NIM model operates directly on a high-dimensional natural image. The image has a high dimensionality because it is treated as a vector, the elements of which are the constituent pixel values. Motivated by eye fixations in human vision, the feature-vector extraction method takes samples from randomly-selected locations along the contours in the image. To emphasize the parallel with human vision, we refer to the samples as 'fixations'. For each fixation, the NIM model extracts features (i.e, a feature vector) from the image area centered at the fixation location. Since the feature vector contains a limited number of features, it is of a much lower dimensionality than the image. The feature-vector extraction method is based on the visual processing generally believed to occur in the visual area V1. The responses of neurons in V1 are modeled by a multi-scale wavelet decomposition (described later). Several studies showed that the biologically-informed multi-scale wavelet decomposition results in a representation space that accurately represents similarities as perceived by humans (e.g., Kalocsai, Zhao, & Biederman, 1998; Lyons & Akamatsu, 1998; Bartlett, Littleworth, Braathen, Sejnowski, & Movellan, 2003). After extraction of feature vectors, principal component analysis represents the feature vectors by their projection onto a number of orthogonal basis vectors which are ordered according to the amount of variance they explain. The dimensionality of the feature vectors is reduced by taking the projection onto the first p basis vectors. The low-dimensional feature vectors reside in a similarity space where visual similarity translates to proximity of feature vectors. Translating a two-dimensional image using a multi-scale wavelet decomposition followed by a principal component analysis, is an often applied method in the domain of visual object recognition to model the first three stages of processing of information in the human visual system (i.e., retina/LGN, V1/V2, V4/LOC; Palmeri & Gauthier, 2004). In contrast, existing memory models lack such a pre-processing method and often make simplifying assumptions about object representations.

Implementation The input image is translated into a multi-scale representation at four spatial scales. At every scale, the image is processed by four oriented filters in the orientations 0°, 45°, 90°, and 105° using the steerable-pyramid transform (Freeman & Adelson, 1991). This processing results in sixteen (four scales times four orientations) filtered images. From each of the sixteen images a 7×7 window is selected at a fixation point and the 16×49 pixel values are placed in a vector. In addition, the pixel values of a 7×7 low-resolution subimage centered at the fixation point are appended to the vector. Fixation points are randomly drawn from the contours of the face. The result is a feature vector for each fixation. As mentioned before, a principal component analysis was used to reduce the dimensionality of the feature vectors by taking the projection onto the first p basis vectors.

The Memory Stage

The Storage Process In the NIM model, the storage process straightforwardly stores an item (i.e., a pre-processed natural image). A pre-processed natural image is represented by a number of low-dimensional feature vectors in the similarity space, each corresponding to an eye fixation. The storage strength, S, is defined as the number of feature vectors stored for an image.

The Recognition Process In the NIM model, the recognition process determines the familiarity of an image to be recognized by comparing feature vectors of the image to be recognized with previously stored feature vectors. Models with a recognition process based on comparing items to previously stored exemplars can provide an accurate quantitative account of recognition performance (Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky, Clark, & Shin, 1989). In the NIM model, the recognition process uses a nearest neighbor classifier method, which takes each feature vector of the image to be recognized and then determines the number of previously stored feature vectors, f, that fall within a hypersphere with radius r, centered around the feature vector of the image. The familiarity, F, of the image is defined as $\sum f_i/T$, with f_i the value of f for the i^{th} feature vector of the image, and T the total number of feature vectors of the image.

We expect that the similarity-space representations employed by the NIM model will deepen our understanding of human recognition memory. Moreover, they may effectively support a number of memory effects often obtained in recognition memory studies. The latter studies are described in the next section.

Human Recognition Memory Studies

Three recognition memory effects often found in recognition memory studies are: the list-strength effect, the list-length effect, and the false memory effect. In general, recognition memory studies provide subjects with a study list of items and test their recognition memory for (some of) the studied items (i.e., targets) and a number of non-studied items (i.e., lures). We will emphasize the relation between the similarity structure of the targets and the lures used in the experiments on the one hand and the memory effects on the other hand.

The List-Strength Effect

A list-strength effect is defined as: a decrease in memory performance for a given set of study list items when other items of the study list are "strengthened" (i.e., the amount of time or the number of times the items are studied is increased) (Ratcliff, Clark, & Shiffrin, 1990). While some researchers failed

to find a list-strength effect for recognition memory (e.g., Ratcliff et al., 1990), recent findings showed that a list-strength effect can be obtained when there is a high degree of similarity between targets and lures. Norman (2002) tested whether strengthening some words of the study list affected a subject's recognition performance for other (non-strengthened) studied words. In the experiments, a significant list-strength effect was obtained only when targets and lures were similar. For dissimilar targets and lures, no list-strength effect was found. Moreover, recognition scores were significantly higher for dissimilar targets and lures than for similar targets and lures.

The List-Length Effect

A list-length effect is defined as: a decrease in memory performance for the items of the study list when additional items are added to the study list (Ratcliff et al., 1990). List-length studies yielded contradictory results. While some researchers failed to find a list-length effect (e.g., Dennis & Humphreys, 2001), others did obtain it (e.g., Cary & Reder, 2003). Recent experimental results indicate that the similarity between targets and lures can affect the degree to which a list-length effect occurs (MacAndrew, Klatzky, Fiez, McClelland, & Becker, 2002). In a study examining the effect of phonological similarity on recognition memory, MacAndrew et al. (2002) tested subjects' recognition memory for letters on a study list of four or six letters. The results showed that a larger list-length effect occurred for similar targets and lures than for dissimilar targets and lures. Also, overall recognition scores were higher for dissimilar targets and lures than for similar targets and lures.

The False Memory Effect

A number of experimental studies showed a false memory effect, which holds that the recognition of a lure (i.e., a false memory or a false alarm) is more likely to happen when the lure is similar to (one of the) studied items (e.g., Postman, 1951; Dewhurst & Farrand, 2004). For instance, the results by Dewhurst and Farrand (2004) show that the number of false memories increases together with the number of targets on the study list that are similar to the lures.

In a similarity space, representations of similar targets and lures show more overlap than representations of dissimilar targets and lures. Similar targets and lures are thus more difficult to discriminate than dissimilar targets and lures. Therefore, we expect that list-strength effects and list-length effects will be more pronounced and there will be more false alarms when targets and lures are similar than when targets and lures are dissimilar.

We hypothesize that the similarity structure of the perceived targets and lures can give rise to the recognition-memory effects discussed above. To test this hypothesis, we conducted a face recognition simulation with the NIM model, which employs similarity-space representations of perceived natural images.

Simulation

In our simulation, we investigated the ability of the NIM model to produce the following effects: (1) the effect of the similarity between targets and lures on the list-strength effect, (2) the effect of the similarity between targets and lures on the list-length effect, and (3) the false memory effect. The NIM model was repeatedly provided with a study list of face images and tested for recognition of the studied images (i.e., targets) and a number of non-studied images (i.e., lures). The images were gray-scale images of human faces taken from the FERET database (Phillips, Wechsler, Huang, & Rauss, 1998) of facial images. Male and female Caucasian faces without beards or glasses were selected. An example of such an image is shown on the left hand side of Figure 1. In this simulation, recognition memory was tested in two different conditions: (1) the dissimilar condition that employed lures dissimilar from the targets, and (2) the similar condition, that employed lures similar to one of the targets. In the NIM model, similar images are separated by a small distance in the similarity space. List-strength effects and list-length effects were assessed in both conditions and compared to determine whether the similarity between targets and lures had affected the degree to which the list effects occurred. Moreover, a comparison of the recognition results in the dissimilar condition and the similar condition revealed whether a false memory effect had occurred. Below we describe the calculation of recognition scores, the paradigms, the conditions, the procedure, and the results.

Calculation of Recognition Scores

The familiarity values, F, were used in a signal detection analysis to determine the recognition scores. The appropriate measure for the recognition score (d_a) was based on the normalized difference between the average F values of the targets ($\overline{F}(I_T)$) and the average F values of the lures ($\overline{F}(I_L)$):

$$d_a = \frac{\overline{F}(I_T) - \overline{F}(I_L)}{\sqrt{\frac{\sigma^2_{[F(I_T)]} + \sigma^2_{[F(I_L)]}}{2}}}$$

(Simpson & Fitter, 1973). Each d_a value was calculated on the basis of the familiarity values for targets ($F(I_T)$) and the familiarity values for lures ($F(I_L)$) of ten recognition tests.

Paradigms

The List-Strength Effect We used the mixed-pure paradigm first proposed by Ratcliff et al. (1990). It is used in many list-strength studies. The mixed-pure paradigm employs three types of study lists: pure weak lists (N weak images), pure strong lists (N strong images), and mixed lists ($N/2$ strong and $N/2$ weak images). A list-strength effect is said to occur (1) when the recognition score for weak images on a pure list is higher than the recognition score for weak images on a mixed list or, (2) when the recognition score for strong images on a mixed list is higher than the recognition score for strong images on a pure list. The pure/mixed ratio for weak images (i.e., the recognition score for weak images on a pure list divided by the recognition score for weak images on a mixed list) thus is an indication for the degree to which a list-strength effect occurs for weak images. Likewise, the mixed/pure ratio for strong images is an indication for the degree to which a list-strength effect occurs for strong images.

The List-Length Effect A list-length effect is said to occur when the recognition score for images on a shorter list is higher than the recognition score for images on a longer list. To assess the occurrence of a list-length effect we compared recognition scores for images on study lists of different lengths.

The False Memory Effect A higher false alarm rate (together with no difference in the hit rate) for the similar condition than for the dissimilar condition is said to indicate the occurrence of a false memory effect. However, using the general performance measure d_a (as described in the previous subsection) to determine recognition scores, the NIM model produces no false memories (and thus no false memory effect), simply because no recognition decisions are made. Most computational memory models, however, make recognition decisions based on the comparison of an obtained familiarity value to a given criterion (e.g., Busey, 2001). When the familiarity value exceeds the criterion, the item is recognized, if not, the item is not recognized. To assess the false memory effect, we also applied a decision criterion to the familiarity values, F, obtained for the dissimilar condition and for the similar condition. As a criterion we used: $C = S \times (0.02 + N/500)$, with S the storage strength of the images, and N the number of images on the study list.

Conditions

We distinguished two conditions: the dissimilar and the similar condition. For the dissimilar condition, recognition performance for targets versus dissimilar lures was tested. Targets and lures were selected from a subset of dissimilar images. The images in the subset of dissimilar images were selected in such a way that the clusters of their feature vectors in the similarity space showed relatively little overlap. Hence, dissimilarity for a subset of images, D, is defined as: $\sum f_{B,A_i}/T_A \leq d_1, \forall A,B \in D$, with f_{B,A_i} the number of feature vectors of image B that fall within a hypersphere with radius r centered around the i^{th} feature vector of image A, T_A the total number of feature vectors of image A, and d_1 a dissimilarity constant. For the similar condition, recognition performance for targets versus similar lures was tested. Pairs of similar targets and lures were selected in such a way that the clusters of their feature-vector representations in the similarity space showed relatively much overlap. Hence, similarity for two images, A and B, is defined as: $\sum f_{B,A_i}/T_A \geq d_2$, with f_{B,A_i} the number of feature vectors of image B that fall within a hypersphere with radius r centered around the i^{th} feature vector of image A, T_A the total number of feature vectors of image A, and d_2 a similarity constant, with $d_2 > d_1$.

Procedure

We provided the NIM model with (1) pure weak study lists, (2) pure strong study lists, and (3) mixed study lists of lengths $N = 4$, 8, and 12, in both the dissimilar and the similar condition. Weak images were stored with storage strength $S = 5$ (i.e., five feature vectors were stored, corresponding to five fixations) and strong images were stored with storage strength $S = 10$. For each feature vector, the first $p = 50$ dimensions were stored. After the last image of a study list had been presented to the model, the N images of the study list (i.e., targets) along with N new images (i.e., lures) were presented for recognition. Lures in the dissimilar condition were selected with dissimilarity constant $d_1 = 0.26$ and lures in the similar condition were selected with similarity constant $d_2 = 0.8$. Recognition tests and the selection of targets and lures were performed using the radius parameter $r = 5.0$.

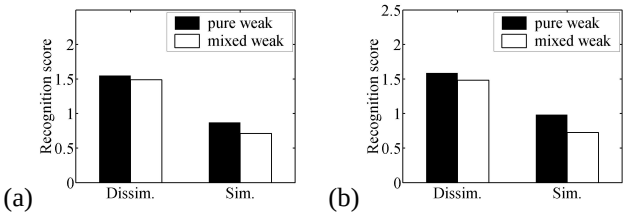


Figure 2: (a) Norman's (2002) results, and (b) Recognition scores (d_a) for weak images on pure lists (black bars) and on mixed lists (white bars) of length $N = 12$ for the dissimilar condition and for the similar condition.

Results

Table 1 presents the results for the dissimilar and similar conditions, respectively. The rows show the recognition results for lists of lengths $N = 4$, 8, and 12. The columns labeled w show the recognition scores for the weak images and the columns labeled s show the recognition scores for the strong images. The columns labeled pw/mw show the pure/mixed ratio for weak images and the columns labeled ms/ps show the mixed/pure ratio for strong images (both of which are indications of the degree to which a list-strength effect occurred). Figure 2(a) presents a bar graph of the results reported by Norman (2002) (described previously). Figure 2(b) presents a bar graph of the recognition scores produced by the NIM model in conditions analogous to the conditions in Norman's experiment. Since results were similar for lists of different lengths N, only the results for lists of length $N = 12$ are shown. A comparison of the graphs in Figures 2(a) and 2(b) reveals a close correspondence between Norman's results and the results produced by the NIM model.

Table 1: The average recognition scores produced by the NIM model for the dissimilar and the similar condition.

Dissimilar condition						
	Pure lists		Mixed Lists		ratios	
N	w	s	w	s	pw/mw	ms/ps
4	1.81	2.39	1.78	2.41	1.01	1.01
8	1.65	2.18	1.54	2.28	1.07	1.05
12	1.59	2.11	1.48	2.15	1.07	1.02
Similar condition						
	Pure lists		Mixed Lists		ratios	
N	w	s	w	s	pw/mw	ms/ps
4	1.38	1.83	1.14	1.97	1.21	1.08
8	1.12	1.52	0.87	1.78	1.29	1.17
12	0.98	1.35	0.73	1.61	1.36	1.20

Similarity and the List-Strength Effect List-strength effects for the dissimilar condition were significantly smaller than list-strength effects for the similar condition as indicated by the higher pw/mw and ms/mw values for the similar condition compared to those for the dissimilar condition. This was supported in a two-way analysis of variance (ANOVA) by the interaction between list type (pure or mixed) and condition. Calculated F values for lists of lengths $N = 4$, 8, and 12 ranged from $F(1,159) = 4.97$ to $F(1,159) = 9.62$, $p < 0.05$ for weak images and $F(1,159) = 4.52$ to $F(1,159) = 12.02$, $p < 0.05$, for strong images.

Similarity and the List-Length Effect The list-length effects for the dissimilar condition were significantly smaller than those for the similar condition. This was indicated in a two-way ANOVA by the interaction between list length and condition for pure weak lists, $F(2,239) = 4.61$, $p < 0.05$, and for pure strong lists, $F(2,239) = 3.68$, $p < 0.05$.

The False Memory Effect Table 2 presents the hit rates and false alarm rates for pure strong lists of lengths $N = 4$, 8, and 12 for both the dissimilar and the similar condition. Since the results were similar for pure weak lists and pure strong lists, we only present the results for pure strong lists. The results show that a false memory occurred: false alarm rates were higher for lists in the similar condition than for lists in the dissimilar condition (while hit rates were not significantly different). In an ANOVA, calculated F values for the false alarm rates ranged from 163.38 to 384.74, $p < 0.05$, while F values for the hit rates ranged from 2.08 to 2.24, $p > 0.05$.

Table 2: The average hit rates and false alarm rates produced by the NIM model for the dissimilar and the similar condition.

	Dissimilar condition		Similar condition	
N	Hit rate	F/A rate	Hit rate	F/A rate
4	0.84	0.01	0.86	0.14
8	0.76	0.02	0.78	0.17
12	0.69	0.02	0.70	0.15

Discussion

Based on recent experimental findings (Norman, 2002), we assumed that the degree to which a list-strength effect and a list-length effect occur varies with the degree of similarity between targets and lures. The NIM model produces this effect, as well as a false memory effect. Below we discuss the single-process NIM model in relation to other memory models and the ability of the NIM model to simulate mirror effects.

Comparison to Other Models

The NIM model differs from existing memory models in that it operates on natural input and employs a single process for recognition.

A Perceptual Process Operating on Natural Input The NIM model encompasses a transformation that yields the similarity structure of natural images. So far, existing memory models have been tested with artificial data (e.g., the REM model, Shiffrin & Steyvers, 1997), with predefined similarity spaces (e.g., the SimSample model, Busey, 2001), or with synthesized natural images (Kahana & Sekuler, 2002). The predictions these models make for recognition memory performance can be similar to the predictions the NIM model makes, provided that a representation space is employed that accurately reflects the similarity structure of the input. However, these models fall short in constructing a representation in an *a priori* manner. In contrast, the NIM model remedies this. Therefore, we expect that the NIM model provides us with a useful tool for making predictions about the effects of varying similarity of natural input on memory.

Single versus dual-process models Several memory models assume two processes for recognition to explain recognition results. These dual-processing models assume that recognition involves (1) a familiarity process, i.e., a context insensitive automatic process, and (2) a recollection process, i.e., a context sensitive strategic process (see Yonelinas, 2002, for a review of dual-processing models). Norman (2002) explains his experimental findings on the similarity effect by a dual-processing approach. He argues that the degree to which a list-strength effect occurs depends on the extent to which recollection drives recognition. While there might be good biological evidence that more than one process is involved in recognition, our results show that a single straightforward process for recognition suffices to produce Norman's and other findings on recognition memory.

Mirror Effects

In addition to the list-strength effect and the list-length effect, memory models are often tested for two related effects consistently obtained in experimental studies: the strength-mirror effect and the length-mirror effect (e.g., Murnane & Shiffrin, 1991). Simulation results, reported elsewhere (Lacroix, Murre, Postma, & Herik, submitted), showed that the NIM model exhibits these effects.

Conclusion

We have seen that the NIM model is able to build a similarity space from perceived natural data. Moreover, it successfully replicated recognition findings on list-strength effects, list-length effects, false memory effects, and mirror effects. Though it is at present not clear to what extent these results emerge from the use of natural images, it does increase the validity of the model by by-passing assumptions about distributional statistics of real-world perceptual features. Future studies aim at extending the NIM model to simulate a wider variety of findings on recognition memory.

Acknowledgments

The research project is supported in the framework of the NWO Cognition Program with financial aid from the Netherlands Organization for Scientific Research (NWO). It is part of the larger project: 'Events in memory and environment: modeling and experimentation in humans and robots' (project number: 051.02.2002).

References

Arkadev, A. G., & Braverman, E. M. (1966). *Computers and pattern recognition.* Washington, DC: Thompson.

Bartlett, M. S., Littleworth, G., Braathen, B., Sejnowski, T. J., & Movellan, J. R. (2003). A prototype for automatic recognition of spontaneous facial actions. In S. Becker & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15). Cambridge, MA: The MIT Press.

Bellman, R. (1961). *Adaptive control processes: a guided tour.* Princeton, NJ: Princeton University Press.

Busey, T. (2001). Formal models of familiarity and memorability in face recognition. In M. Wenger & J. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges.* Hillsdale, NJ: Erlbaum Associates.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*, 231-248.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452-478.

Dewhurst, S. A., & Farrand, P. (2004). Investigating the phenomenological characteristics of false recognition for categorised words. *European Journal of Cognitive Psychology*, *16*, 403-416.

Edelman, S., & Intrator, N. (1997). Learning as extraction of low-dimensional representations. In R. Goldstone, D. Medin, & P. Schyns (Eds.), *Mechanisms of perceptual learning* (Vol. 36, pp. 353-380). San Diego, CA: Academic press.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *13*, 891-906.

Hubel, D. H. (1988). *Eye, brain, and vision.* New York, NY: WH Freeman.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research*, *42*, 2177-2192.

Kalocsai, P., Zhao, W., & Biederman, I. (1998). Face similarity space as perceived by humans and artificial systems. In *Proceedings, third international conference on automatic face and gesture recognition* (pp. 177-180). Nara Japan.

Lacroix, J. P. W., Murre, J. M. J., Postma, E. O., & Herik, H. J. van den. (submitted). Modeling recognition memory using the similarity structure of natural input. *Psychological Review*.

Lyons, M., & Akamatsu, S. (1998). Coding facial expressions with gabor wavelets. In *Proceedings, third international conference on automatic face and gesture recognition* (pp. 200-205). Nara Japan.

MacAndrew, D. K., Klatzky, R. L., Fiez, J. A., McClelland, J. L., & Becker, J. T. (2002). The phonological-similarity effect differentiates between two working memories. *Psychological Science*, *13*, 465-468.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724-760.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*, 855-874.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 1083-1094.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 282-304.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* Cambridge, MA: The MIT Press.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*, 291-303.

Pearson, K. (1901). On lines and planes of closest fit to systems of points is space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, *2*, 559-572.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing Journal*, *16*, 295-306.

Postman, L. (1951). The generalization gradient in recognition memory. *Journal of Experimental Psychology*, *42*, 231-235.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93-134.

Rao, R. P. N., & Ballard, D. H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 461-505.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). The list-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*, 163-178.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem: Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin*, *80*, 481-488.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (in press). Word association spaces for predicting semantic similarity effects in episodic memory. In A. Healy (Ed.), *Cognitive psychology and its applications: Festschrift in honor of lyle bourne, walter kintsch, and thomas landauer.* Washington, DC: American Psychological Association.

Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319-2323.

Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum Press.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441-517.

Hierarchical Skills and Cognitive Architectures

Pat Langley (langley@csli.stanford.edu)
Kirstin Cummings (kirstinc@ccrma.stanford.edu)
Daniel Shapiro (dgs@stanford.edu)
Computational Learning Laboratory, CSLI
Stanford University, Stanford, CA 94305

Abstract

In this paper, we examine approaches to representing and utilizing hierarchical skills within the context of a cognitive architecture. We review responses to this issue by three established frameworks – ACT-R, Soar, and PRODIGY – then present an alternative we have developed within ICARUS, another candidate architecture. Unlike most earlier systems, ICARUS lets skills refer directly to their subskills and communicate within a single recognize-act cycle. This assumption has implications for the number of cycles required to complete complex tasks. We illustrate our approach with the domain of multi-column subtraction, then discuss related methods and directions for future work in this area.

Introduction and Overview

Human skills are organized in a hierarchical fashion. There seems to be general agreement with this claim, as it is consistent not only with experimental findings about the execution and acquisition of skills, but also with introspection about our everyday behavior. Upon request, most people can describe their complex skills at successive levels of aggregation, whether these involve how they drive to work each day, how they cook a meal, or how they write a technical article.

What remains an open question is how we should model such skill hierarchies in computational terms. Alternative approaches to modeling cognition encode the notion of hierarchy in distinct ways that have different implications for performance and learning. The most interesting positions are those which are embedded in theories of the human cognitive architecture, such as Soar (Laird et al., 1987), ACT-R (Anderson, 1993), and EPIC (Kieras & Meyer, 1997). These frameworks make strong commitments to both the mental representations of knowledge and to the processes that operate on them.

In the pages that follow, we consider the challenge of modeling hierarchical skills within a unified theory of cognition. We begin with a brief review of three such architectures and their responses to this issue, then turn to ICARUS, an alternative framework that approaches hierarchical skills from a different perspective. The key issue involves whether one can traverse levels of a hierarchy within a single cognitive cycle. Our illustrative example comes from a familiar cognitive skill that has a hierarchical organization – multi-column subtraction – but we also consider other models that incorporate multi-level skills. We conclude by discussing related work and our plans to extend the architecture's capabilities.

Previous Research on Hierarchical Skills

A cognitive architecture (Newell, 1990) specifies the infrastructure for an intelligent system, indicating those aspects of a cognitive agent that remain unchanged over time and across different application domains. Many proposals for the human cognitive architecture take the form of a production system, which stores long-term knowledge as a set of condition-action rules, encodes short-term elements as declarative list structures, and relies on a recognize-act cycle that matches against, and executes rules that alter, the contents of short-term memory. Soar, ACT-R, and EPIC are all examples of the production-system paradigm, although other architectural frameworks are also possible.

Despite the general agreement that cognitive skills are organized hierarchically, there exist different ways to implement this basic idea. Within a production-system architecture, the most natural scheme involves communication between skills and their subskills through the addition of elements to short-term memory. For instance, ACT-R achieves this effect using production rules that match against a generalized goal structure in their condition sides, such as the desire to prove that two triangles are congruent, and, upon firing, create new subgoals, such as the desire to prove that two lines have the same length. This approach requires the explicit addition of goal elements to short-term memory, since this is the only mechanism through which other production rules can be accessed.

Soar takes a somewhat different approach to encoding hierarchical skills that relies on multiple problem spaces. For instance, Jones and Laird (1997) report a detailed model of ying an aircraft in combat training scenarios. This system organizes its capabilities in a hierarchical manner, with each node implemented as a Soar problem space with associated states, operators, and goals. To invoke a lower-level problem space, the system adds a new element to short-term memory that refers to that space, and it must take similar action in order to exit. The details differ from those in ACT, but the passing of messages through short-term memory is similar.

The PRODIGY architecture (Minton et al., 1989) produces cognitive behavior in yet another manner. The system represents knowledge about actions as STRIPS-like operators, and the central module utilizes means-ends analysis to decompose problems into subproblems. This gives PRODIGY the ability to generate hierarchical

structures dynamically for each problem it encounters, and it can use control rules to select among candidate operators, states, and goals that determine the decompositions. However, these hierarchical structures do not remain in memory after a problem has been solved; instead, the system stores its experience in generalized control rules that let it reconstruct them for new problems. Again, this requires adding explicit elements to short-term memory that serve as mediators.

Thus, despite their diversity, these three frameworks share a basic assumption about how communication occurs among skills and subskills. This tenet has implications for the cognitive behavior of systems cast within them, including whether intermediate results are available, the time required to execute complex skills, and the effects of learning on hierarchical structure. We will see shortly that another response to this issue is possible.

Hierarchical Skills in ICARUS

ICARUS (Choi et al., in press) is a cognitive architecture that shares some central features with its predecessors. For instance, it relies on symbolic list structures to encode information, and it makes a clear distinction between long-term and short-term memories, with the former storing generic knowledge and the latter containing specific beliefs. Moreover, ICARUS assumes a recognize-act cycle, which relies on matching patterns in long-term memory against elements in the short-term store, to determine behavior over time. The framework also comes with a programming formalism for building knowledge-based systems.

However, ICARUS has other characteristics that distinguish it from earlier frameworks in significant ways. One such feature is its commitment to embodied cognition, which requires that all symbols in ICARUS programs be grounded in sensori-motor primitives.[1] A second key assumption is that each long-term memory structure may have not only a symbolic description but also an associated numeric function that computes its value as a function of sensory attributes. A third novel feature is that ICARUS contains distinct long-term memories for skills, which store its knowledge about action, and for concepts, which specify its knowledge about states and relations. Yet another contribution lies in the strong correspondence between long-term and short-term memory, specifically that every element in the latter must be an instance of some structure in the former.

A fifth important assumption, which is our focus here, is that hierarchical structures play a central role in cognition. Although production systems and related architectures allow hierarchies, many do not encourage them, and we maintain that ICARUS supports them in a sense that is stronger and more plausible psychologically than do frameworks like ACT and Soar. To understand this claim, we must first examine the architecture's representation for skills and the relations among them.

[1]Some recent architectural variants, like ACT-R/PM and EPIC-Soar, incorporate sensori-motor modules, but these were grafted onto systems based on theories of human problem solving, whereas ICARUS included them from the outset.

ICARUS skills bear a strong resemblance to production rules, but they have an even closer kinship to the operators in PRODIGY and STRIPS. Each skill has a name, arguments, and some optional fields. These include:

- a `:start` field, which encodes the situation that must hold to initiate the skill;
- a `:requires` field, which must be satisfied throughout the skill's execution across multiple cycles; and
- an `:effects` field, which specifies the desired situation the skill is intended to achieve.

For example, Table 1 shows a complete set of ICARUS skills for the domain of multi-column subtraction, including *borrow*, which has the objective of getting *?digit1* to be greater than *?digit2*. This skill can start only if *?digit2* is greater than *?digit1* and if a third element, *?digit3*, is nonzero. Moreover, its execution requires that *?digit1* be above *?digit2*, that *?digit3* be in the top row, and that *?digit3* be left of *?digit1*.

In addition, each ICARUS skill includes another field that specifies how to decompose it into subskills or actions. This may be either:

- an `:ordered` field, which indicates the agent should consider the components in a specific order;
- an `:unordered` field, which identifies a choice among skill components; or
- an `:actions` field, in which a primitive skill specifies one or more actions that are directly executable.

For example, *borrow* states that one should invoke *decrement* on one digit and call *add-ten* on another, in that order. These are both primitive skills that play the same role as STRIPS operators in a traditional planning system, with their `:start` fields serving as preconditions and their `:effects` fields specifying the desired results of execution.

The table also clarifies that ICARUS can specify multiple ways to decompose each skill in this manner, much as a Prolog program can include more than one Horn clause with the same head. Different decompositions of a given skill must have the same name, number of arguments, and effects. However, they can differ in their start conditions, requirements, and subskills. The skill *borrow* has two such expansions, one for borrowing from nonzero elements and another for borrowing across zero, which involves a recursive call to itself.

Each skill decomposition may also include a numeric function that encodes the utility expected if it executes this decomposition. This function is specified by a `:percepts` field that matches against the values of objects' attributes and a `:value` field that states an arithmetic combination of these quantities. The expected utility for a skill decomposition is a linear function of the numeric descriptors matched by that skill. Such functions are not required when the available skills specify deterministic behavior, as do those in the table, but we have used them in other domains and we mention them here for completeness.

We should note that ICARUS also organizes its long-term conceptual memory in a hierarchy, with higher-level concepts being defined in terms of lower-level ones, which

Table 1: ICARUS skills for multi-column subtraction.

```
(subtract ()
 :percepts ((digit ?digit))
 :requires ((top-row ?digit) (right-col ?digit))
 :unordered ((process ?digit))

(subtract ()
 :requires ((processed ?digit1) (top-row ?digit2)
            (left-of ?digit2 ?digit1))
 :unordered ((process ?digit2))

(process (?digit)
 :ordered  ((borrow ?digit)
            (find-difference ?digit))
 :effects  ((processed ?digit)))

(borrow (?digit1)
 :percepts ((digit ?digit2) (digit ?digit3))
 :start    ((greater ?digit2 ?digit1)
            (nonzero ?digit3))
 :requires ((above ?digit1 ?digit2)
            (top-row ?digit3)
            (left-of ?digit3 ?digit1))
 :ordered  ((decrement ?digit3)
            (add-ten ?digit1))
 :effects  ((greater ?digit1 ?digit2)))

(borrow (?digit1)
 :percepts ((digit ?digit2) (digit ?digit3))
 :start    ((greater ?digit2 ?digit1)
            (zero ?digit3))
 :requires ((above ?digit1 ?digit2)
            (top-row ?digit3)
            (left-of ?digit3 ?digit1))
 :ordered  ((borrow ?digit3))
 :effects  ((greater ?digit1 ?digit2)))

(decrement (?digit)
 :percepts ((digit ?digit val ?val))
 :start    ((nonzero ?digit))
 :actions  ((*cross-out ?digit)
            (*decrement ?digit ?val))
 :effects  ((crossed-out ?digit)))

(add-ten (?digit1)
 :percepts ((digit ?digit1 val ?val1)
            (digit ?digit2) (digit ?digit3))
 :start    ((above ?digit1 ?digit2)
            (top-row ?digit3)
            (left-of ?digit3 ?digit1)
            (crossed-out ?digit3))
 :actions  ((*add-ten ?digit1 ?val1))
 :effects  ((greater ?digit1 ?digit2)))

(find-difference (?digit1)
 :percepts ((digit ?digit1 col ?col val ?val1)
            (digit ?digit2 col ?col val ?val2))
 :start    ((above ?digit1 ?digit2)
            (greater-or-equal ?digit1 ?digit2))
 :actions  ((*add-difference ?col ?val1 ?val2))
 :effects  ((processed ?digit1)))
```

ultimately connect to perceptual elements. The literals that appear in the `:start`, `:requires`, and `:effects` fields must be defined concepts, thus linking the skill and conceptual memories in an interleaved manner. The architecture also incorporates three short-term memories. These include a perceptual buffer, updated on each cycle, that contains descriptions of perceived objects, a conceptual short-term memory that contains inferences derived from the perceptual buffer, and an intention memory that contains instances of skills the system intends to execute. The elements in these memories are simple literals but, because their predicates correspond to hierarchical structures in long-term memory, they encode substantial information about the agent's beliefs and intentions.

Distinctive Aspects of ICARUS Hierarchies

From the preceding discussion, it should be clear that ICARUS utilizes a more structured representation of knowledge than traditional cognitive architectures, but the implications of this structure depend directly on the processes that operate over them. Together, they enable cognitive processing that exhibits important differences from that in older frameworks.

One such characteristic involves the ability of ICARUS' skills to reference subskills by their names, rather than through the indirect references used in Soar, ACT, and PRODIGY. For example, the *borrow* skill in Table 1 calls directly on *decrement* and *add-ten*. ICARUS' approach has some aspects of subroutine calls in procedural programming languages but, when combined with multiple expansions for each skill (such as two variants for *borrow*), effectively embeds these calls within an AND/OR search. This makes our formalism a close relative of logic programming languages like Prolog, which uses a very similar syntax to support logical reasoning. But like other cognitive architectures, ICARUS is concerned with agents that exist over time, so it situates these computations within a recognize-act cycle.

As a result, ICARUS retains the overall avor of a production system but gains the ability to invoke subskills directly, rather than through the creation of short-term memory elements. This lets it execute complex skills in a top-down manner without having to descend through the hierarchy one step at a time. ICARUS can take advantage of this hierarchical organization without requiring the generation of explicit intermediate goal structures that are needed by production systems.

Recall that ICARUS includes a short-term memory that contains skill instances the agent considers worth executing. On each cycle, for each skill instance, the architecture retrieves all decompositions of the general skill and checks whether they are applicable. A skill is applicable if, for its current variable bindings, its `:effects` field does not match, the `:requires` field matches, and, if the system has not yet started executing it, the `:start` field matches the current situation. Moreover, at least one of its subskills must also be applicable. Since this test is recursive, a skill is only applicable if there exists at least one acceptable path downward to executable actions.

For each such path, the architecture computes the expected value and selects the candidate with the highest utility for execution. For a given path, it uses the value function stored with each skill and the numeric values matched in that skill's `:percepts` field to calculate the expected value at each level, summing the results along the path to compute the overall score. For instance, for the path *((subtract), (process digit11), (borrow digit11), (decrement digit21))*, the system would sum the expected value for all four levels to determine the utility of decrementing. This means that the same action can

have different values on a given cycle depending on which higher-level skill invokes it, providing a natural way for the hierarchy to incorporate the effect of context.

The architecture treats a skill expansion differently depending on whether its components appear in an `:unordered` set or an `:ordered` list. If they are unordered, the module considers each of the subskills and selects the one that yields the highest scoring subpath. If they are ordered, it instead treats the list as a reactive program that considers each subskill in reverse order. If the final subskill is applicable, then it expands further only down paths that include that subskill. Otherwise, it considers the penultimate skill, the one before that, and so forth. The intuition is that the subskills are ordered because later ones are closer to the parent skill's objective, and thus should be preferred when applicable.

We should clarify that ICARUS' consideration of alternative paths through its skill hierarchy does not involve generative planning. On each cycle, the architecture finds the best pathway through a set of exible but constrained structures. The process is much closer to the execution of a hierarchical task network (e.g., Myers, 1999) than to the construction of a plan from primitive operators. Such computation can be done efficiently within a single recognize-act cycle, at least for well-structured skill hierarchies. One can craft ICARUS programs that are expensive to evaluate, but the same holds for production systems with complex conditions on their rules.

An Illustrative Example

We can best clarify ICARUS' operation with an example that uses the skills from Table 1 on the subtraction problem 305 147. The system interacts with a simulated environment that, on each cycle, deposits perceptual elements such as *(digit digit11 col 1 row 1 val 5)* and *(digit digit12 col 1 row 2 val 7)* into the perceptual buffer. A conceptual recognition process draws inferences from these elements, such as *(greater digit12 digit11)* and *(above digit11 digit12)*, which it adds to conceptual short-term memory.

The model also begins with the single top-level intention *(subtract)*, which focuses cognitive behavior on the skills in the table even if others are present in long-term memory. On the first cycle, the system would consider both expansions of *subtract*, selecting the first one and binding *?digit* to *digit11*, the object in the top row and right column. This skill instance would in turn invoke *(process digit11)*, which would consider its subskills *find-difference* and *borrow*. Only the latter skill has its `:start` and `:requires` fields met, specifically by its second expansion, which handles situations that require borrowing from zero.

This skill instance, *(borrow digit11)*, then invokes itself recursively, producing the call *(borrow digit21)*, where the argument denotes the digit 0 in the top row and second column. Because the digit to its left is nonzero, this instantiation can only utilize the first expansion of *borrow*, which in turn calls on *(decrement digit31)* in its `:ordered` field, since its preconditions are satisfied, but those for *add-ten*, which occurs later in this ordering, are not. Because *decrement* is a primitive skill, it becomes the terminal node for an acceptable path through the skill hierarchy. Also, because this is the only such path ICARUS finds, it executes the instantiated actions *(*cross-out digit31)* and *(*decrement digit31 3)*.

These actions alters the environment and lead to another execution cycle. This time ICARUS again finds a single acceptable path that shares all but the last skill instance with that from the first round. The difference is that *digit31* has been crossed out, making *(decrement digit31)* inapplicable but enabling the skill instance *(add-ten digit21)*. Again this is the only acceptable path through the hierarchy, so ICARUS executes the action associated with this primitive skill, thus altering the simulated display so that *digit21*'s value is increased by ten.

On the third cycle, the architecture again finds only one path, in this case *((subtract), (process digit11), (borrow digit11), (decrement digit21))*, since the top number in the second column is no longer zero and can be safely decremented. This action enables execution of the path *((subtract), (process digit11), (borrow digit11), (add-ten digit11))* on the fourth cycle, after which (on the fifth cycle) the model selects the path *((subtract), (process digit11), (find-difference digit11))*, which writes the two digits' difference in the rightmost column.

This altered situation leads ICARUS to add the inference *(processed digit11)*, which on the sixth cycle causes it to select the second expansion of *subtract*; this invokes the skill instance *(process digit21)* on the revised top digit in the second row. Because this digit has already been incremented by ten, it is greater than the one below it, so the skill instance *(find-difference digit21)* is now applicable. Execution of this path produces an answer in the second column, which leads on the next cycle to processing of the third column and to an answer there as well. No paths are satisfied on additional cycles, so the system idles thereafter, waiting for new developments.

General Implications

The sample trace above does not illustrate all of ICARUS' capabilities, since it ignores details about the conceptual inference process and it does not utilize value functions. However, it should make clear that ICARUS operates over its skill hierarchy in a different manner than frameworks like Soar and ACT-R. They can model behavior on complex tasks in two distinct ways. One scheme assumes hierarchical rules or problem spaces, which require additional cycles for stepping downward through each level of the hierarchy. Another assumes that learning has produced compiled rules that eliminate the hierarchy and thus the need for intermediate goal structures. Such compilation methods have been used to explain both the power law of learning and reduced access to subgoals (e.g., Neves & Anderson, 1981).

However, it seems unlikely that the hierarchical structure of skills disappears entirely with practice, and ICARUS offers an account that avoids the above dichotomy. An architecture that traverses levels in a skill hierarchy within a single recognize-act cycle also predicts that intermediate structures will be inacessible, and the

power law follows from the construction of the hierarchy itself, which we discuss later. This account also makes different predictions than traditional schemes about the number of cycles, and thus the time, required to accomplish tasks with hierarchical and recursive structures.

Again, we can use multi-column subtraction to illustrate this point. A standard production-system model for this domain, like that described by Langley and Ohlsson (1984), finds the difference between two numbers in a column when the top one is larger but otherwise adds a goal to process or borrow from the adjacent column. Analysis reveals that such a model will take $5 \cdot b + 2 \cdot n$ cycles to complete a problem with b columns that require borrowing and n columns that do not. In contrast, the ICARUS model in Table 1 requires $3 \cdot b + 2 \cdot n$ cycles on the same problems. The two frameworks both indicate that solution time will increase with the number of borrow columns, but they predict quite different slopes.

Experiments with human subjects should reveal which alternative offers a better explanation of skilled behavior in this arena. We have not yet carried out such studies, but to test our framework's generality, we have developed ICARUS models for behavior in other domains that appear to have hierarchical organizations. One involves the well-known Tower of Hanoi puzzle, which can be solved using a hierarchical strategy. Our model for behavior on this task includes three primitive skills for lifting, lowering, and moving a disk sideways, along with one high-level skill for moving a disk to a target peg that, in two expansions, calls itself recursively. However, the Tower of Hanoi is like subtraction in that the environment changes only when the agent takes some action.

To ensure that ICARUS can also support behavior in more dynamic domains, we have developed two additional models. One involves hierarchical skills for balancing an upright pole by moving its lower end back and forth. This system includes two high-level skills with knowledge about the order in which the four primitive skills should be invoked. As described elsewhere (Choi et al., in press), we have also constructed a system that drives a vehicle and delivers packages in a simulated in-city environment. This model includes 46 skills that are organized in a hierarchy five levels deep. The high-level skills let the agent drive straight in lanes, get into rightmost lanes, slow for intersections, drive through intersections, turn at intersections, and make U turns. These terminate in actions for changing speed and altering the wheel angle. Another 13 skills support the delivery of packages to target addresses.

We have not attempted to compare the details of these systems' operations to human behaviors on the same tasks. Nor have we attempted to show that ICARUS produces more robust behavior than programs cast in earlier frameworks like Soar and ACT-R. Rather, our goal has been to demonstrate the same broad functionality as we observe in humans, including their apparent organization of complex skills into hierarchies. We have also aimed to show that ICARUS constitutes a viable point in the space of cognitive architectures, which we believe has been too thinly explored to date.

Related Efforts and Future Research

Although ICARUS incorporates a number of features that distinguish it from typical cognitive architectures, some related ideas have appeared elsewhere under different guises. For instance, our framework has much in common with the 'reactive planning' movement, which often utilizes hierarchical procedures that combine cognition, perception, and action in physical domains.[2] Examples include PRS (Georgeoff et al., 1985), teleoreactive programs (Nilsson, 1994), and the 3T robotic architecture (Bonasso et al., 1997), and some case-based planners (e.g., Hammond, 1993) embody similar notions.

However, within this paradigm, only Freed's (1998) APEX has been proposed as a candidate architecture for human cognition. This framework shares ICARUS' commitment to hierarchical skills, but it has a more procedural flavor and it does not incorporate a separate conceptual memory or enforce a correspondence between long-term and short-term structures. Another kindred spirit is Albus and Meystel's (2001) RCS architecture, which organizes knowledge hierarchically and makes a clear distinction between logical structures and value judgments. ICARUS and RCS share many common features, but they also retain many differences due to their origins in cognitive modeling and control theory, respectively.

We should clarify that, as a cognitive architecture, ICARUS still requires some development. The framework lacks many processing assumptions that would make it more plausible as a general theory of human behavior. For instance, it lacks any limits on perceptual bandwidth that require attention, which arises even on routine tasks like subtraction. We intend to model such behavior by introducing an action that focuses the agent's attention on an object and deposits its description in the perceptual buffer, with ICARUS being able to apply this action to only one visible object per cycle. This should produce longer subtraction runs that require additional steps for attentional events, much as in Anderson, Matessa, and Lebiere's (1997) ACT-R model of visual attention.

We should also extend our models for subtraction and other domains to handle lower levels of behavior more faithfully. The skills in Table 1 terminate with actions for decrementing, adding ten, and finding a difference, but these can be further decomposed into subskills for writing specific digits and even drawing individual lines. Our claim to model embodied cognition would be stronger if we extended the skill hierarchy downward in this fashion. We should also extend the hierarchy upward to model choices about which problem to tackle when many are present on the page. Such an expanded system would model subtraction behavior in a more complete way than do most accounts.

However, a more important omission concerns the origin of ICARUS' hierarchical skills. To handle this, we hypothesize a mechanism that is related to chunking in Soar and production composition in earlier versions of ACT. Although humans prefer to use routine behaviors when possible, they can, within limits, combine

[2] Our model of subtraction skills also has similarities to VanLehn's (1990) hierarchical treatment of this domain.

knowledge elements when needed to solve novel problems. Means-ends analysis has been implicated in such situations, so we plan to incorporate this method into future versions of the architecture. The generalized results of means-ends problem solving would be cached as a new skill. However, unlike chunking and composition, which produce at rules that eliminate structure, ICARUS would store a new hierarchical skill that refers to the original ones as components. This method should lead to the construction of skill hierarchies in a gradual, bottom-up manner as an agent learns about a domain.

Concluding Remarks

In closing, we should review the main claims we have made about hierarchical skills and their treatment within various cognitive architectures. There seems to be general agreement that skills are organized in some hierarchical fashion, but most existing models implement the invocation of subskills through goal structures that are deposited in short-term memory. In contrast, ICARUS produces hierarchical behavior by letting skills communicate directly with their subskills, as in procedural languages and logic programming formalisms.

We illustrated this idea by presenting an ICARUS model for multi-column subtraction and tracing its behavior on a specific problem. We saw that this system takes the same basic steps as a production-system model, but that steps involved in traversing the skill hierarchy occur within a single recognize-act cycle rather than across successive cycles. This theoretical difference leads to different predictions about the time required to execute complex skills. Future research should test these predictions and extend ICARUS to incorporate mechanisms for attention and construction of skill hierarchies.

Acknowledgements

This research was funded in part by Grant IIS-0335353 from the National Science Foundation, by Grant NCC 2-1220 from NASA Ames Research Center, and by Grant HR0011-04-1-0008 from Rome Labs. Discussions with Stephanie Sage, David Nicholas, and Seth Rogers contributed to many of the ideas presented in this paper.

References

Albus, J. S., & Meystel, A. M. (2001). *Engineering of mind: An introduction to the science of intelligent systems*. New York: John Wiley.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its application to visual attention. *Human-Computer Interaction*, *12*, 439–462.

Bonasso, R. P., Firby, R. J., Gat, E., Kortenkamp, D., Miller, D., & Slack, M. (1997). Experiences with an architecture for intelligent, reactive agents. *Journal of Experimental and Theoretical Arti cial Intelligence*, *9*, 237–256.

Choi, D., Kaufman, M., Langley, P., Nejati, N., & Shapiro, D. (in press). An architecture for persistent reactive behavior. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*. New York: ACM Press.

Freed, M. (1998). Managing multiple tasks in complex, dynamic environments. *Proceedings of the National Conference on Arti cial Intelligence* (pp. 921–927). Madison, WI: AAAI Press.

Georgeff, M., Lansky, A., & Bessiere, P. (1985). A procedural logic. *Proceedings of the Ninth International Joint Conference on Arti cial Intelligence*. Los Angeles: Morgan Kaufmann.

Hammond, K. (1993). Toward a theory of agency. In S. Minton (Ed.) *Machine learning methods for planning*. San Francisco: Morgan Kaufmann.

Jones, R. M., & Laird, J. E. (1997). Constraints on the design of a high-level model of cognition. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 358–363). Stanford, CA: Lawrence Erlbaum.

Kieras, D., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, *12*, 391–438.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Arti - cial Intelligence*, *33*, 1–64.

Minton, S., Carbonell, J. G., Knoblock, C. A., Kuokka, D., Etzioni, O., & Gil, Y. (1989). Explanation-based learning: A problem solving perspective. *Arti cial Intelligence*, *40*, 63–118.

Myers, K. L. (1999). CPEF: A continuous planning and execution framework. *AI Magazine*, *20*, 63–70.

Newell, A. (1990). *Uni ed theories of cognition*. Cambridge, MA: Harvard University Press.

Nilsson, N. (1994). Teleoreactive programs for agent control. *Journal of Arti cial Intelligence Research*, *1*, 139–158.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Langley, P., & Ohlsson, S. (1984). Automated cognitive modeling. *Proceedings of the Fourth National Conference on Arti cial Intelligence* (pp. 193–197). Austin, TX: Morgan Kaufmann.

Shapiro, D., Langley, P., & Shachter, R. (2001). Using background knowledge to speed reinforcement learning in physical agents. *Proceedings of the Fifth International Conference on Autonomous Agents* (pp. 254–261). Montreal: ACM Press.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

The Role of Prior Learning in Biasing Generalization in Artificial Language Learning

Jill Lany (lany@u.arizona.edu)
Department of Psychology, The University of Arizona, Tucson, AZ 85721-0068

Rebecca Gómez (gomez@u.arizona.edu)
Department of Psychology, The University of Arizona, Tucson, AZ 85721-0068

Abstract

Because unbiased learners are unlikely to arrive at the appropriate generalizations of their language (Gold, 1967), accounts of acquisition must examine the nature of learning biases. One form of bias is learners' prior learning experience. Adult participants familiarized with a category-induction language learned a language with the same underlying structure but novel vocabulary much more rapidly than naïve learners (Lany, Gómez, & Gerken, 2004). In the present experiments, we extend our investigations of prior learning experience by manipulating whether learners were initially exposed to fully- or partially-cued structure. Generalization is hindered by prior exposure to fully-cued structure, but enhanced by prior exposure to structure that is partially-cued. The results are important for understanding of the role of prior experience in constraining language acquisition.

Introduction

The syntax of natural languages is highly complex. Without corrective feedback, unconstrained learners are unlikely to converge on the grammar of their linguistic community (Gold, 1967). However, language learners regularly manage to discover the underlying patterns of their language while ignoring irrelevant structure, suggesting that they are constrained. Recent work in artificial language learning has begun to investigate forms of learning constraints, or ways in which learning processes might be guided.

Saffran (2002) demonstrated modality constraints on learning an artificial language by exposing adult learners to a phrase-structure language in which the presence of one item in a phrase predicted the presence of another item (Language P), or to a phrase-structure language in which predictive relationships were absent (Language N). Participants were presented with either an auditory or visual version of Language P or N. Learning of Language P was better when presentation was auditory. Visual language learners acquired the predictive and non-predictive languages equally well, suggesting that learners' sensitivity to patterns and relationships varies as a function of modality.

Recent work by Gómez (2002; Gómez, Welch, & Lany, 2003) speaks to another way learning may be guided. Gómez' studies suggest that learners selectively tune in to regularities by seeking out the most reliable structure in their input. While learners are highly sensitive to conditional probabilities of adjacent elements (e.g., Saffran, Aslin, & Newport, 1996), Gómez demonstrated that when conditional probabilities between adjacent elements are low or unreliable, learners attend to relationships between non-adjacent elements. These findings indicate that learners are biased to take the statistical reliability of a structure into account. Importantly, languages make use of various cues highlighting relevant structure, not just statistical ones[1].

Saffran's (2002) work suggests that learners are biased before they even begin to learn language. Gómez's work (2002; Gómez et al., 2003) suggests that statistical characteristics of the input itself can bias learning. Research also suggests learners can be biased by prior experience.

Saffran and Thiessen (2003) provided evidence that once infants form phonological generalizations based on regularities in their input, those generalizations influence how they parse new speech materials. Lany, Gómez, and Gerken (2004) demonstrated that generalization occurs at the abstract level of syntax-like structure. In these experiments, adults were exposed to a language consisting of categories of words, with restrictions on how categories could combine (see Braine, 1987; Frigo & MacDonald, 1998; Gerken, Wilson, & Lewis, in press; Gómez & Lakusta, in press). The language is composed of words belonging to the categories *a*, *b*, *X*, and *Y*. As in natural languages, the rules

[1] For example, languages are rich with prosodic cues to syntactic structure, and learners are sensitive to such cues beginning in early infancy (see Jusczyk, 1997 for an overview).

involve relationships between word categories. Specifically, the language has restrictions on how categories of different types can be combined within a string, such that *a* elements are paired with *X* elements and *b* elements with *Y*s, but not vice versa. (See Figure 1.)

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$
$\mathbf{a_1}$	a_1X_1	a_1X_2	a_1X_3	a_1X_4	a_1X_5
$\mathbf{a_2}$	a_2X_1	a_2X_2	a_2X_3	??	a_2X_5

	$\mathbf{Y_1}$	$\mathbf{Y_2}$	$\mathbf{Y_3}$	$\mathbf{Y_4}$	$\mathbf{Y_5}$
$\mathbf{b_1}$	b_1Y_1	b_1Y_2	b_1Y_3	b_1Y_4	b_1Y_5
$\mathbf{b_2}$	b_2Y_1	b_2Y_2	b_2Y_3	??	b_2Y_5

Figure 1. A typical *aX bY* paradigm. Learners are exposed to a subset of the grammatical pairings of markers and content-words then are tested for generalization to the withheld (??) pairings.

Learners exposed to 18 or 6 minutes of the *aX bY* language successfully acquired the language in the former but not the latter condition. However, learners exposed to 18 minutes of one language, and then transferred to a second language with the same underlying pattern, but none of the same words, learned the pattern with just 6 minutes of exposure.

The findings of Lany et al. (2004) demonstrate that learners do not remain the same over the course of acquisition. Rather, the learning process changes them, constraining the ways they perceive and learn about subsequent input. Thus, prior experience represents an additional constraint enabling learners to successfully acquire language.

Using this procedure, we can begin to investigate other ways learning might be facilitated by prior experience. In doing so, it is important to explore the extent to which this process might be useful in natural language acquisition.

In English, consistency in head direction results in a co-occurrence relationship between determiners and nouns, and also between auxiliaries and verbs. These relationships both involve a functional element preceding a lexical one, and restrictions on co-occurrences of categories of functional and lexical elements (similar to the *aX bY* structure used by Lany et al., 2004). Importantly, learners only truly acquire categories, and their co-occurrence restrictions, when there are cues indicating category membership (Frigo & McDonald, 1998; see also Gerken et al., in press). However, cues in natural language are variable, differing with strength according to category. For example, a corpus analysis (Lany et al., 2004) showed that in infant-directed speech, a greater proportion of nouns are marked with morpho-syntactic cues than verbs – e.g. nouns were fully cued by a determiner and a plural or diminutive ending 20% of the time and partially cued by either a determiner or an ending 60% of the time. Verbs were fully cued both by an auxiliary and inflectional ending 1% of the time and by one or the other 20% of the time. Thus languages contain different degrees of fully- and partially-cued structure.

In this study, we asked whether learners with prior exposure to a well-cued pattern have an advantage over naïve learners in acquiring a version of that pattern in which the cues highlighting relevant structure are diminished. In Experiment 1, we exposed a control group to 18 minutes of a partially-cued *aX bY* language, in which only 60% of the *X* and *Y* words had cues to category membership. We exposed an experimental group to 18 minutes of a fully-cued *aX bY* language, in which 100% of the *X* and *Y* words had cues to category membership, and then transferred them to 18 minutes of the partially-cued *aX bY* language. Interestingly, learners with prior exposure to a fully-cued language did not subsequently learn a partially cued language better than the control group. However, in Experiment 2, learners with prior exposure to a partially-cued language learned a second partially-cued language better than naïve learners. Why might this be the case? Learners initially exposed to the fully-cued language may have learned the perfectly predictive surface regularities resulting from the underlying structure, as opposed to the category relationships. These surface regularities were probabilistic in the partially-cued language, and thus a focus on this aspect of structure would not result in successful learning at transfer. In other words, perhaps learners of the fully-cued language were hindered by their focus on the surface regularities of the strings. Learners of the partially-cued language were not led to rely exclusively on a cue that would later be less reliable. We speculate on factors facilitating this group's transfer in the discussion.

Experiment 1

Method

Participants Ninety-five University of Arizona undergraduate students participated for course credit, forty-eight in the experimental condition.

Materials Learners in these experiments were exposed to category-induction languages of the form *aX bY*. We constructed both a fully-cued and a partially-cued *aX bY* language. The fully-cued language had two versions which shared the same underlying structure, but had none of the same words (two versions were necessary to test transfer). The vocabulary of each version of the language consisted of two *a* elements (*ong* and *rud* in Version A, and *ush* and *dak* in Version B) and two *b* elements (*alt* and *pel* in Version A, and *erd* and *vot* in Version B). The vocabulary of each version also consisted of six bisyllabic *X* words and six bisyllabic *Y* words. The *X* words all ended with the same syllable ("-ul" in Version A and "-it" in Version B). Similarly, all *Y* words shared the same final syllable ("-ee" in Version A and "-oo" in Version B. Additionally, each version had two variants, or grammars. The two grammars of a version were composed of the same set of words, but differed in how they were combined – strings from Grammar 1 (G1) took the form *aX bY* and strings from Grammar 2 (G2) took the form *aY bX*. For example, strings from Version A, G1were *ong bivul* and *erd suffee*, and strings from G2 were *ong suffee* and *erd bivul*. In Version B, G1 strings were *ush zamit* and *alt wifoo*, and G2 strings were *ush wifoo* and *alt zamit*. The fully-cued versions of the language consisted of 24 possible strings (12 *aX* strings and 12 *bY* strings), however, 4 strings (2 *aX* and 2 *bY* strings) were withheld from familiarization to be presented at test, so the familiarization set consisted of 20 strings.

The test materials for the fully-cued languages consisted of 16 strings, half grammatical and half ungrammatical. Four grammatical strings had been withheld during familiarization, four strings had been presented during training, and eight strings were ungrammatical strings (these were from the unheard language). Strings that were grammatical for one group of participants were ungrammatical for the other.

The materials for the partially-cued languages were the same as those of the fully-cued languages, with the exception that only 60% (four of six) of the *X* and *Y* words had endings cueing their category membership. Uncued words were bisyllabic, and each had a distinct ending (i.e. an ending that was not present on any of the other words, cued or uncued). Examples are *jeeloff*, *skyjer*, *bowda*, and *pefto*. These uncued words replaced four of the cued words from the fully-cued version of the language. The partially-cued versions of the language consisted of 24 possible strings (12 *aX* strings and 12 *bY* strings), however, 8 strings (4 *aX* and 4 *bY* strings) were withheld from familiarization to be presented at test, so the familiarization set consisted of 16 strings.

The test materials for the partially-cued languages consisted of 32 strings, half grammatical and half ungrammatical. Eight grammatical strings had been withheld from familiarization (four were cued and four were uncued). Eight strings had been presented during familiarization (four were cued, four were uncued). There were also 16 strings from the grammar of the unheard language.

The process underlying successful learning of this language is twofold (Braine, 1987; Frigo & MacDonald, 1998). Learners must first discover that there are different categories of words, which requires that words from different categories be differentiable based on their semantic or phonological characteristics. Once learners are sensitive to the categories, they can then learn that there are restrictions on how categories co-occur. Learners with knowledge of co-occurrence restrictions can generalize to novel combinations that respect these restrictions. When cues to category membership are present, generalizations can be accomplished through attention to the pairing of *a*s and *b*s with the endings of the *X* and *Y* words. Additionally, learners exposed to the partially cued language, can generalize to novel combinations involving uncued words by noting that if an *X*-word is paired with a particular *a* word, it can co-occur with other *a* words, but not with *b* words.

Procedure There were eight conditions in this experiment, resulting from the between-subjects manipulations of familiarization type (Transfer vs. Control), version (Version A vs. Version B), grammar (Grammar 1 vs. Grammar 2), Aside from instructions at the start of the familiarization phase, which were delivered by the experimenter, the entire experiment was conducted on a Hewlett Packard Brio PC running SuperLab 2.01 software.

In the Transfer condition, participants listened over headphones to 18 blocks (approximately 18 minutes) of randomly ordered strings from their fully-cued training language, and then answered two iterations of 16 test

questions, each in a different random order and separated from each other by a brief pause. Using the "Y" and "N" keys on their keyboard, participants made yes/no judgments on the grammaticality of each string. They then repeated this familiarization and test procedure for a partially-cued version of the language with novel vocabulary. In the Control condition, participants were familiarized with 18 blocks of the partially-cued language before test.

For the fully-cued language, participants sensitive to the *aX bY* structure should endorse grammatical test strings, including those withheld during training, more often than they endorse ungrammatical ones. Similarly, for participants familiarized with a partially-cued language, sensitivity to the *aX bY* structure would be indicated by higher endorsement rates to withheld grammatical strings (both cued and uncued) than to ungrammatical ones.

Results and discussion Preliminary analyses indicated that there were no differences in learning as a function of language version or grammar, so we collapsed across these variables. Mean endorsement rates to test strings are found in Table 1.

We tested whether participants in the Transfer group learned the partially-cued language better than Control participants. We did a three-way mixed ANOVA, with a between-participant factor of familiarization type (Transfer vs. Control), and the within-participant factors of test string familiarity (heard vs. unheard), and test string cues (cued vs. uncued). The dependant measure was the difference in endorsement rates to grammatical test strings and their ungrammatical counterparts.

There was a main effect of test string familiarity, $F(1, 92) = 88.89$, $p < .001$, with the difference in endorsement rates to heard grammatical strings and ungrammatical ones ($M = .27$, $SE = .026$) more than to unheard ones ($M = .05$, $SE = .022$). There were no other main effects or interactions. Thus, both Transfer and Control participants endorsed heard grammatical test strings more often than unheard grammatical strings, but the Transfer group did not perform better than the control group. Neither group showed differences in endorsement rates to *unheard* grammatical and ungrammatical strings, ts ≤ 1.82, ps $\geq .076$. Thus there was no generalization to unheard items.

These findings suggest that Transfer learners did not benefit from their prior exposure to a fully-cued *aX bY* language. One explanation is that learners cannot acquire an *aX bY* pattern with only partial cues, regardless of their prior experience. However, given that other studies provide evidence of learning partially cued *aX bY* structure (e.g. Frigo & McDonald, 1998), this explanation seems unlikely. An alternative explanation is that learners exposed to a fully-cued language focused only on the surface relationship between the *a* and *b* elements and endings on the *X* and *Y* words, essentially learning a co-occurrence relationship between the first word in the string and the ending on the second. If this were the case, they would learn only the surface regularities resulting from the underlying structure as opposed to the category relationships. Experiencing a perfectly predictive relationship between the first word and the ending of the second may have led Transfer learners to tune-in to this aspect of the partially-cued language as opposed to the abstract structure.

We next tested whether learners would transfer from a partially-cued *aX bY* language, by exposing them to a partially-cued *aX bY* language before transferring them to another version of this partially-cued language. Because the initial language does not have a perfect correspondence between the initial word and ending of the final word of a string, learners would not likely focus solely on it, and thus might perform differently than the Experiment 1 learners at transfer.

Table 1: Endorsement Rates to Test Items with Standard Errors in Parentheses

	Grammatical Heard		Grammatical Unheard		Ungrammatical	
Expt 1	Cued	Uncued	Cued	Uncued	Cued	Uncued
Transfer	.76 (.028)	.77 (.031)	.54 (.029)	.60 (.028)	.50 (.031)	.55 (.032)
Control	.80 (.020)	.75 (.026)	.62 (.025)	.56 (.031)	.56 (.029)	.59 (.026)
Expt 2						
Transfer-2	.83 (.028)	.80 (.026)	.59 (.036)	.54 (.031)	.48 (.033)	.48 (.032)

Experiment 2

Method

Participants Forty-eight University of Arizona undergraduates participated for course credit.

Materials The language materials were the two partially-cued versions of the language used in Experiment 1.

Procedure The procedure for Experiment 2 participants was the same as for the Transfer group in Experiment 1, such that participants were familiarized and tested on one version of the language in Phase 1, and then transferred to the other version in Phase 2. Half of the participants were exposed to Version A of the partially-cued language and then Version B, and half to Version B and then A. This group is referred to as the Transfer-2 group.

Results and discussion We wanted to determine whether participants in the Transfer-2 group learned the partially-cued language they heard in Phase 2 better than the control participants from Experiment 1. Thus, we did a three-way mixed ANOVA, with the between-participant factor of familiarization type (Transfer-2 vs. Control), and the within-participant factors of test string familiarity (heard vs. unheard), and test string cues (cued vs. uncued). The dependant measure was the difference in endorsement rates to grammatical and ungrammatical test strings. Mean endorsement rates to test items can be found in Table 1.

We found that the difference in the Transfer-2 learners' endorsement rates to grammatical vs. ungrammatical test items ($M = .21$, $SE = .030$) was higher than that of the control group from Experiment 1($M = .11$, $SE = .031$), $F\,(1, 93) = 5.69$, $p = .019$. There was an effect of test string familiarity, with the difference in endorsement rates to heard grammatical test strings and ungrammatical ones ($M = .27$, $SE = .026$) greater than the difference between unheard grammatical test items and ungrammatical ones ($M = .05$, $SE = .022$), $F\,(1, 93) = 102.45$, $p < .001$. There was also an effect of test string cues. The difference in endorsement rates to grammatical test strings with cues and ungrammatical strings ($M = .20$, $SE = .026$) was significantly greater than the difference to grammatical test strings without cues and ungrammatical strings ($M = .13$, $SE = .023$), $F\,(1, 93) = 6.50$, $p = .012$. There were no interactions between any of the three variables, suggesting that learning in the Transfer-2 group was generally better than the Control group.

Paired-sample *t* tests comparing endorsement rates for each of the four types of grammatical strings and ungrammatical ones indicate that the Transfer-2 group learned the underlying structure of the language (see Table 1 for means and standard errors). Endorsement rates to grammatical heard test strings, cued and uncued, were higher than those to ungrammatical ones, ts $(47) \geq 6.36$, ps $\leq .001$. Critically, endorsement rates to *unheard* grammatical strings with cues were higher than those to ungrammatical strings, $t\,(47) = 2.10$, $p = .04$. (Recall that control subjects in Experiment 1 did not show such learning). Endorsement rates to grammatical unheard strings without cues did not differ from those to ungrammatical ones, $t\,(47) = 1.68$, $p = .1$.

In sum, learners with prior exposure to a partially-cued language subsequently learn a new version of such a language better than naïve learners. Thus, these findings shed light on how learners might acquire patterns in the absence of robust cues typically necessary for successful learning.

General discussion

In this set of experiments, we demonstrated that what learners can acquire from their input changes as they gain experience with a particular type of structure. Our results suggest that learners exposed to a fully-cued category-induction language become sensitive to a phonological pattern in the form of a perfect correspondence between *a* and *b* words and the endings on the *X* and *Y* words. Because they have not become sensitive to the underlying category relationships, only to the surface correlates of this pattern, their learning of a partially-cued language in which this relationship is probabilistic is not enhanced relative to controls. However, exposure to a partially-cued language facilitates subsequent acquisition of another partially-cued language. What is the basis for generalization in these learners? Recall from Experiment 1 that control participants exposed to a partially-cued language do not generalize to new strings. Tranfer-2 learners, whose initial learning phase was identical to controls', are also unlikely to have learned the *aX bY* structure. While neither the Transfer nor the Transfer-2 group appeared to successfully

acquire the structure of their initial training language, unlike learners transferred from a fully-cued language, Transfer-2 learners were not led to rely on a cue that would later be disrupted. While we cannot precisely determine what aspect of their exposure to the partially-cued language facilitated generalization, it is clear that strong sensitivity to the underlying category relationships does not drive the effect. Generalization may be driven by sensitivity to the underlying pattern learners acquire in their initial exposure. Generalization may also be influenced by other similarities between languages. In these experiments, strings from both versions of the language were composed of two words, the first monosyllabic and the second bisyllabic, and most *X*s and *Y*s in both versions of the language had distinctive features in the form of endings on the words. If some or all of these similarities were absent, transfer of structure might be less likely to occur, thus raising important questions about constraints on transfer. We might test this by transferring learners to languages in which some of these similarities are removed.

These results add to our previous work (Lany et al., 2004) by providing information about how prior learning experience can interfere with generalization (Exp. 1) or enhance it (Exp. 2). We plan to extend this work by asking whether prior exposure to one of the partial-structure languages facilitates acquisition of a language in which the cues are even further diminished. This would be analogous to asking whether the higher incidence of partially cued noun phrases found in English child-directed speech could help learners acquire verb phrases (in which category structure is less reliably marked). This manipulation should also provide information about how prior learning affects the flexibility of later learning.

In conclusion, learners' prior experience can be instrumental in shaping what they acquire from their input. Learning biases of other sorts have also been proposed, such as constraints on the kinds of computations likely to be performed based on the modality of the input (Saffran, 2002), and the tendency to acquire the most reliable or statistically predominant structure from indeterminate input (Gómez, 2002; Gómez et al., 2003). Prior experience is yet another source of bias constraining language learners. Understanding the scope of these biases will contribute importantly to our theories of language acquisition.

Acknowledgements

This research was supported by NIH R01 HD42170-01 to RLG.

References

Braine, M.D.S. (1987). What is learned in acquiring word classes – A step toward an acquisition theory. In B. McWhinney (ed.), *Mechanisms of Language Acquisition.* Hillsdale, NJ: Erlbaum.

Frigo L., & MacDonald, J. (1998). Properties of phonologicasl markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language, 39,* 448-457.

Gerken, L., Wilson, R., & Lewis, W. (in press). Seventeen-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language.*

Gold, E.M. (1967). Language identification in the limit. *Information and Control, 10,* 447-474.

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431-436.

Gómez, R.L., & Lakusta, L. (in press). A first step in form-based category abstraction in 12-month-old infants. *Developmental Science.*

Gómez, R.L., Welch, K., & Lany, J. (2003). Statistical determinants of learning. Paper presented at BUCLD, Boston, MA.

Jusczyk, P.W. (1997). *The Discovery of Spoken Language.* Cambridge, MA: The MIT Press.

Lany, J., Gómez, R.L, & Gerken, L. (2004). Generalization to parallel structure in language acquisition. Manuscript submitted for publication.

Saffran, J.R. (2002). Constraints on statistical learning. *Journal of Memory and Language, 47,* 172-196.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1928.

Saffran, J., & Thiessen, E. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39,* 484-494.

Fuzzy Cognitive Quantification

Anne Laurent (laurent@lirmm.fr)
CNRS - UMR 5506 LIRMM - Université Montpellier II
France-34392 Montpellier Cedex 5

Charles Tijus (Tijus@idf.ext.jussieu.fr)
CNRS - FRE 2627 Université Paris VIII
2, rue de la Liberté France-93526 Saint-Denis Cedex

Bernadette Bouchon-Meunier (Bernadette.Bouchon-Meunier@lip6.fr)
CNRS - UMR 7606 LIP6 - Université Paris VI
8, rue du Capitaine Scott - France-75015 Paris

Abstract

This article presents the results of a two stage investigation about how linguistic quantifiers are used to summarize expressions of quantity. In the first step subjects were asked to give verbal descriptions of arrays of percentages. Then a second group of subjects was asked to reproduce the original array from the verbal description. The second group produced quantities extremely similar to the original percentages even though the verbal descriptions they used did not describe all categories within the arrays. We shall show that quantifiers have implied meaning (e.g., between *a large number* and *most*) and that similar linguistic constructions may refer to amounts that are noticeably different (e.g., *the principal* vs. *principally*). Finally we highlight the importance of the implicit, topic-related meaning in the choice of spoken complements by showing how the concept of fuzzy quantifiers can be applied not only to their modeling, but also to their use in multi-dimensional data searches.

Cognitive semantic approaches of Quantifiers

For Cognitive semantic approaches of Quantifyers have also been quite rare. Just, (1974) defined cognitive traits within three dimensions (Universal-Specific, Large-Small, Negative, Positive) which serve to categorize and give meaning to quantifiers. Quantifiers also project a representation of quantity which influences how information is perceived. The statement " few dots are blue (or red)" with a visual image of two dots of one color and 12 dots of another, will result in attention being focussed on the two dots. If the statement were "a lot of dots are blue (or red)", attention would be focussed on the group of 12 dots (Just & Carpenter, 1971). Quantifiers also carry implied meaning and lead to making inferences about other quantities. This is the case for distinctions based on the positive ("a few of") or negative ("few of") polarity of quantifiers (Paterson, Sanford, Moxey & Dawydiak, 1998). The statement, "there are a few people in the train" designates the amount of people in relation to an empty train, whereas "there are few people in the train" refers to the number of people missing in relation to how many the train should carry.

How much meaning can be inferred from quantifiers? How many bottles of soda should you buy when a friend asks you to pick up "a few"? The analogous approach of Holyoak & Glass, (1978) is based on the idea that there is a direct correspondence between terms and a numerical scale. They showed that confusion arises when the quantifiers are very close in scale, as did Anderson (1981). Furthermore, with their approach it is not possible to evaluate the question when maximum values are unknown (the maximum amount of bottles of soda, for example). The adverbs "generally" and "usually" can be accepted as covering almost all the people or things in consideration. The adverb "often" refers to either the majority of the individuals ("children are often bright") or the majority of occurrences and can thus convey repetition ("demonstrations are often violent"). Continuing in this vein, Hörmann, Cascio & Bass O' Connor, (1974) sought to define how quantifiers would be spread on a scale with class intervals. Unfortunately quantifiers denoted values that varied in relation to what was being quantified. For example "frequently" corresponded to 70% (on average), when referring to how often Miss Sweden was judged attractive, but only to 30% when used to refer to the frequency of airplane crashes (Newstead, 1988). In natural language semantics, where quantifiers denote relations between groups (Geurst, 2003), "a lot" corresponds to the majority. Thus "a lot of A's are B's" can mean that there a more A's that are B's than there are A's that are not B's. This interpretation nevertheless is not valid for the statement "in the last elections a lot of (A: electors) were (B: electors that didn't vote)" which actually refers to the number of non-voters (Barwise & Cooper, 1981). The adverbs "sometimes"' and "rarely" appear to designate respectively a small, but non-negligible frequency, and a quite negligible frequency. Still, it is difficult to say what is "small" or "negligible", as in (1) and (2) where "sometimes" probably is not the same. What a quantifier denotes is therefore dependant on the various elements of the situation being described.

(1) Sometimes I watch the evening news.
(2) Sometimes I go to the movies.

The vast majority of studies have confined themselves to the quantifiers "all", "none" and "some", but even with these most simple cases the process which generates inferences from their meanings remains unclear. In function of given statements 49 to 75% of respondents infer from "all A's are B's" that "all B's are A's" (Newstead, 1988; Chater & Oaksford, 1999). Applying pragmatic linguistics theory,

Grice (1975) similarly found that while "all" should logically encompass "some", it is possible for "some" to be restrictive and exclude "all". Interestingly enough, it has been observed that children do not have the same understanding of quantifiers as adults. Although at 3 years of age children do not differentiate between "all" and "some" (Hollander, Gelman & Star, 2002), by the time they are seven and even though they have not learned all rules of conversation, their answers are logically more valid than adults' are (Smith, 1980).

A question remains as to the information the adult is using to make an inference. Are they saying "some" because they know they cannot include "all" the objects (3), in which case the restrictive interpretation is pragmatically valid, or, are they saying "some" because they have no more information about the other objects (4), in which case a non-restrictive interpretation is more accurate.

(3) There are some broken eggs in this carton

(4) I took some eggs from this carton and they weren't fresh

In contrast to studies which have used ad hoc statements, in this article we have sought to potentially include all the values and expressions that "some - not all" can have in relation to a real situation (results of the 1998 French high school graduation standardized examination, the "baccalauréat"). There were 1,277,282 students in the class of '98. The data from this group are multidimensional; they can be arranged according to many factors: pass rate, gender, age, nationality, region, type of school, presence during the exam, which specialized baccalauréat was taken, which foreign languages were studied, etc.

From the 18 independent dimensions used (each containing 2 to 13 terms), more than 300 million relations are possible when only considering the intersection of the dimensions' terms, because relations may also derive from the fusion of exclusive terms. Our objective was to investigate quantifiers both in terms of their production and their interpretation.

The first step was to observe an initial group of students producing quantifiers (experiment 1). In step two a second group of students was given the opposite task of assigning numerical percentages to the verbal descriptions which were summarized by group 1.

Experiments

In order to assess the wide range of quantifiers used, in experiment 1 we allowed the free production of verbal descriptions for the percentage distributions. In order to assess how often certain quantifiers were used for each distribution, participants then selected terms from a list

Experiment 1: production and choice of quantifiers

Method

Participants. 83 university students in psychology and computer sciences responded to the questionnaires either in writing or via the Internet. In the later case students had to access the research lab's site. All the participants had passed their *baccalauréat* examination one or two years previouslySecond-level headings should be 11 point, initial caps, bold, and flush left. Leave one line space above and 1/4 line space below the heading.

Questionnaires. 18 percentage distributions generated from a database concerning the 1998 baccalauréat, constituted the basic information in the questionnaires. The distributions were selected to include the largest range possible of cases in terms of dimensions, number of variables (from 2 to 13) and values. They were as follows:

D-1: very close values all around a half,
D-2: very different values; one very large, one very small,
D-3: many values; one very large, one small and all others very close and very small,
D-4: many values, but none large (i.e. near a half) and all others very small but not as close as in D-3,
D-5: five not very high values; two quite close, one not very distant and two others quite close and very low,
D-6: three values close to a third and two other very weak ones (almost null),
D-7: data in absolute values (to compare with the first distribution),
D-8: one very strong value, one very weak,
D-9: two strong proportions (above 50%), neither very, nor too distant,
D-10: on two lines, a strong imbalance between columns with a rather large difference between lines with intersections as well,
D-11: two rather close proportions near 50%, but one was 10 points over and the other 10 points below,
D-12 : many very small and close values with two other values close to 20% and a third near 40%,
D-13: two close values, both quite average,
D-14: two high and close values,
D-15: two very weak, close values,
D-16: five values; one near a half, two others near 20%, and two very weak others
D-17: three very weak and close values (two almost equal and the third a little higher)
D-18: two strong, very close values (in comparison to D-14).

*Procedure.*The instructions on page one of booklet one were as follows: "Without using any of the figures in the tables, in a few lines, write what you can say about the table. You have 20 minutes for this task." The second booklet had the same instructions with the constraint to use words from the quantifiers list given: "Without using any of the figures in the tables, in a few lines, use words from the list below to say what you can about the table. If a non-included term seems absolutely necessary you may add it to the list. You have 20 minutes for this task".

Results and Discussion

Quantifier production. The participants' responses fell into two contrasting types of verbal statements summarizing the given distributions. The first type consisted mostly of describing the dominant term(s) and its(their) impact by using quantifiers, like "half". Modifiers like "very" were used extensively. These kinds of responses were particularly common when the distribution table had only two variables and/or when the table had one very high level, very different from the other levels.

The other type of verbal summary was used when there were many variables and when no one term was prominent. The statements first purpose was to organize and structure the data, that is, to make comparisons (especially of near equality) between the different terms, for example by contrasting the girls' results to the boys'. This process can be seen as a relative analysis of the results. Sequencing terms like "first" and "then" were prevalent and no indication in either quantitative or qualitative terms was given for evaluating proportions. When statements of this type set hierarchical relations between variables, they did not on the other hand, establish orders of magnitude. For D-13 and D-14 for example, all the respondents ordered the success rates, but only one gave an indication of the actual values of the scores.

The principal terms found fall into four categories:

Ordering numerical values
- More ___ than ___ and less ___ than ____
- first ... next/then ... the rest ...
- best/better than ...
- many ... few

Quantifiers
- principally, in general, the majority, frequently, often, almost/almost all, a good part
- (very) few of, a minority of
- a third, (nearly) half, more than a million
- about x%

Relations (fuzzy
- about the same, (a little) more often, almost as much, no major difference, lower, raise, constantly, bigger, below, above, …

Modifiers (fuzzy)
- just about, almost, lightly, clearly, noticeably, approximately, (very) (low) minority, large majority, …

It quickly appeared that certain quantifiers were not used as often as would have been assumed. This was the case for the expressions of proportion other than "about half" (i.e. "about a quarter", "about a third") and it was also true for "most of" which wasn't often produced spontaneously, but was chosen when included on the list (see next section). In general the respondents preferred the expression "a lot of". Comparisons and ordering seemed of primary interest and thus, the term "more" was used extensively.

Finally, we observed in most cases that the verbal summary descriptions were incomplete: the respondents only focussed on certain variables in each distribution. It is true that for inter-dimensional variables, orders of magnitude cannot be deduced. However, our distributions were for the most part intra-dimensional. For this reason it was often possible to infer and reconstruct the numerical values of the proportions in the tables (all proportions summed to 100%), from the verbal summaries.

The choice of quantifiers. The quantifiers the respondents chose for summarizing the data are as follows. When more than 75% of the participants chose a certain quantifier, it was then associated with the distribution

If we consider only the data compiled from the Internet questionnaires, only distributions 4 and 10 generated a disparate choice of quantifiers. For the majority of the other cases, three groups of quantifiers stand out; the first being simply "almost all", the second being proportional ("about half", "about a quarter", "about a third", etc.), and the third being composed of quantifiers that describe either the general case ("most of"") or particular cases ("few").

The option "If a non-included term seems absolutely necessary you may add it to the list" was chosen most often to differentiate (and provide a substitute for) descriptions that used the terms "as many/much as", "equals", "similar", "almost equal", "more than", etc., found with test booklet version 1

Items (D-1) and (D-7) presented the same information in numerical values and as a proportion, respectively. For D-7 the participants massively chose the same quantifiers as in the distribution, whereas for the absolute values there was a much greater diversity in the quantifiers selected.

Items (D-14) and (D-18) represent approximately the same numerical data (strong proportions near 75%), but with different dimensions. The temporal dimension is more salient with quantifiers like "constant over time" or "rose by" for D-18, (which described baccaulareat rates by years), than for D-14 (pass rates vs. gender), even though for version 2 the results were quite similar.

Finally, it was noted that although version 2 had different quantifiers added to its list which did not necessarily appear on version 1, the answers concerning the same subject were often very close, for both versions of the questionnaire**.**

Experiment 2: understanding quantifiers

The objective of this second investigation was to determine whether respondents would be able to reconstruct tables of numerical data from the series of verbal summaries containing quantifiers which the subjects in experiment 1 defined.

Method

Participants. As with the first test, the participants were university students in psychology or computer sciences (N=116), who all had passed their baccalauréat one or two years previously. All students used printed questionnaires (none responded via the Internet***).***

Questionnaires. The questionnaire used in experiment 2 contained nine (D-1, D-2, D-3, D-4, D-11, D-13, D-15, D-16 and D-17) of the original 18 distribution tables, without their numeric values. Hereafter they will be designated as D-1 through D-9. Each distribution is associated with a verbal description containing the quantifiers that the majority of the subjects in test one chose. The descriptions are as follows:

D-1: "There are a few more girls than boys",

D-2: "There are very few foreigners enrolled",

D-3: "The test candidates mainly came from public schools, if not, they came principally from state accredited private schools",

D-4: "About half the candidates have no possible area of specialization. Most of the others chose mathematics or a modern foreign language",

D-5: "The majority of test passers were girls",

D-6: "The pass rate was higher (quantifier=more) for the French than for the foreigners and represents the majority",
D-7: "The pass rate was higher (quantifier=more) for the girls than the boys",
D-8: "A quarter of the test takers fail. Half receive no honors. 1/5 receive the honorable mention - good. Very few receive the highest distinction - excellent".
D-9: "The excellent honors rate was very weak for the baccalauréats specialized in economics and social sciences. It is 3 times higher (quantifier=more) for the literary baccalauréats (a little less than double) and it is 4 times higher (quantifier=more) for the scientific baccalauréats".

The tables and their descriptions were on one sheet of paper. On the top of the paper the following instructions appeared: "Use the corresponding summaries to reconstruct tables of numerical data (percentages)".

Procedure. Each participant received the questionnaire sheet with the nine distributions to fill in according to its corresponding verbal summary. They were allotted a maximum of 30 minutes to complete the task.

Some of the tables submitted did not present a total of 100, so they were proportionally corrected to reach 100. In this manner the quantitative values obtained in experiment 2 could be compared to the initial numeric information the subjects used in experiment 1.

Results and Discussion

Item D-7 introduced a comparison with a conditional frequency. The number of boys receiving highest honors was 7.6%; for girls it was 7.8%. This relationship was interpreted by the respondents as being complementary and thus, they furnished values that summed to 100%. Setting aside this distribution, the results demonstrate that the subjects were able to reconstruct numeric tables from the textual summaries. Variation around the average response was quite low in the majority of answers (Figure 1).

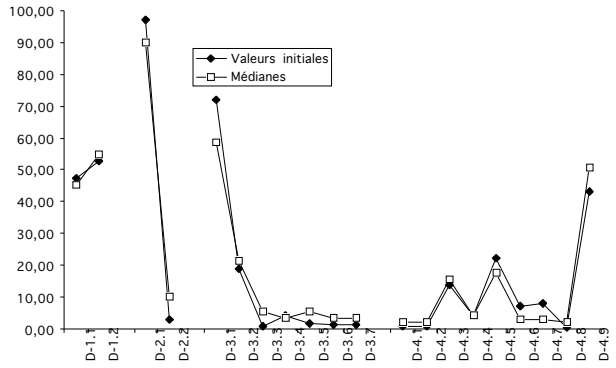


Figure 1: Comparison of the initial values and the reconstructed values.

It is clear that although the verbal descriptions were often incomplete, the participants were capable of filling in the distribution tables with values quite similar to the original ones. Aside from D-7, the averages of the quantities generated by the respondents correlate to the initial values at 0.89 ($p<0.0001$) and the medians correlate at 0.875 ($p<0.0001$). The greatest differences occurred with items D-6 and D-9. In D-6 as in D-7, the participants produced complementary percentages that totaled to 100%. In D-9, they greatly overestimated the percentage of *excellent* honors.

General Discussion

Despite the variances in D-6, D-7 and D-9, the students were able to produce a very good approximation of the initial quantitative data and the question remains as to how they managed this using just the quantifiers in the verbal descriptions.

References

Anderson, J.R. (1981). Memory for logical quantifiers. *Journal of verbal learning and verbal behavior, 20*, 306-321.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy, 4*, 159-219.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 28*, 191-258

Cordier F. & Tijus C. (2001). - Object properties: A Typology. *Current Psychology of Cognition*, 20, 445-472.

Geurst, B. (2003).*Reasoning with quantifiers. Cognition, 86*, 223-251.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds), *Syntax and semantics*, vol. 3: Speech acts. New York : Seminar Press.

Hollander, M.A., Gelman, S.A., & Star, J. (2002). Children's interpretation of generic noun phrases. *Developmental Psychology, 38*, 883-894.

Holyoak, K. J., & Glass, A. L. (1978). Recognition confusions among quantifiers. *Journal of verbal learning and verbal behavior, 17*, 249-264.

Hörmann, Bass, B/M., Cascio, W.F., & O'Connor, E.J. (1974). *Journal of applied psychology, 59*, 313-320.

Just, M. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive psychology, 6*, 216-236.

Just, M., & Carpenter, P. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10*, 244-253.

Newstead, S.E. (1988). Quantifiers as fuzzy concepts. In T. Zetenyi (Ed.). *Fuzzy sets in psychology*. Amsterdam : Elsevier Science Publishers.

Paterson, K.B., Sanford, A.J., Moxey, L.M. & Dawydiak, E.J. (1998). Quantifier Polarity and Referential Focus during reading, *Journal of Memory and Language, 39*, 290-306.

Poitrenaud, S. (1995). The Procope Semantic Network: an alternative to action grammars. *International Journal of Human-Computer Studies*, 42, 31-69.

Smith, C.L. (1980). Quantifiers and question ansérine in young children. *Journal of Experimental Child Psychology, 30*, 191-205.

Understanding Knowledge Models: Modeling Assessment of Concept Importance in Concept Maps

David Leake, Ana Maguitman and Thomas Reichherzer
Computer Science Department, Indiana University
Lindley Hall 215, Bloomington, IN 47405, USA
{leake, anmaguit, treichhe}@cs.indiana.edu

Abstract

Concept mapping is widely used in educational and other settings to aid knowledge construction, sharing, and comparison; concept maps are also used as a vehicle for assessing understanding. To aid the concept mapping process, projects at Indiana University and the Institute for Human and Machine Cognition (IHMC) are developing "intelligent suggesters" to support users as they build concept maps, by presenting them with relevant information from existing knowledge models and the Internet. This depends on identifying important concepts in the concept map under construction. This paper presents and evaluates models of the influence of concept map layout and structure on the selection of concepts expected to be relevant to the topic of concept maps. It presents and assesses a set of potentially-relevant structural factors and evaluates how these factors combine to affect human judgments of concept importance. Twenty subjects were asked to judge the relative importance of concepts in concept maps selected to highlight particular characteristics, and three models were compared to their judgments. Analysis of the results shows that subjects were significantly influenced by concept map topology, but little influenced by other aspects of concept map layout. The results suggest that layout-independent models of concept maps can provide a suitable representation for guiding retrieval of topic-relevant information to support concept map construction, provided that the representation reflects topologically-based influences. The results are applied in the design of the suggesters' similarity assessment procedures for retrieving relevant concept maps.

Introduction

Concept mapping [Novak and Gowin, 1984] has been widely used to elucidate humans' knowledge and to facilitate knowledge elicitation, construction, and comparison and sharing. In concept mapping, users construct a two-dimensional, visually-based representation of concepts and their relationships. The concept map representation encodes propositions describing two or more concepts and their relationships, in simpli ed natural language sentences. In educational settings, concept mapping exercises have been used to encourage students to actively construct an understanding of concepts and relationships within domains of interest. To facilitate concept map construction and sharing, the Institute for Human and Machine Cognition (IHMC) has developed CmapTools, publicly-available tools to support generation and modi cation of concept maps in an electronic form (http://cmap.ihmc.us/). CmapTools enable interconnecting and annotating maps with material such as other concept maps, images, diagrams, and video clips, providing rich, browsable knowledge models available for navigation and collaboration across geographically-distant sites. The CmapTools software has been downloaded by users in approximately 150 countries, and has been used in major educational initiatives, such as the Quorum project [Canas et al., 1995], which involved more than one thousand schools in Latin America. It has also been used for modeling and sharing the knowledge of human experts, for example, for modeling NASA experts' knowledge of Mars (http://cmex-www.arc.nasa.gov/).

CmapTools provides a convenient framework for knowledge construction, but users may have dif culty nding relevant resources, remembering speci c aspects of a domain to include, or locating relevant concept maps to compare. To alleviate this problem, projects are under way at Indiana University and the IHMC to develop intelligent suggesters to support users by retrieving resources such as prior concept maps and multi-media materials [Leake et al., 2003]. Figure 1 shows a screenshot of a Mars knowledge model under construction, with suggestions of propositions, resources, and topics to consider. The suggesters' effectiveness depends on their ability to retrieve topic-relevant information, which in turn depends on modeling users' own judgments as they examine concept maps. Thus modeling users' judgments of the importance of concepts to a map's topic has practical value for suggester software to support concept mapping and scienti c value, for better understanding what in uences human understanding of the knowledge that concept maps convey.

The assessment of concept importance may depend on the concepts they include (based on their labels in the concept map), on the concept map topology, or on layout differences between isomorphic maps. Especially for users unfamiliar with a domain, we would expect topology and layout to play an important role in their assessment of the topic of a concept map. However, to our knowledge, no previous studies have investigated whether/how the topology and layout of a concept map actually in uence judgments of its topic. To hypothesize candidate topological and layout factors that might in uence decisions of which concepts are most topic-relevant, we considered general structure and layout guidelines for building good concept maps in the concept mapping literature, as well as methods for identifying important nodes from the structure of hyperlinked environments. These were used to develop candidate models for the in uence of structural features on identifying the concepts most important to the topic of a concept map. We then performed experiments in which twenty paid subjects judged the relative importance of concepts in concept maps selected to investigate particular structural in uences. We used this data to set pa-

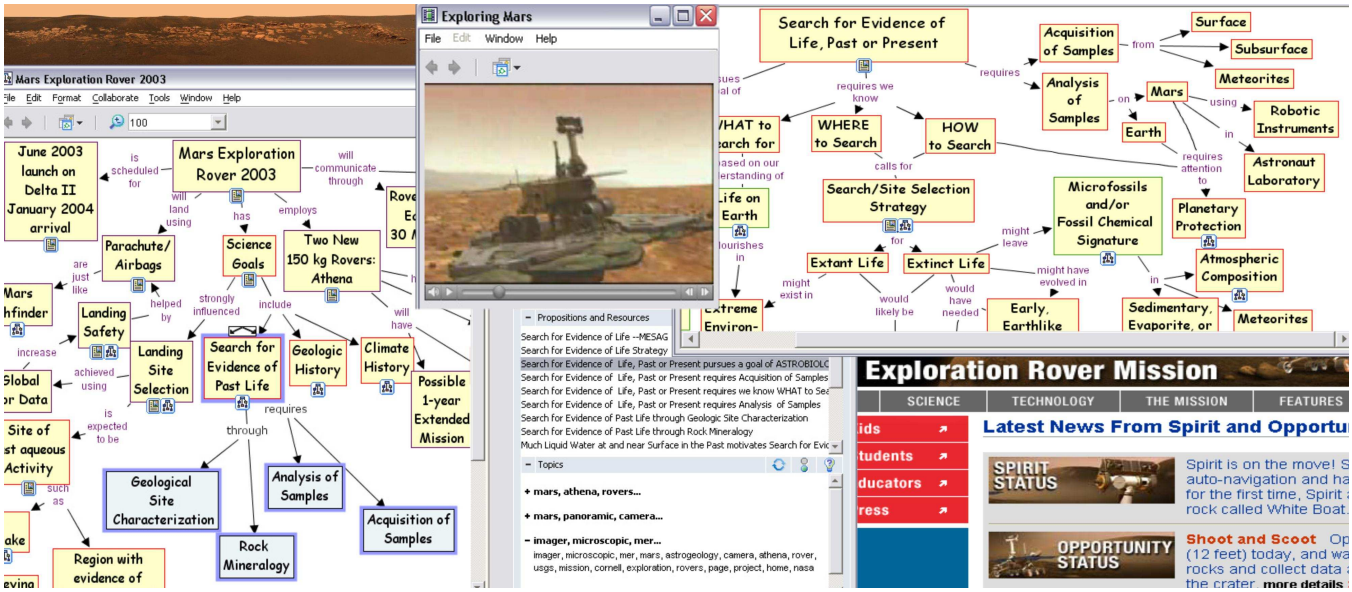


Figure 1: Portion of a Knowledge Model developed by the NASA Center for Mars Exploration, with Sample Suggestions.

rameters in the models and to assess the ability of the models to predict the subjects' performance. Our results suggest that topology is important; the structure of concept maps plays an important role in assessments of concept importance. However, they also suggest that layout plays a less important role. Methods suggested by the models have been implemented in the suggesters to provide support for students and experts' concept map construction.

Modeling Concepts and their Relationships

Concept mapping was developed in an educational setting by Joseph Novak, in an effort to design better teaching and learning activities [Novak and Gowin, 1984]. Novak based the approach on Ausubel's cognitive learning theory[Ausubel, 1963], which proposes that meaningful learning requires deliberate effort by the learner to connect new concepts to relevant preexisting concepts and propositions in the learner's own cognitive structure. Concept mapping was designed to support the learner's effort by externalizing concepts and propositions known to the student, making them visually apparent to facilitate their connection with newly acquired concepts. Concept maps have been used by teachers to assess students' understanding, by students to compare their knowledge and collaboratively re ne their understanding, and by experts as a vehicle for modeling and sharing their knowledge.

Concept maps relate to several other frameworks developed in cognitive psychology and arti cial intelligence to model concepts and their relationships. Schemes based on graphs or networks are commonly used as models of human memory organization, to account for phenomena such as similarity judgments or hierarchical category structure. Early examples include the hierarchical network model [Collins and Quillian, 1969], semantic memory [Tulving, 1972] and conceptual structures [Ausubel, 1963]. More formal approaches to graph-based representations, such as conceptual graphs [Sowa, 1984] or semantic networks [Quillian, 1968], attempt to provide a representation suitable for machine processing. Proposals for non graph-based representations to model concepts and their relationships include formal concept analysis [Ganter and Wille, 1999], which models the organization of concepts in terms of lattice theory, and the geometric structure of conceptual spaces [Gärdenfors, 2000].

Despite the many differences among theories of knowledge organization, they share a fundamental assumption that knowledge can be modeled in terms of a set of components and their relationships. Concept mapping is a method for externalizing such a structure in an individual, making concepts and relationships explicit. Thus examination of concept maps can be used to assess subjects' knowledge [West et al., 2002], and support for the usefulness of this approach has been provided by empirical studies[Aidman and Egan, 1998, Michael, 1994]. However, there has been little study of what affects subjects' judgments of the topic of a concept map, how to determine topic similarity from concepts maps, and the types of representations that may support computer models of concept map retrieval. In previous studies using similar types of representations, topological information about graphs has been used to de ne measures of graph similarity [Goldsmith and Davenport, 1990] and for concept clustering [Esposito, 1990]. These frameworks are based on the premise that the closer the relationship of two concepts the closer they are in cognitive structure the closer they will be in the graph representation. This has been used to induce concept proximity or relatedness. Our study investigates a complementary question, the in uence of other structural factors, such as the numbers of incoming and outgoing links. How graph topology and layout affect assessments of concept importance is central to understanding the information conveyed by concept map structure, as well as for developing models of topic similarity for concept maps.

Models for Analyzing Concept Maps

We developed four candidate models of the in uence of structural and layout characteristics on expectations for the importance of particular concepts to the topic of concept maps. In the models, concepts are represented as nodes in the concept map graph. The baseline model treats map topology and layout as unimportant. The three remaining models use the topology of the concept map to compute a weight predicting each concept's importance in describing the topic of the map.

To determine which factors to include in the models, we rst considered factors from the concept mapping literature. Novak proposed that meaningful learning is facilitated when new concepts or concept meanings are subsumed under broader, more inclusive concepts, which suggests that concept maps should have a hierarchical structure. All of the non-baseline models can re ect such a structure, with weighings re ecting that important concepts are at the top of the map, and less important at the bottom. However, the models are parameterized so that the actual contribution of hierarchical structure if any can be determined empirically. We also considered the applicability of topological analysis methods from other domains, in particular, Kleinberg's algorithm[Kleinberg, 1999] for topological analysis of graphs, used to identify important nodes in a hyperlinked environment. Kleinberg's work characterized nodes on the World Wide Web as hubs and authorities based on their interconnections. When applied to concept maps, we expected hub and authority concepts to be especially important to determining the topic of concept maps.

Connectivity Root-Distance Model (CRD)

The connectivity root-distance model is based on two observations. First, concepts that participate in more than one proposition, as indicated by their connectivity the number of incoming and outgoing connections may be more important in de ning a map's content than concepts with lower connectivity. Second, Novak argues that concept maps are best constructed if a focus question or a single root concept guides the selection of concepts and their hierarchical organization in the map. The root concept, typically located at the top of a map, tends to be the most general and inclusive concept and to specify the map's topic. This suggests that concept importance may increase with proximity to the root concept.

The CRD model determines proximity by counting the number of direct links between the map's root concept and a given concept. For example, in gure 2, the concept masses of ice has a connectivity of four (one outgoing and three incoming links) and a distance of one to the root concept glaciers . If concept k in a map has o outgoing and i incoming connections to other concepts and is d steps distant from the root concept of the map, then the weight assigned to k by the CRD model is

$$W(k) = (\alpha \cdot o(k) + \beta \cdot i(k)) \cdot (1/(d(k)+1))^{1/\delta}$$

The model parameters α, β, and δ determine in uence of the incoming connections, outgoing connections, and distance to the root concept. The formula implies that the higher a concept's connectivity and the shorter its distance to the root concept, the larger its weight and therefore relevance in the topic of the map.

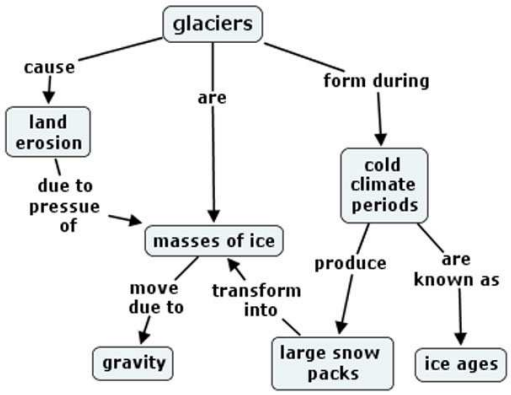


Figure 2: A simple concept map about glaciers.

Hub Authority and Root-Distance Model (HARD)

The Hub Authority and Root-Distance Model also explores the importance of the root node and the hierarchical organization of concepts in maps. However, while CRD performs a local analysis, only taking immediate neighbors into account, HARD performs a global analysis on the in uences of the concepts on each other. Its analysis centers on three different types of concepts that may be found in a concept map:

- *Authorities* are concepts that have multiple incoming connections from hub nodes.
- *Hubs* are concepts that have multiple outgoing connections to authority nodes.
- *Upper* nodes include the root concept and concepts closest to the root concept.

To determine a node's role as a hub or authority, we adapted Kleinberg's algorithm for analyzing hyperlinked graphs to concept maps. Our algorithm, described in detail in [Canas et al., 2001], associates each concept with three weights between 0 and 1, each re ecting the concept's role as a hub, authority, or upper node. A given concept may simultaneously have properties of all three, but in Figure 2, glaciers is primarily a hub concept, due to the number of outgoing connections, and masses of ice is primarily an authority, due to its mostly incoming connections. Among the three concepts with outgoing links to the concept masses of ice , glaciers is the one with the greatest in uence in making masses of ice an authority node, because of the comparative strength of glaciers as a hub.

In the HARD model, the three weights of a selected concept k are combined into a single weight as follows:

$$W(k) = (\alpha \cdot h(k) + \beta \cdot a(k) + \gamma \cdot u(k))$$

In the above formula a, h, and u are the corresponding authority, hub, and upper node weights of a concept in a map and α, β, and γ are the model parameters. As above, the parameters re ect the in uences of the different roles that a concept may play.

Path Counter Model (PC)

The Path Counter Model, like the CRD model, re ects the expectation that concepts participating in more propositions will tend to be more important to the topic of a map. However, instead of considering only a concept node's immediate connectivity, like the CRD model, the PC model considers indirect relationships as well. It counts all possible paths that start from the root and either (1) end on a concept with no outgoing connections, or (2) end on a concept that has already been visited in a path. We note that if a concept has high connectivity (which allows for many paths through the map to include that concept), then the number of paths crossing concepts indirectly linked to the high-connectivity concept increases as well. For example, the PC value for the concept gravity in gure 2 is three, because there are three paths extending from the root concept to gravity, due to masses of ice which is well connected in the map. Formally, to determine the weight $W(k)$ of a concept k in a map, assume that n is the number of paths crossing k. Then the weight is computed as $W(k) = n$. Unlike the previous two models, this model considers only a single in uence on concept weight, and consequently requires no parameters.

Experiments and Results

We conducted a human-subjects experiment to study the in uences of the hypothesized factors on human judgments of concept importance, and the overall t of the four models' predictions to human judgments, with the parameter settings that best t the CRD and HARD models to the subject data.

Method

Twenty paid subjects, all students admitted to Indiana University, were recruited by postings on electronic message boards and bulletin boards for a one-hour experiment conducted on the Web. In a training phase, participants were given a brief description of concept maps and their applications and asked to write a short summary of two concept maps from different domains. In the test phase, subjects answered 56 questions about a total of 12 small concept maps (fewer than 15 concepts each). The maps were designed with controlled differences in their topological structure and layout, to investigate the presence or absence of in uences from particular types of changes (e.g., changing position of a node without affecting topology). Each question presented a concept map and two concepts selected from that map. Participants were asked to examine a map and to answer which of the two concepts best described the map's topic, or whether both described it equally well.

To allow participants to rst practice decision making on regular concept maps, the rst 2 of the 12 concept maps used regular words in the concepts. To prevent domain knowledge from in uencing participants' decisions, concept labels were replaced with arti cial terms in the remaining 10 maps, and only responses concerning the latter 10 test maps were used in evaluating the models. To minimize the in uence of previously-seen concept maps on new responses, different arti cial labels were used for each map, and both the ordering of options for questions and ordering of topological and layout changes between successive concept maps was randomized. The concept maps in the experiment were designed to test speci c hypotheses about the topological and layout factors that may in uence subjects' evaluation of relevance of concepts to a concept map's topic; the absence of domain information forced subjects to rely entirely on topology and layout.

Results

To test whether subjects' judgments of the importance of two concepts changed signi cantly from one map to another, we used a χ^2 test of independence when comparing the subjects' selections from two different maps. Table 1 summarizes the statistical results.

Distance to root concept: To test the in uence of distance to the root concept, subjects evaluated two concept maps in which the distance from a test concept to the root concept was changed from 2 to 1, by inserting an intermediate node. In a series of questions, subjects were asked to compare importances of the test concept, which was moved in the map's hierarchy, to the root concept and neighboring concepts of the moved concept. The results show that the root concept was considered most important compared to the other concepts, and that the importance of the test concept increased as it moved up the hierarchy. The differences in the selection of the moved concept over its neighboring concepts between the two concept maps were statistically signi cant.

Connectivity of a concept: To test the in uence of connectivity, we used two concept maps which differed by increasing a test concept's connectivity the number of incoming and outgoing connections to neighboring concepts from 1 in the rst map to 6 in the second. Subjects were asked to compare importances of the test concept to the root concept and the neighboring concepts of the modi ed concept. When the test concept's connectivity was increased, participants favored it over neighboring concepts and sometimes even over the root concept. All differences were statistically signi cant except for the preference over the root concept.

Layout of a map: To test whether a difference in layout affects subject's selections, two concept maps were constructed with identical topology but substantially different layout. The layout changes primarily involved horizontal organization, but in one instance a single concept was moved from the center right to the bottom left position. The questions asked for both layouts compared the concept that changed its position to its neighboring concepts. The statistical evaluation revealed that the layout changes had no signi cant affect on the concept ratings.

Direct and indirect influences of hub and authority nodes in a map: To test the effects of direct and indirect in uences, a total of four concept maps were constructed with strong hub and authority concepts connected to other concepts in the map. The results showed that hub and authority concepts have an in uence on the selection of concepts, and that authorities play a stronger role than hubs. However, the indirect in uence of either a hub or authority concept on other concepts (when a hub or authority is indirectly connected to a test concept) did not signi cantly affect concept importance.

Fitting the Models to the Data

A hill-climbing algorithm was used to determine the parameter settings for the CRD and the HARD models which gave

In uence	Signi cant	χ^2 Test of Independence
distance to root concept	yes	$(1, N = 40) = 17.04, p < 0.05$
concept connectivity	yes	$(1, N = 40) = 19.37, p < 0.05$
map layout	no	$(1, N = 40) = 0.23, p > 0.05$
direct, hub concept	yes	$(1, N = 40) = 7.74, p < 0.05$
direct, authority concept	yes	$(1, N = 40) = 15.93, p < 0.05$
indirect, hub concept	no	$(1, N = 40) = 3.73, p > 0.05$
indirect, authority concept	no	$(1, N = 40) = 3.73, p > 0.05$

Table 1: Statistical evaluation of in uences on concept importance.

Model	Parameters for Best Fit			RMSE	Cumul. Error
	α	β	γ / δ		
CRD	0.930	4.959	3.603	0.072	27.5%
HARD	0	2.235	1.764	0.1487	32.8%
PC	N/A	N/A	N/A	0.170	27.8%
Baseline	N/A	N/A	N/A	0.564	66.8%

Table 2: Summary of model parameters and RMSE.

the best t between the models and user data. Table 2 summarizes the chosen parameter values, the root-mean-square error (RMSE) of user and model data, and the cumulative error. The cumulative error is the percentage of the total questions (44 questions per subject, involving the 10 test concept maps) for which the models determine different responses from the subjects. To determine a model's preference between two concepts in a concept map, we compared the model's importance values for the two nodes. The model was considered to treat the concepts as equally relevant when their relevance values were within a xed threshold of each other, for a threshold distance determined by hill-climbing. The last row of the table shows the RMSE and the cumulative error for a baseline model. In this model each concept in a map is rated equally important by assigning it a weight of 1.

The results show that the CRD model provides the best t to the user data, followed by HARD and PC. All models except the baseline agree with more than 67% percent of the decisions reached by the participants, who were in a few cases strongly divided in their vote for the best topic-describing concepts. For the remaining 33%, in most cases the models' predictions match the decisions of some subjects. Only once for the CRD model, twice for the HARD model, and four times for the PC model were model and user predictions entirely disjoint. Overall, CRD, HARD, and PC perform better than the baseline model.

Further analysis of the best- t parameters for the CRD and HARD models supports the importance of authority nodes (nodes with incoming connections). For the CRD model, nodes with incoming connections (nodes that play the role of an authority) are more relevant than nodes with outgoing connections (nodes that play the role of a hub) because their β is greater than α. With the best- t parameters for the HARD model, hub nodes are not considered relevant when computing the weight of a node. However, we note that hub nodes still play an important role when computing the level of authority of other nodes in the map.

Discussion

The experiments studied how topology and layout affect assessments of the importance of concepts within concept maps. They compared four candidate models which, using only analysis of a map's topology, compute a weight for each concept in a map. The computed weights provide an estimate of the importance of each concept as a descriptor of the topic of the map, according to subjects' judgments of topic importance.

The studies highlighted the importance of topological information; to our knowledge, this is the rst study to show this effect. They also suggested that speci c layout does not have a signi cant effect. This is important for being able to recognize similarity across concept maps developed by different individuals, despite super cial differences that might affect user judgments. It is also interesting to note that despite the importance of topology, local information alone was suf cient to account for the observed results. The CRD model, which considers distance from the root node and local connectivity, outperformed the HARD model, which takes indirect in uences into account as well.

The current study did not examine interobserver variation; this is an interesting area for future work. Also, the experiment used small concept maps, and considered only the topological and layout factors of the maps, rather than their content. We are conducting additional studies to explore the role of content in assessments of concept importance. However, preliminary results suggest that structure plays a surprisingly strong role, with structural information alone often suf cient to make high-quality predictions.

Application in the Suggesters

The experimental results are re ected in the design of the CmapTools suggesters, two of which are shown in use in the lower center of Figure 1. The rst suggester uses the calculated importance values to weight keywords from concept labels in a concept map, in order to retrieve similar prior concept maps for comparison and to suggest propositions from those maps. This approach to supporting concept map generation is inspired by case-based reasoning [Kolodner, 1993]; concept maps constructed by different users are considered as case-bases of their concept-mapping activity, with each concept map considered to be a separate case. When a user wants to extend a concept to add a new connected concept the system draws upon prior concept maps that include the original concept, as examples of how that concept was extended in similar past contexts. The second suggester uses the similarity weighting to weight keywords for Web search,

to derive topics for the user to consider when starting a new concept map to broaden the knowledge model. These and other implemented suggesters are described in detail in [Leake et al., 2003].

Conclusion

This paper explores factors affecting human judgments of concept importance in determining the topic of concept maps. Modeling such judgments helps elucidate the knowledge captured in concept maps, provides information to guide the design of concept maps in educational settings, and aids the development of intelligent support systems to provide relevant material during concept mapping. Our experiments assessed the in uence of speci c factors and examined the ability of four different models to re ect human assessments of concept importance.

Among the three models, the CRD model, which considers connectivity and distance to the root concept, provided the best match to human data: Its predictions were consistent with the average predictions made by the participants for forty-three out of forty-four questions. The results highlight the importance of local topology and suggest that human topic decisions are robust to layout differences, which is encouraging for the generality of concept mapping for knowledge sharing and the development of support tools to retrieve similar concept maps and topic-relevant information. We are performing followup studies to examine the role of domain content and the t between the predictions of these models and the concept maps developed by domain experts for sample domains.

Principles suggested by the results have been applied to intelligent suggesters to aid the human knowledge modeling process, and the implemented systems appear to give good results in practice. We consider the type of evaluation presented here as important step for guiding the design of such tools, and are now designing experiments to more formally test the relevance of the suggester systems' recommendations during the concept map construction process.

Acknowledgments

This research is supported by NASA under award No NCC 2-1216. We thank Alberto Canas and the IHMC CmapTools development team for their many contributions to the project, Tei Laine for her comments on a draft of this paper, and John Kruschke for assistance on experimental design and analysis.

References

[Aidman and Egan, 1998] Aidman, E. and Egan, G. (1998). Academic assessment through computerized concept mapping: validating a method of implicit map reconstruction. *Int. Journal of Instructional Media*, 25(3):277 294.

[Ausubel, 1963] Ausubel, D. (1963). *The Psychology of Meaningful Verbal Learning*. Grune and Stratton, NY, NY.

[Canas et al., 1995] Canas, A., Ford, K., Brennan, J., Reichherzer, T., and Hayes, P. (1995). Knowledge construction and sharing in quorum. In *World Conf. on Arti cial Intelligence in Education, AIED'95*, pages 218 225. AACE.

[Canas et al., 2001] Canas, A., Leake, D., and Maguitman, A. (2001). Combining concept mapping with CBR: Experience-based support for knowledge modeling. In *Proc. of the Fourteenth Int. Florida Arti cial Intelligence Research Society Conf.*, pages 286 290. AAAI Press.

[Collins and Quillian, 1969] Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8:240 248.

[Esposito, 1990] Esposito, C. (1990). A graph-theoretic approach to concept clustering. In *Path nder associative networks: studies in knowledge organization*, pages 89 99. Ablex Publishing Corp.

[Ganter and Wille, 1999] Ganter, B. and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations.* Springer, Berlin - Heidelberg - New York.

[Gärdenfors, 2000] Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

[Goldsmith and Davenport, 1990] Goldsmith, T. E. and Davenport, D. M. (1990). Assessing structural similarity of graphs. In *Path nder associative networks: studies in knowledge organization*, pages 75 87. Ablex.

[Kleinberg, 1999] Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *J. of the ACM*, 46(5):604 632.

[Kolodner, 1993] Kolodner, J. (1993). *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, CA.

[Leake et al., 2003] Leake, D., Maguitman, A., Reichherzer, T., Canas, A. C., Carvalho, M., Arguedas, M., Brenes, S., and Eskridge, T. (2003). Aiding knowledge capture by searching for extensions of knowledge models. In *Proceedings of the Second Int. Conf. on Knowledge Capture (K-CAP)*, New York. ACM Press, pages 44 53.

[Michael, 1994] Michael, R. S. (1994). *The Validity of Concept Maps for Assessing Cognitive Structure*. PhD thesis, School of Education, Indiana University.

[Novak and Gowin, 1984] Novak, J. and Gowin, D. (1984). *Learning How to Learn*. Cambridge University Press, NY.

[Quillian, 1968] Quillian, M. R. (1968). Semantic memory. In Minsky, M., editor, *Semantic Information Processing*, pages 216 270. MIT Press.

[Sowa, 1984] Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley.

[Tulving, 1972] Tulving, E. (1972). Episodic and semantic memory. In *Organization of Memory*. Academic Press.

[West et al., 2002] West, D., Park, J., Pomeroy, J., and Sandoval, J. (2002). Concept mapping assessment in medical education: a comparison of two scoring systems. *Medical Education*, 36(9):820 826.

Processes of Artistic Creativity: The Case of Isabelle Hayeur

Jude Leclerc (jude.leclerc@umontreal.ca)
Frédéric Gosselin (frederic.gosselin@umontreal.ca)
Department of Psychology, University of Montreal
C.P. 6128, Succ. Centre-Ville, Montreal, QC, H3C 3J7, Canada

Abstract

What are the problems faced by artists in real-life contexts? By what processes do they solve these problems? In this paper, work on scientific discovery (e.g, Klahr, 2000; Kulkarni & Simon, 1988; Langley, Simon, Bradshaw, & Zytkow, 1987) and a situated perspective on creative cognition (e.g., Csikszentmihalyi, 1988, 1999; Nersessian, 2004) are brought together into a unifying framework for studying the processes of artistic creativity in real-life. Within this framework, artistic creativity is viewed as *situated problem solving*. We illustrated our approach by applying it to Isabelle Hayeur, a successful Canadian visual artist.

Introduction

In this paper, work on scientific discovery and a situated perspective on creative cognition are brought together into a framework for studying the processes of artistic creativity. Everybody is interested in art, but up until now few have examined the problem-solving processes that provide support for the artistic process and the production of works of art; almost no one has looked at real-life artistic practices.

We begin by reviewing work on scientific discovery processes. We then present a current definition of creativity that parallels work on situated and distributed cognition. We then go on to describe part of an ongoing field study we are conducting, a study of creative artistic processes in a contemporary visual arts practice, within our framework. Finally, we discuss the potential of this approach for future studies of artistic creativity.

Artistic Practice and Creativity as Situated Problem Solving

Artistic Creativity as Problem Solving

There is now a tradition of studying creativity from a problem-solving viewpoint (e.g., Klahr, 2000; Kulkarni & Simon, 1988; Langley, Simon, Bradshaw, & Zytkow, 1987; Newell, Shaw, & Simon, 1962). The processes of scientific discovery, especially, have been studied from this perspective.

In an excellent review, Klahr and Simon (1999) present the four major approaches of these studies: historical, laboratory, direct observation, and computational. What Klahr and Simon note is that all these approaches to the study of scientific creativity have led to convergent findings about discovery processes.

Klahr and Simon propose that by using the concepts and vocabulary of human problem-solving theory "we may be able … to converge toward a common account of discovery in many areas of human endeavor: practical, scientific and artistic, occurring both in everyday life and in specialized technical and professional domains" (p.524). Here, these concepts and vocabulary are those of problem spaces – states, operators and goals –, heuristic rules, weak and strong search methods – hill-climbing, means-end analysis, planning (Newell & Simon, 1972). Discovery is thus viewed as a search process in a problem-solving space, composed of goals, rules and other aspects of the task and situation.

Up until now, artistic creativity had almost never been studied from a problem-solving perspective. There are a few exceptions (e.g., Weisberg, 1993), but a lot of groundwork still needs to be done. So far, the studies of artistic creativity based on this approach have mainly addressed creative processes in relatively general terms; they have not produced specific descriptions of problem spaces and heuristics in specific artistic practices.

Search Spaces in Scientific Discovery *Search spaces* or *problem spaces* are abstract – representational, conceptual – spaces explored by a 'problem solver' during the problem-solving process. In the case of scientific discovery, scientists have been found to work in two, three, four, and even in search spaces of greater dimensionality (e.g., Klahr & Dunbar, 1988; Kulkarni & Simon, 1988; Schunn & Klahr, 1995; Thagard, 1998; Wolf & Beskin, 1996; see also Klahr & Simon 1999; Klahr, 2000). The traditional two-space view of scientific discovery has its origins in Simon and Lea's (1974) work on problem solving and rule induction; it was first proposed by Klahr and Dunbar (1988). According to this model, in the process of scientific discovery, search happens in two coordinated spaces: (1) the hypothesis space, and (2) the experiment space. Thus, scientific discovery involves generating new hypothesis and experiments; then these experiments serve to evaluate the hypothesis and further generate new ones. This can be considered a problem-solving process.

Similarly, we may ask: what problem space is explored by an artist in the course of the artistic work and practice? In what problem space, and by what processes, is this search conducted? And, of course, there is the possibility that the artist is working through multiple search spaces, corresponding to diverse subproblems involved in artistic creativity.

Artistic Creativity as Situated Activity

According to Csikszentmihalyi (1999), "For creativity to occur, a set of rules and practices must be transmitted from the domain to the individual. The individual must then

produce a novel variation in the content of the domain. The variation then must be selected by the field [the social organization of the domain] for inclusion in the domain" (p. 315; see also Feldman, Csikszentmihalyi, & Gardner, 1994).

From this point of view, creative cognition is not just "in the head" (Norman, 1993a), it is a computational process involving domain and field, as well as the individual. The parallels with situated or distributed approaches to cognition are obvious (e.g., Hutchins, 1995; Nersessian, Kurz-Milcke, Newstetter, & Davies, 2003; Thagard, 1999). Nersessian et al. (2003), for example, studied innovation – creativity – in biomedical engineering research laboratories as a situated and distributed process. The view of creativity as situated, contextual, points toward individual, field, and domain-specific studies of creativity (Csikszentmihalyi, 1988, 1999; Li, 1997; Mace & Ward, 2002).

A lot of recent work in cognitive science explores the situated nature of cognition and action (e.g., Clancey, 1997; Hutchins, 1995; see Nersessian, 2004; Norman, 1993b). Nersessian summarizes the challenges posed to traditional cognitive science by this *environmental perspective* with three interrelated questions: "1) What are the bounds of the cognitive system? 2) What is the nature of the processing employed in cognition? and 3) What kinds of representations – internal and external – are used in cognitive processing?" This perspective effectively poses challenges to cognitive science; the same challenges are also implicit in current models of creativity.

Thus, as with the problem-solving approach, situated and distributed cognition approaches have been used to study processes of scientific discovery (e.g., Nersessian et al., 2003).

Within our framework for studying processes of artistic creativity, in accord with problem-solving theory, recent approaches to situated and distributed cognition, and with current definitions of creativity, we view artistic creativity as *situated problem solving*. We are interested in finding out what problem-solving processes are involved in artistic creativity and in situating these – computations, rules – within the larger system involved in an artistic practice.

Contemporary Visual Arts Practice: The Case of Isabelle Hayeur

To illustrate this approach, we will briefly present preliminary results obtained from the study of a contemporary Canadian visual artist's work and practice. The main focus of this first phase of analysis is on determining the *search spaces* involved in a real-life artistic work and practice.

Isabelle Hayeur[1] is a professional Canadian artist. She is a *professional artist* in the sense of Quebec's law on the *Professional status of artists in the visual arts, arts and crafts and literature, and their contracts with promoters* (R.S.Q., c. S-32.01); she has received multiple grants from both the Canada Council for the Arts and the Conseil des Arts et des Lettres du Québec, and her work has been shown nationally and internationally.

Isabelle Hayeur works mainly with digital photography and video. Her digital photomontages and videos have been shown in solo and group exhibitions, and festivals, in Australia, Belgium, Canada, Chile, Croatia, Denmark, England, Estonia, France, Finland, Germany, Italy, Japan, Malaysia, Mexico, Portugal, Poland, Serbia, Spain, and the United States. She also produces Internet art projects and site-specific works. Her artistic work deals mainly with the impact of the Western model of development on the environment. Her images often display landscapes, part idyllic, part disenchanted, amid man's interventions. Based on a major sociological survey[2] of Québec's visual artists' conditions of practice (Bellavance, Bernier, & Laplante, 2001), she can be considered representative of other successful visual artists in that context.

At the time of writing we had been conducting a field study of this artist's creative processes and practice for a ten-month period; the study is ongoing. Kulkarni and Simon (1988) discussed the use of different kinds of data for building models of processes that span many months or years (e.g., discovery processes in science), where gathering continuous protocols is not practical; in such contexts, recourse to other kinds of data is required. Data about this artist's creative processes were collected on-site, at the artist's studio, through interviews, recording of her artistic activity at the computer, and photographs taken of her work space and tools. Extensive field notes were also taken. The combined data collection allows for the recording of cognitive processes involving a distributed set of activities and tools (see Clancey, 2001). All data was digitally recorded (except for the field notes); the total archived data volume amounts to close to 30 gigabytes.

Our study is at the crossroads of the observational and computational approaches to discovery and creativity processes (Klahr & Simon, 1999); we are using observational and interview data to build a computational description and model of processes of artistic creativity.

Here we will focus on the interview data. Eight semi-structured interviews were conducted over a six-month period, at the artist's studio (Leclerc & Gosselin, 2003). We took inspiration from the traditional protocol analysis methodology (Ericsson & Simon, 1993) for eliciting verbal reports; interviews were thus conducted with the goal of producing information resembling what Ericsson and Simon call "Level 2 verbalizations". This type of verbalization involves descriptive information; Isabelle Hayeur was therefore asked for descriptions of her activities as an artist, not for explanations. Interviews were digitally recorded and were 30 to 60 minutes long each. These were transcribed verbatim and represent a total of 74,507 words. Interviews were organized, stored, and analyzed using Atlas.ti, a computer package designed for qualitative data analysis.

[1] Her work, artist's statement, and resume can be found on her Web site: isabelle-hayeur.com.

[2] This study was commissioned by Québec's main group of professional visual artists, the *Regroupement des Artistes en Arts Visuels du Québec*.

Problem Spaces in Isabelle Hayeur's Creative Process

Viewing artistic creativity as a special case of human problem solving, we have to ask what are the problems solved by an artist? More precisely, what are the problems solved by Isabelle Hayeur? Finding what problems she solves means finding out what problem spaces she explores in the course of her artistic practice. Problem spaces are defined in terms of states, operators, goals, and constraints (Klahr & Simon, 1999, citing Newell & Simon, 1972); we have coded and analyzed our interviews in these terms.

Recently, some researchers have also started to redefine the concept of *problem space*, putting the emphasis not just on internal representations, search, and operations on these representations, but also on the physical space, and the context, involved in real-life problem-solving activity. For example, Nersessian et al. (2003), in their study of innovative practices in biomedical research laboratories, considered the "lab-as-problem-space"; the laboratory, with its resources, people, technology, equipment, etc., is thus considered as a 'problem space'. Similarly, in contemporary visual arts practice, the artist, work space, tools, technologies, technical knowledge and skills, environment, partnerships with other artists, relations with galleries, art centers, funding agencies… constitute the problem space of an arts practice (e.g., see Figure 1). Our analysis is thus based on identifying states, goals, operators, and constraints, in this sense – in a situated artistic practice.

Figure 1: The artist's "studio-as-problem-space"

Criteria for Proposing New Search Spaces Schunn and Klahr (1996; see also Klahr, 2000) have suggested three criteria for proposing new problem spaces: (1) logical, (2) empirical, and (3) implementational. The logical criterion refers to logical coherence of the categories – spaces – proposed; spaces must be mutually exclusive. The empirical refers to the fact that there must be some activity going on in the proposed spaces. And, the implementational criterion allows precise characterization of the proposed problem spaces.

Given the preliminary nature of our research, we have relied mainly on the empirical and on the logical criteria. We have analyzed transcripts from the interviews and coded those in terms of rules (i.e., production rules), condition-action rules. And based on these rules, we have identified goals and heuristic operators[3]; these define the spaces searched by Isabelle Hayeur in the course of her creative process.

Artistic Practice and Career Search Spaces Following coding and analysis of the interviews, two main spaces emerged as the ones most actively searched in the course of Isabelle Hayeur's account of her creative activity: the *artistic practice space* and the *career space*. Throughout the interviews she describes both areas of activity. For example:

> Interview 2
> (28:30) I always plan, I plan moments where I concentrate on my [artistic] production. And there are moments where I put together my artist's dossiers; it is rather dull, but it has to be done. I put together those dossiers [for submission calls]. *You see, there really is the creative work, you know what this is, and there is also everything surrounding that, which takes about half my time* [italics added][4].
>
> Interview 3
> (29:49) I find myself putting more time on my artistic work… the artistic work, and the career.
>
> Interview 5
> (01:08) Already, I am very busy, with things related to the dissemination [of the artistic work], but which I must do, everything surrounding the artistic practice.

In the following sections, we will look at the organization and role of the artistic practice and the career search spaces.

Artistic Practice Search Space: Goals and Heuristics Table 1 shows the main set of goals found to operate in the artistic practice space. These high-level goals shape Isabelle Hayeur's artistic practice. Heuristic operators searching through the artistic practice space apply these goals; these play a role in many heuristic rules used by Isabelle Hayeur to accomplish the tasks associated with this space.

Table 1: Artistic practice space main goals

[gR2-12; gR2-13] Doing my work as an artist seriously, full time.
[gR2-14; gR2-15] Living with less money, in order to put more time into my artistic practice (and less time into 'bread-and-butter' jobs).
[gR2-17; gR2-28] Having more time for my artistic practice.

[3] Goals are labeled gR and rules R. A rule is given the number of the interview in which it first appeared; goals are constitutive parts of rules.

[4] For this paper, interview excerpts, goals, and rules were translated from the French language.

[gR2-30; gR2-32] Working on my images, especially after a few weeks of not working on them.
[gR2-31] After a long time working on my images, taking some time away from the work, doing something else.
[gR2-33; gR-35] Putting time into my practice, taking up and continuing work on projects, planning time when I concentrate on my production.
[gR2-38] Art must remain a calling, it must remain research; the career side must not take too much time.
[gR2-38; gR3-06] I want my images, series, artistic work and career to succeed.
[gR2-39] Doing art for the knowledge it brings in my own life, and for what it may teach or give to others.
[gR3-04] Not stopping my artistic work.
[gR3-07] Being an artist; doing this my entire life.
[gR5-01; gR5-02] Creating strong works, strong images; saying things in a strong way, a stronger way.

Among the goals defining the artistic practice space, some appear to play a major role because they call upon many other goals to search the problem space. For example, gR3-07 calls on a host of activities to reach its aim (see Tables 1 & 2, [gR2-33; gR2-35; gR2-36; gR2-37; gR2-38; gR3-01, gR3-02; gR3-05; gR3-06], for an example of subgoals – heuristics – called by gR3-07).

The artistic practice space is divided further in a number of important subspaces. Among these figures the *image-generation space*. Of great importance, it is the very basis of Isabelle Hayeur's artistic practice; this subspace includes all the knowledge and skills actually involved in the image production activity. Another subspace would be a 'project-management' subspace. We will not expand on these here.

The task achieved through the artistic practice problem space is the task of being an artist, of focusing on one's artistic practice and of producing art works.

Career Search Space: Goals and Heuristics This is the main set of goals found to operate in the career space. Goals gR2-36, gR3-01, gR3-06, gR3-07, and their associated heuristic rules, are the most significant; these actually call upon every other goal and heuristic in the career space.

Table 2: Career space main goals

[gR2-01; gR2-03] Sitting on panels, juries, etc., with other artists.
[gR2-01; gR2-02] Learning. Learning how other artists talk about their work, getting ideas about how to present your work, how art councils work, etc.
[gR2-26; gR2-27] Being represented by a private art gallery, in order to sell my work.
[gR2-34; gR6-01] Sending my work to art centers, galleries, and obtaining exhibitions.
[gR2-35; gR3-05] Putting together "artist's dossiers" – artistic projects and related documents about my practice (to be sent to art centers when there are calls for submissions).
[gR2-36] Doing the things that make a difference in an artist's career, in order to have a successful career.
[gR2-37; gR3-02; gR3-05] Taking care of the career side – everything that surrounds the creative work: searching for submission-call deadlines, new places and centers to show my work, residencies, putting together my "artist's dossiers" according to grants and submission-call deadlines, also answering specific requests for exhibitions and, at some point, sending a set of dossiers – around ten at a time – to art centers I want to reach in a given year, etc. I put half of my total work time as an artist on these activities (when I do not have other contracts, 'bread-and-butter' jobs to do).
[gR2-38] Not putting too much time on the career side; leaving aside some activities if necessary, even if I miss out on some opportunities.
[gR3-01; gR3-06] Being entrepreneurial; sending out a lot of artist's dossiers in order to have exhibitions.
[gR3-01; gR3-07] Having success as an artist; being an artist my entire life.

The career space is further subdivided in two main subspaces: the dissemination and the promotion spaces. These serve to solve the 'problem' of making the artistic work known and seen.

Some of the career goals are related to the same rules as some of the artistic practice goals. This is because certain heuristics operators mediate activity between these two search spaces. The task achieved through the career problem space is the task of making one's work known and seen, thereby building up a successful career.

Heuristics Coordinating Search We have found some heuristic operators, some rules, to be of special importance in Isabelle Hayeur's creative process because they coordinate the search between the artistic practice and career spaces. Here is such a heuristic operator associated with the recurrent goal gR3-07:

> [R3-07] If I want to be an artist and I want to be an artist my entire life, and I know what I have to do, then I do it immediately (i.e., gR2-33, gR2-35, gR2-36, gR2-37, gR2-38, gR3-01, gR3-02, gR3-05, gR3-06).

R3-07 coordinates a lot of activity related to the artistic practice and the career spaces. In fact, it links vocational goals, wanting to live the life of an artist, with very practical career goals and activities.

Here is another example of a heuristic operator linking artistic practice and career:

> [R2-38] If I want art to remain a calling as it must, and if at a certain point I realize that the 'career' side takes too much of my time, then I just don't do it, that activity (even if it means missing opportunities).

Identifying coordination between search spaces is an important part of modeling problem-solving processes; as Klahr (2000) notes, "One must … distinguish search in a particular space from coordination among multiple spaces" (p. 215).

An Additional Search Space This additional space, the *economic space* is not directly part of Isabelle Hayeur's process of artistic creativity, although, as we will see, it is essential for it. This space could also be called the 'working for a living' or the 'bread-and-butter job' space. Table 3 shows a sample of goals from this space.

Table 3: Some economic space goals

[gR2-07] Taking small jobs, contracts, especially in my domain or related to my practice, the arts, and the art milieu.
[gR2-07] Trying to find more gratifying, better paid, and a little bit more interesting jobs.
[gR2-13] 'Bread-and-butter' jobs must not take away from my hours of artistic work.
[gR2-14; gR2-16] To live with less money, in order to need to work less (in order to have more time for my practice).
[gR2-23; gR2-24] Not putting time into searching for bread-and-butter jobs; taking what comes.
[gR2-27; gR2-28] To sell my art work, in order to spend more time on my production and less on contracts outside my practice.

Some rules related to these goals show an interaction between the artistic practice, career, and economic spaces. Here are some examples:

> [R2-07] If you are an artist, and you (necessarily) need to pay for your own production (e.g., the high cost of printing large format photographs), and you have the chance to work in your own domain, then generally you accept these small jobs.
> [R2-14] If I cannot live solely from my art, and I have to take 'bread-and-butter' jobs, and I do not want this to replace my hours of artistic practice, then I decide to live with less money, in order to need to work less.
> [R2-28] If I sell my art work, even if just one image a month, then for each picture sold, I have one less contract to do, and I have more time for my practice.

These rules show that artistic creativity – artistic practice and career – is supported by the economic space. Search, minimal search in Isabelle Hayeur's case, in this space aims at finding the necessary resources to allow most of the artist's activity to be focused on her professional life and artistic production. The main task achieved through the economic problem space is finding (minimal) financial resources to support artistic and career related activities.

In Isabelle Hayeur's life and developmental trajectory as an artist, less and less time is spent on the economic space and more is spent on the actual artistic practice and career (see goals gR2-12, gR2-13, gR2-14, gR2-15, gR2-16, gR2-17, gR2-23, gR2-24, gR2-27, gR2-28, gR2-33, gR2-35, gR2-38, gR3-04, in Tables 1, 2, & 3). The rules that coordinate the artistic practice, career, and economic spaces aim at: (1) diminishing economic space activity, (2) maintaining career activity at a balanced level, and (3) maintaining or augmenting artistic practice activity level. According to Bellavance, Bernier, and Laplante's (2001) survey, few professional artists manage to achieve these goals in Québec's and Canada's socio-cultural and economic context. One measure of an artist's success, at least in regard to the interplay between practice, career, and economic spaces, seems to be his or her ability to do just that, focus on the artistic life rather than just on sheer survival.

Our situated problem-solving perspective on artistic creativity has shown that two main spaces are directly involved in a contemporary visual artist's creative process: the artistic-practice-as-problem-space and the career-as-problem-space. When Nersessian (2004) describes the challenges posed by the environmental perspective to the traditional view of cognition, she mentions considerations of the boundaries of cognitive systems; according to this perspective, cognition is situated and distributed in a complex cognitive system, a system that includes environment and individual. In this first part of Isabelle Hayeur's case study, we found a number of environmental elements playing a role in the artistic practice space (e.g., the artist's studio, equipment, time and financial resources, knowledge and skills needed to produce art works, etc.) and the career space (e.g., relationships with other artists, art centers and galleries, funding agencies, etc.), and defining the complex cognitive space of her artistic creativity.

Conclusions

The project of modeling Isabelle Hayeur's processes of artistic creativity is ongoing. What was outlined here is meant as an illustration of our framework for studying real-life artistic creativity; our preliminary results suggest a well-integrated set of search spaces and processes involved in real-life, situated, artistic practice and cognition. Further work will involve collecting verbal protocols related to the image-generation search space – the actual picture producing process; we have already recorded more than 100 hours of her image-generation activity.

Within the *artistic creativity as situated problem solving* framework, it is possible to study real-life artistic practices. The product is a descriptive model of search spaces, goals, and heuristic operators involved in artistic creativity. Dasgupta (1994, see also 2003) has done something of the kind in the context of science and technological innovation. The type of studies provided by our framework may lead to computational models of historical instances of artistic creativity, as studies in science have led to computational models of historical scientific discoveries (Langley, Magnani, Cheng, Gordon, Kocabas, & Sleeman, 2001). Such studies may also serve an educational purpose by providing information about real-life processes of artistic creativity.

Acknowledgments

We thank Isabelle Hayeur for her generous participation in this research project and for graciously giving her consent to the disclosure of confidential information. This research was supported by an Excellency Scholarship from the University of Montreal awarded to Jude Leclerc, and by an NSERC

(R0010085) and an NATEQ (R0010287) grant awarded to Frédéric Gosselin.

References

Bellavance, G., Bernier, L., & Laplante, B. (2001). *Les conditions de pratique des artistes en arts visuels: Rapport d'enquête, phase 1*. Montréal, Canada: INRS Urbanisation, Culture et Société.

Clancey, W. J. (1997). *Situated cognition.* New York: Cambridge University Press.

Clancey, W. J. (2001). Field science ethnography: Methods for systematic observation on an expedition. *Field Methods, 13,* 223-243.

Csikszentmihalyi, M. (1988). Society, culture, and person: A systems view of creativity. In R. J. Sternberg (Ed.), *The nature of creativity* (pp. 325-339). New York: Cambridge University Press.

Csikszentmihalyi, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 313-335). New York: Cambridge University Press.

Dasgupta, S. (1994). *Creativity in invention and design: Computational and cognitive explorations of technological originality.* New York: Cambridge University Press.

Dasgupta, S. (2003). Multidisciplinary creativity: The case of Herbert A. Simon. *Cognitive Science,* 683-707.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Feldman, D. H., Csikszentmihalyi, M., & Gardner, H. (1994). *Changing the world: A framework for the study of creativity.* Westport, CT: Praeger Publishers.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Klahr, D. (2000). Exploring science: The cognition and development of discovery processes. Cambridge, MA: MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual space search in scientific reasoning. *Cognitive Science, 12,* 1-55.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125,* 524-543.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12,* 139-175.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes.* Cambridge, MA: MIT Press.

Langley, P., Magnani, L., Cheng, P. C.-H., Gordon, A., Kocabas, S., & Sleeman, D. H. (2001). Computational models of historical discoveries. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (p. 3). Mahwah, NJ: Erlbaum.

Leclerc, J., & Gosselin, F. (2003). [Interviews with Isabelle Hayeur]. Unpublished raw data.

Li, J. (1997). Creativity in horizontal and vertical domains. *Creativity Research Journal, 10,* 107-132.

Mace, M.-A., & Ward, T. (2002). Modeling the creative process. *Creativity Research Journal, 14,* 179-192.

Nersessian, N. J. (2004). Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In M. Gorman, R. Tweney, D. Gooding, & A. Kincannon (Eds.), *New directions in scientific and technical thinking* (pp. 17-56). Mahwah, NJ: Erlbaum.

Nersessian, N. J., Kurz-Milcke, E., Newstetter, W. C., & Davies, J. (2003). Research laboratories as evolving distributed cognitive systems. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In H. E. Gruber, G. Terrel, & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking* (pp. 63-119). New York: Atherton Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Norman, D. A. (1993a). Cognition in the head and in the world: An introduction to the special issue on situated action. *Cognitive Science, 17,* 1-6.

Norman, D. A. (Ed.) (1993b). Special issue on situated action [Special issue]. *Cognitive Science, 17*(1).

Schunn, C. D., & Klahr, D. (1995). A 4-space model of scientific discovery. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 106-111). Mahwah, NJ: Erlbaum.

Schunn, C. D., & Klahr, D. (1996). Integrated yet different: Logical, empirical, and implementational arguments for a 4-space model of inductive problem solving. In G. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 25-26). Mahwah, NJ: Erlbaum.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. Gregg (Ed.), *Knowledge and cognition* (pp. 105-128). Hillsdale, NJ: Erlbaum.

Thagard, P. (1998). Ulcers and bacteria I: Discovery and acceptance. Studies in History and Philosophy of Science. Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 29, 107-136.

Thagard, P. (1999). *How scientists explain disease.* Princeton, NJ: Princeton University Press.

Weisberg, R. W. (1993). *Creativity: Beyond the myth of genius.* New York: Freeman.

Wolf, D. F., & Beskin, J. R. (1996). Task domains in N-space models: Giving explanation its due. In G. Cottrell (Ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 27-28). Mahwah, NJ: Erlbaum.

An Efficient Method for the Minimum Description Length Evaluation of Deterministic Cognitive Models

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

The ability to evaluate competing models against noisy data is central to progress in cognitive science. In general, this requires advanced model selection criteria, such as the Minimum Description Length (MDL) criterion, that balance goodness-of-fit with model complexity. One limiting property of many of these criteria, however, is that they cannot readily be applied to deterministic models. A solution to this problem, developed by Grünwald (1999), involves a process called 'entropification' that associates deterministic models with probability distributions, and allows MDL criteria to be calculated. However, a potential practical difficulty with this approach is that it requires a multidimensional summation over the data space that can be prohibitively computationally expensive in realistic situations. This paper derives a simpler version of the MDL criterion for deterministic models in the important special case of 0-1 loss functions that is computationally feasible. Two concrete applications of the simpler MDL criterion are presented, demonstrating its ability to consider model fit and complexity in selecting between competing models of cognitive processes. The first application involves three different heuristics for a problem solving task, while the second involves three different models of forced-choice decision making.

Introduction

To a large extent, progress in cognitive science relies on the development of better models of cognitive phenomena. Models provide a formalized representation of theoretical explanations, and make predictions that can be tested empirically. For this reason, the ability to evaluate competing cognitive models against noisy data in a complete and meaningful way has been a central concern recently in mathematical psychology (e.g., Myung & Pitt 1997; Myung, Balasubramanian & Pitt 2000; Myung, Forster, & Browne 2000; Pitt, Myung, & Zhang 2002).

In particular, there has been a strong (and overdue) focus on balancing the goodness-of-fit of models with their complexity. These ideas have been applied to core topics in cognitive science such as models of psychophysical discrimination (e.g., Myung *et al.* 2000), stimulus representation (e.g., Lee 2001; Navarro & Lee 2003; in press), inference and generalization (e.g., Tenenbaum & Griffiths 2001), and decision-making (e.g., Myung & Pitt 1997).

Probabilistic Models

For the most part, however, these recent development have been restricted to considering *probabilistic* cognitive models. This class of models has the property that any parameterization (or, more generally, any probability distribution over the parameter space) corresponds to a probability distribution over the data. That is, the model corresponds to a parametric family of probability distributions over the data. This means that considering a probabilistic model at a particular set of parameter values makes some data quantifiably more likely than others. In turn, for probabilistic models the likelihood of any observed data having arisen under the model at any parameterization of interest can be evaluated.

Many cognitive models are probabilistic in this way. For example, models of memory retention (e.g., Rubin & Wenzel 1996) usually consist of parameterized functions that specify the probability an item will be recalled correctly after a period of time. As another example, the ALCOVE model of category learning (Kruschke 1992) also produces a probability, that depends upon the values of a number of parameters, for each possible category response on any trial. For these models, their probabilistic nature allows likelihood to be measured against any pattern of observed data.

Many advanced model selection criteria, such as Bayes Factors (e.g., Kass & Raftery 1995), Minimum Description Length (MDL: e.g., Grünwald 2000), Stochastic or Geometric Complexity (Myung, Balasubramanian & Pitt 2000; Rissanen 1996), and Normalized Maximum Likelihood (Rissanen 2001), rely on this property. This is because they integrate the probabilities of the data across the parameter space of the models, or the maximum likelihoods across all possible data sets, and so require non-zero probabilities over a subset of the parameter space that has measure greater than zero to be meaningful.

Deterministic Models

As Myung, Pitt and Kim (in press) note, however, there are many important cognitive models that belong to the alternative class of *deterministic* models. These models specify differently how to assess the relationship between data on the one hand, and model predictions at different parameterizations on the other. For example, a sum-squared loss or error function might be proposed, so that increasingly large differences between model predictions and observed data are penalized more heavily in evaluating the model. Alternatively, a 0-1 loss function might be proposed, so that models are evaluated as being correct only if they predict data exactly, and are wrong otherwise. What deterministic models do not specify, however, is an error theory that describes the likelihood of data that differ from model predictions. This means that, when a deterministic model makes incorrect predictions, it is not possible to assign the probabilities needed by many modern model selection criteria.

A good example of a deterministic cognitive model is the 'Take the Best' model of decision making (Gigerenzer & Goldstein 1996). This model takes the form of a simple algorithm, searching a fixed stimulus environment in a deterministic way, so that it will always make the same decisions. One way of interpreting the model in relation to empirical data is that it has probability one when it makes the same decision as that observed, but probability zero when it makes a different decision. Adopting this approach, however, any evaluation of the model against human data involving multiple decisions is very likely to find an overall probability of zero, because at least one of the model's decisions will disagree with the data.

Other deterministic models that face similar problems include the memory models surveyed by Pietsch and Vickers (1997), axiomatic theories of judgment and choice (e.g., Luce 2000), and various lexicographic decision models (e.g., Payne, Bettman & Johnson 1990). For these sorts of models, the natural assessment is in terms of the proportion of correct decisions it makes, or some such error function, but this measure is not the same as the probabilities from likelihood functions used in probabilistic model selection. In particular, it is not clear how the error function measuring goodness-of-fit should be combined with measures of model complexity to undertake model selection.

Recently, however, Grünwald (1999; see also Myung, Pitt, & Kim in press), has developed a model selection methodology that overcomes these difficulties. He provides a principled technique for associating deterministic models with probability distributions, through a process called 'entropification', that allows MDL criteria for competing models to be calculated. There is a potential practical difficulty, however, in using this approach to evaluate cognitive models. The MDL criterion involves multidimensional summations over the data space that could be prohibitively computationally expensive in some realistic situations. This paper derives and demonstrates a reformulation of the MDL criterion for deterministic models in the important special case of 0-1 loss functions that is much less computationally expensive.

The MDL Criterion

In this section, Grünwald's (1999) formulation of the MDL criterion based on entropification is described, and a computationally simpler form is then presented. In one sense, the reformulation is just a straightforward algebraic manipulation, and has probably been noted (but not published, as far as we are aware) by others. In another sense, making the reformulation explicit, and demonstrating its advantages, is a useful contribution. There are many cognitive models that are deterministic and naturally assessed under 0-1 loss[1], for which the MDL method described here ought to find wide application.

Original Formulation

Suppose a deterministic model M is being evaluated using a dataset D that has n observations, $D = [d_1, \ldots, d_n]$. Each of the observed data are discrete, and can assume only k different values. The model uses P parameters $\theta = (\theta_1, \ldots, \theta_P)$ to make predictions $Y = [y_1, \ldots, y_n]$. To evaluate any prediction made by the model, a 0-1 loss function is defined as $f(D, Y) = \sum_{i=1}^{n} \gamma_i$, where $\gamma_i = 0$ if $d_i = y_i$ and $\gamma_i = 1$ otherwise. By considering all possible parameterizations, the model makes a total of N different predictions. In other words, there are N different predictions, $Y_1, \ldots, Y_N$, the model is able to make about the data by choosing different parameter values. In general, the relationship between parameterizations and predictions will be many-to-one. This means that every unique model prediction is naturally associated with one or more parameterizations of the model.

Under these assumptions, Grünwald (1999) shows that using entropification the model making prediction Y can be associated with a probability distribution, parameterized by the scalar w, as follows:

$$p(D \mid M, Y, w) = \frac{e^{-wf(D,Y)}}{\sum_{x_1=1}^{k} \cdots \sum_{x_n=1}^{k} e^{-wf(D,[x_1,\ldots,x_n])}}.$$

Determining the MDL criterion for the model requires finding the model predictions Y^* and scalar w^* that *jointly* maximize $p(D \mid M, \theta, w)$ to give the value p^*.

[1] All of the deterministic decision making, memory and judgment models already mentioned effectively have 0-1 loss when they are restricted to two choices. There are other models, such as the optimal stopping models considered later, that are also naturally associated with 0-1 loss despite having a larger number of choices.

Once this is achieved the MDL criterion for the model is given simply by $\text{MDL} = -\ln p^* + \ln N$.

Besides automatically balancing the competing demands of model fit and complexity, this MDL criterion has at least two attractive properties for model selection in cognitive science. First, differences in MDL values, through their natural probabilistic interpretation, can be assessed as odds, in much the same way as Bayes Factors. This allows the assessment the 'significance' of different MDL values for different models to be done meaningfully as a question of the standards of scientific evidence required for the problem at hand, using a scale that is calibrated by betting. Secondly, as Grünwald (1999, pp. 24-28) discusses, the information theoretic or coding approach used by MDL means that results are available for cases where the data generating process that is being modeled has statistical properties that are not perfectly represented by the models being considered. We would argue this is inevitably the case for cognitive models, and so the ability of the MDL approach to address this problem is an important one.

Despite these attractions, however, there is an obvious difficulty in maximizing $p(D \mid M, \theta, w)$. The problem is that the denominator given by $Z = \sum_{x_1=0}^{k} \cdots \sum_{x_n=0}^{k} e^{-wf(D,[x_1,\ldots,x_n])}$ involves considering every possible data set that could be observed, which involves a total of k^n terms. In cognitive science, where it is possible for a deterministic model to be evaluated using many data points, each of which can assume many values, the repeated calculation of Z may be too computationally demanding to be practical.

A Simpler MDL Computation

A simpler form for Z can be derived by noting that $f(D, Y)$ can only take the values $0, \ldots, n$, in accordance with how many of the model predictions agree with the data. Since Z considers all possible data sets, the number of times $n-x$ matches (i.e., x mismatches) will occur is $\binom{n}{x}(k-1)^x$. For a prediction Y that has $n-m$ matches with the data (i.e., there are m mismatches and $f(D, Y) = m$), this leads to the simplification

$$p(D \mid M, Y, w) = \frac{e^{-wm}}{\sum_{x=0}^{n} \binom{n}{x} (k-1)^x e^{-wx}},$$

which has a denominator that sums $n+1$ rather than k^n terms.

The computational efficiency offered by this reformulation means it will generally be possible to find the w_i^* that maximizes $p(D \mid M, Y_i, w_i)$, giving p_i^*, for all N model predictions. The p^* required for MDL calculation is then just the maximum of $p_1^*, \ldots, p_N^*$.

Finding each w_i^* can also be done efficiently by observing that

$$\partial p / \partial w = \frac{1}{Z^2} e^{-wm} \sum_{x=0}^{n} \binom{n}{x} (k-1)^x (x-m) e^{-wx}.$$

This derivative is clearly always positive if $m = 0$ and always negative if $m = n$. This means, if a model predicts all of the data correctly, $w_i^* \to \infty$, and if a model fails to predict any of the data correctly $w_i^* \to -\infty$. Otherwise, if $0 < m < n$, the substitution $u = e^{-w}$ allows w_i^* to be found from the positive real roots of the degree n polynomial

$$\sum_{x=0}^{n} \binom{n}{x} (k-1)^x (x-m) u^x.$$

by standard numerical methods (e.g., Forsythe, Malcolm, & Moler 1976).

Grünwald (1999, pp. 98-99) notes, with particular reference to the 0-1 loss function, that the case $w < 0$ corresponds to 'inverting' models. For example, if a model only makes two choices, and so considers binary data (i.e., $k = 2$), the inverted model changes all of the model predictions to the alternative possibility. We would argue it will generally be the case in cognitive modeling that it is not appropriate to consider inversion, because this manipulation will require the model to be interpreted in a substantively different and unintended way. If this is the case, it is necessary to restrict consideration to $w \geq 0$ in finding the MDL value.

With this restriction in place, the Y^* and w^* learned from data for qualitative model selection convey useful information in their own right. In particular, as Grünwald (1999, pp. 94-95) explains carefully, the value of w^* measures the 'randomness' of the data with respect to the model Y^*, so that smaller values of w^* indicate that the the model provides relatively less information about the data.

Demonstrations of the MDL Criterion

In the remainder of this paper, we present two concrete examples of the MDL criterion evaluating cognitive models, in situations where there is a clear need to assess whether the better goodness-of-fit of some models warrants their additional complexity. The first involves different heuristics for a problem solving task, while the second involves different models of forced-choice decision making.

Optimal Stopping Problem

As a first demonstration of the MDL criterion for deterministic models, consider three different account of human decision-making on an optimal stopping task sometimes known as the full-information secretary problem (see Ferguson 1989 for a historical overview).

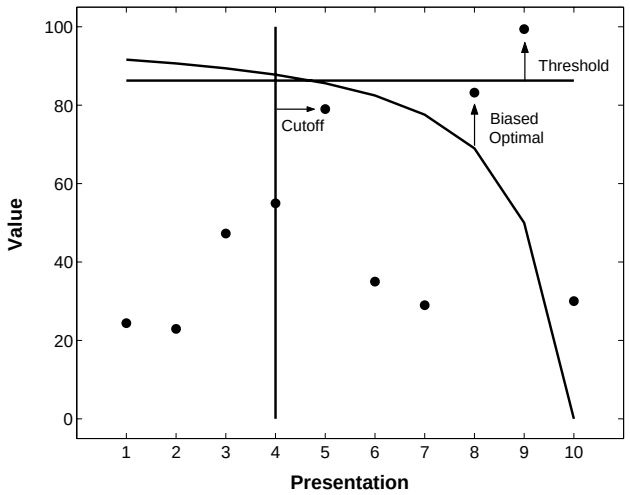


Figure 1: An optimal stopping problem of length 10, with the sequence of values shown by circles, demonstrating the operation of the biased optimal (curved line), threshold (horizontal line) and cutoff (vertical line) models.

Background In these problems, a person presented with a sequence of numerical values, and told to select the maximum. They must decide whether to accept or reject each possibility in turn and, if a possibility is rejected, they cannot select it at some later point. The number of choices in the complete sequence is fixed and known, and the distribution from which the values are drawn (usually a uniform distribution on the interval $[0, 1]$) is also known. Performance is assessed using a 0-1 loss function, so that if choosing the maximum is regarded as correct, but any other choice is regarded as incorrect.

From the mathematical (e.g., Gilbert & Mosteller 1966) and psychological (e.g., Seale & Rappoport 1997) literature, there are at least three plausible accounts of how people might make decisions on these problems. The first 'threshold' model assumes people simply chooses the first value that exceeds a fixed threshold. The second 'biased optimal' model assumes people choose the first value that exceeds a threshold level, where the threshold level changes for each position in the sequence. The threshold levels correspond to the mathematically optimal values (see Gilbert & Mosteller 1966, Tables 7 and 8), for the given problem length, all potentially biased by shifting by the same constant. The third 'cutoff' model assumes people view a fixed proportion of the sequence, remember the maximum value up until this cutoff point, and then choose the first value that exceeds the maximum in the remainder of the sequence. Each of these models has one parameter, giving the threshold, the bias, or the cutoff proportion respectively. For all three models, if no value meets the decision criterion, the last value presented becomes the forced choice.

Figure 1 summarizes the three models on a secretary problem of length 10. The sequence of values presented is shown by the filled circles. The horizontal line shows the constant level used by the threshold model. The threshold levels for the optimal model with no bias follow the solid curve. The vertical line shows the proportion used by the cutoff model. Under these parameterizations, the biased optimal, threshold, and cutoff models choose, respectively, the eighth, ninth, and fifth values presented.

Application of MDL Lee, O'Connor and Welsh (this volume) administered $n = 20$ problems of length $k = 10$ to a number of subjects. For this set of problems, the threshold, biased optimal, and cutoff models are able to predict, respectively, 60, 78, and 9 data sets by varying their parameters. As a concrete example of how the MDL criterion can balance these different model complexities against the fit they are able to achieve, consider the decisions made by one subject from the experiment. For this subject, the best-fitting parameterizations of the threshold, biased optimal, and cutoff models correctly predict, respectively 14, 17, and 10 of the 20 decisions. This is an interesting case to consider, because increases in model complexity lead to increases in model fit.

The MDL criteria values for each model, in relation to this subject's data, are 29.5, 19.4 and 38.0 respectively, showing that, despite its increased complexity, the biased optimal model provides a better account than the threshold and cutoff models. This superiority can be quantified in terms of naturally interpretable odds, because differences between MDL values lie on the log-odds scale. For example, the biased optimal model provides an account that is about $e^{29.5-19.4} \approx 24,000$ times more likely than that provided by the threshold model.

Sequential Sampling Processes

As a second example, we consider the sequential sampling model of decision making developed by Lee and Cummins (in press).

Background Lee and Cummins (in press) proposed that an evidence accumulation approach can unify the 'Take the Best' (TTB: Gigerenzer & Goldstein 1996) model with the 'rational' (RAT) alternative to which it is usually contrasted. The cognitive process being modeled involves choosing between two stimuli on the basis of the cues or features that each does or does not have. In essence, TTB searches the cues until it finds one that only one stimulus has, and then simply chooses that stimulus. The RAT model, in contrast, forms weighted sums across the cues for both stimuli,

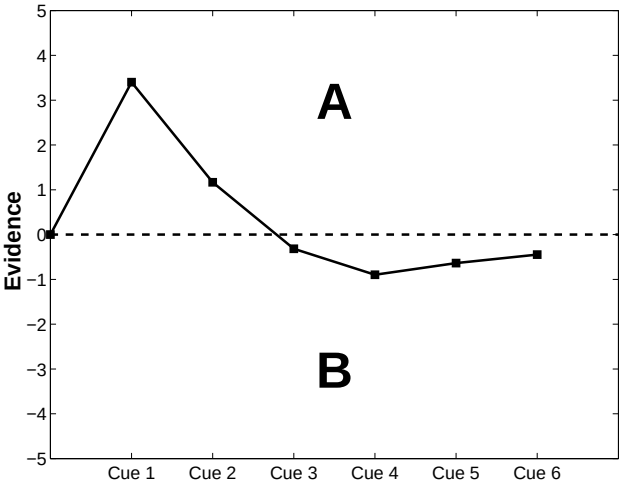


Figure 2: A sequential sampling process using evidence accumulation to decide between choices A and B. Successive evidence values are shown as cues are examined from highest validity to lowest. A decision is made once the evidence exceeds a threshold value.

and chooses the one with the maximum sum.

Figure 2 shows a sequential sampling process accruing information in making this sort of decision. Each of the cues is examined and the evidence provided by that cue is used to update the state of the random walk in favor of choosing stimulus A or stimulus B. If stimulus A has the cue and stimulus B does not, the random walk moves towards choosing A. If stimulus B has the cue and stimulus A does not, the random walk moves towards choosing B. If both stimuli either have or do not have the cue, the state of the random walk is unchanged.

The important observation about Figure 2 is that the TTB and RAT models correspond simply to different required levels of evidence being accrued before a decision is made. If a very small evidence threshold were set, the sequential sampling process would choose stimulus A, in agreement with the TTB choice. Alternatively, if a very large evidence threshold were set, the sequential sampling process would eventually choose stimulus B (because the final evidence is in its favor), in agreement with the RAT model. In general, if a threshold is small enough that the first discriminating cue is guaranteed to have evidence that exceeds the threshold, sequential sampling corresponds to the TTB decision model. If a threshold is large enough that it is guaranteed never to be reached, the final evidence is used to make a forced decision, and sequential sampling corresponds to the RAT decision model.

Application of MDL For the 200 decisions collected from 40 subjects by Lee and Cummins (in press), the TTB model made 36% correctly, while the RAT model made 64% correctly. The sequential sampling model, at the best-fitting value of its evidence threshold parameter, made 84.5% of the decisions correctly. Of course, the sequential sampling model, through its use of the parameter, is more complicated than both the TTB and RAT decision models, which are parameter-free. This raises the issue of whether the extra complexity is warranted by the improved accuracy. Using the model selection method developed here, Lee and Cummins (in press) found MDL values of 87.6, 138.6 and 130.7 for the sequential sampling, TTB and RAT models respectively. The much smaller MDL value for the unified model indicates that it provides a better account of the data, even allowing for its additional complexity.

Conclusion

These demonstration of the MDL criterion provides clear practical examples of how it can be used to evaluate competing deterministic models of human cognitive processes. It also highlights the contribution of this paper, which is a simpler form of the MDL criterion for the special case of 0-1 loss functions. For the optimal stopping problem example, the original MDL formulation involves summing 10^{20} terms in the denominator to find $p(D \mid M, Y, w)$ for each combination of m and Y that needs to be evaluated in optimization. The simpler form given here requires summing only $n + 1 = 21$ terms each time. For the sequential sampling problem, the original formulation involves $2^{200} \approx 10^{60}$, while the simplification involves 201 terms. As these comparisons make clear, the drastic reduction in computation offered by the simplification developed here makes the MDL evaluation of deterministic cognitive models under 0-1 loss feasible for most (if not) all empirical data collected in cognitive science.

Acknowledgments

I thank Jay Myung, Peter Grünwald and three anonymous reviewers for particularly helpful comments.

References

Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science 4* (3), 282–296.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1976). *Computer Methods for Mathematical Computations*. New York: Prentice-Hall.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal Way: Models of bounded rationality. *Psychological Review, 103* (4), 650–669.

Gilbert, J. P., & Mosteller, F. (1966). Recognizing the maximum of a sequence. *American Statistical Association Journal 61*, 35–73.

Grünwald, P. D. (1999). Viewing all models as 'probabilistic'. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT' 99)*, Santa Cruz. ACM Press.

Grünwald, P. D. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology 44*(1), 133–152.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99* (1), 22–44.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45* (1), 149–166.

Lee, M. D., & Cummins, T. D. R. (in press). Evidence accumulation in decision making: Unifying the 'take the best' and 'rational' models. *Psychonomic Bulletin & Review.*

Lee, M. D., O'Connor, T. A., & Welsh, M. B. (this volume). Human decision-making on the full-information secretary problem. *Proceedings of the 26th Annual Conference of the Cognitive Science Society.*

Luce, R. D. (2000). *Utility of Gains and Losses: Measurement Theoretical and Experimental Approaches.* Mahwah, NJ: Erlbaum.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences 97*, 11170–11175.

Myung, I. J., Forster, M., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology 44*, 1–2.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review 4*(1), 79–95.

Myung, I. J., Pitt, M. A., & Kim, W. J. (in press). Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *Handbook of Cognition.* Thousand Oaks, CA: Sage.

Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun., & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 59–66. Cambridge, MA: MIT Press.

Navarro, D. J., & Lee, M. D. (in press). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review.*

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1990). *The Adaptive Decision Maker.* New York: Cambridge University Press.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*(5), 1712–1717.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103* (4), 734–760.

Seale, D. A., & Rapoport, A. (1997). Sequential decision making with relative ranks: An experimental investigation of the "Secretary Problem". *Organizational Behavior and Human Decision Processes 69*(3), 221–236.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, Similarity, and Bayesian Inference. *Behavioral and Brain Sciences, 24* (4), 629–640.

Pietsch, A., & Vickers, D. (1997). Memory capacity and intelligence: Novel techniques for evaluating rival models of a fundamental information processing mechanism. *Journal of General Psychology, 124*, 229–339.

Creative strategies in problem solving

N. Y. Louis Lee (ngarlee@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544-1010 USA

P. N. Johnson-Laird (phil@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544-1010 USA

Abstract

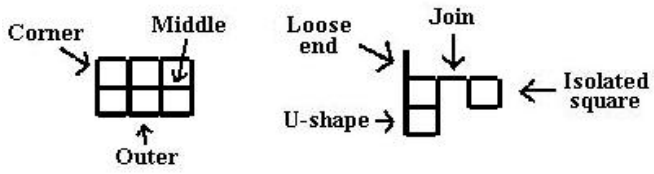
Three experiments investigated how individuals solve "shape" problems. These problems are not susceptible to a means-ends strategy. They consist of a configuration of squares, whose sides consist of separate pieces; and the task is to remove a given number of pieces to leave behind a given number of squares. The paper presents a theory of how individuals develop strategies for these problems. Experiment 1 explored the constraints of symmetry and visual saliency in shape problems. Experiment 2 corroborated the theory's prediction of a major shift in which knowledge acquired during the evaluation of tactical steps comes to govern the generation of these steps. Experiment 3 showed that participants could be biased to adopt strategies making use of specific tactical steps.

Introduction

How do individuals develop strategies to solve problems? The question arises crucially for those problems that come in a series of different instances (e.g. Luchins's, 1942, water jug problems). Our aim was to answer this question for problems that do not have a unique solution and for which individuals cannot develop a simple deterministic strategy guaranteeing an error-free solution. We therefore studied what we refer to as "shape" problems (see Katona, 1940). Figure 1 presents an example of such a problem. There is an initial shape made out of separate pieces (matchsticks) and the goal is to remove a given number of pieces to leave a given number of squares. There are two constraints: the resulting squares should be of the same size as the initial squares, and the solution should not have any loose ends (pieces with an end not connected to any other piece).

Figure 1: On the left is a shape problem in which the task is to remove five matches so that only ten squares remain. A solution is shown on the right.

An important feature of shape problems is that naïve individuals cannot use a means-ends strategy in which they work backwards from the desired goal (Newell & Simon, 1972). The goal merely specifies how many squares should remain, but not how they are arranged. Likewise, individuals cannot always tell if a tactical step in a shape problem makes progress towards the goal. The discovery of the tactical steps in shape problems is accordingly a discovery of the problem space. There are, in fact, seven distinct tactical steps for removing pieces, which are summarized in Figure 2.



1. To remove 1 piece & 0 squares, remove *loose end*
2. To remove 1 piece & 0 squares, remove *join*
3. To remove 1 piece & 1 square, remove *outer*
4. To remove 1 piece & 2 squares, remove *middle*
5. To remove 2 pieces & 1 square, remove *corner*
6. To remove 3 pieces & 1 square, remove *U-shape*
7. To remove 4 pieces & 1 square, remove *isolated-square*

Figure 2: The seven tactical steps for shape problems.

In what follows, we outline a theory of how individuals explore these tactical steps, and how they use these explorations to develop strategies. We then report three experiments that test the predictions of this theory.

The theory

Problem solving is a creative process, and we distinguish three main sorts of algorithm for creativity (e.g., Johnson-Laird, 1993). First, a *neo-Darwinian* algorithm consists of a stage in which ideas are generated followed by a stage in which they are evaluated. Generation depends on arbitrary combinations and modifications of existing elements; evaluation depends on the use of knowledge as constraints to filter out useless results. Any ideas that survive can be recycled recursively through the generative stage again, and so on. Second, in a *neo-Lamarckian* algorithm, all the knowledge acquired from experience constrains the generation of ideas. If alternatives are created, then choice amongst them can only be arbitrary, because all the

constraints have already been used in their creation. When individuals have the requisite knowledge, the algorithm is highly efficient, because there is no need for recursion. Third, in a *multi-stage* algorithm, some knowledge is used to constrain the creation of ideas and some knowledge is used to evaluate the results – with the option of recursion. In sum, according to this account, constraints govern the evaluation of ideas, or their generation, or both.

The algorithm that individuals use to solve shape problems should depend on their experience. Naïve individuals are likely to tackle their initial problems using a strategy that is close to neo-Darwinian. They should be constrained solely by the statement of the problem, the problem shape itself, and their existing perceptual and cognitive processes. As they try out the various possible tactical steps, they learn their consequences, which are summarized in Figure 2. Learning occurs whether or not a tactical step turns out to be useful in solving a problem. Any problem allows only a limited set of tactical options, and so individuals should gradually narrow down the steps that are left to explore. Likewise, granted that the problem is within their competence, they should at length hit upon a sequence of steps that leads to a solution. In addition, some pieces in the problem shape are visually salient, and they may bias participants to attempt certain tactical steps first. Saliency is likely to depend on the perimeter of the shape. Any piece in the perimeter should be *visually salient* if it has at least one adjacent piece that is also in the perimeter and is at right angles to it. In addition, a piece should be more salient if both adjacent pieces are at right angles. A visually salient *component* comprises a group of such visually salient pieces that are adjacent to each other. These principles are probably special cases of broader factors governing visual salience. The acquisition of tactical knowledge depends on perceptual abilities, e.g., subitizing a small number of squares, and conceptual and inferential abilities, e.g., a grasp of the concepts of *squares* and *pieces*, and relevant arithmetical operations.

According to the theory, as tactical knowledge is acquired, it shifts from the evaluative stage of the creative process to the generative stage. Individuals accordingly shift from using a neo-Darwinian algorithm to a multi-stage algorithm, and may even converge on a neo-Lamarckian algorithm. This strategic shift enables them to avoid useless tactical steps and thereby to make more efficient progress towards solutions. They may proceed at once to correct tactical steps. A central component of an efficient strategy for shape problems is the *ratio* of the number of pieces to remove to the number of squares to remove (henceforth, the "p/s" ratio). It constrains the appropriate tactical steps from those afforded by the current configuration of the problem (see Figure 2). But, its optimal use depends on knowledge of the full variety of tactical steps. Conversely, a limited knowledge of these steps yields limited strategies for coping with the problems. Yet, the shift of tactical knowledge to the generative stage of problem solving should still occur, albeit with a restricted repertoire of tactical steps. Hence, it should be possible to bias the development of strategies by giving individuals only a limited experience of tactical steps in an initial set of problems.

Experiment 1

This experiment explored two factors that should affect shape problem solving: whether the initial shape is symmetrical or asymmetrical, and whether the solution is salient or not in the shape.

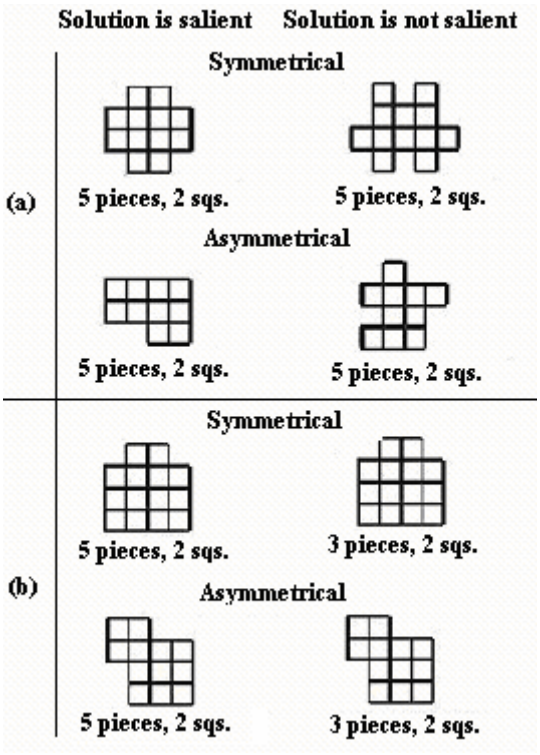


Figure 3: The eight problems used in Experiment 1. We manipulated symmetry and the presence of a salient solution.

Method and procedure

Twenty Princeton University students carried out eight problems, which manipulated symmetry and the presence or absence of salient solutions. Figure 3 presents the eight problems used in the experiment. In order to counterbalance the manipulation, one set of four problems (see Figure 3a) called for the same number of pieces and squares to be removed. This condition is feasible only by changing the shapes from the cases in which the solution is salient to the cases in which it is not. Hence, a second set of four problems (see Figure 3b) used the same shapes in these two cases but changed the number of pieces and squares that had to be removed. The experiment employed a block design: half of the participants carried out the four problems in Figure 3a first, and the other half carried out the four problems in Figure 3b first. The assignment of block

presentation, as well as the order of problems within each block, was random.

On each trial, the participants constructed the shape in a given diagram using matchsticks. They then tried to solve the problem. They were told that they should not leave any loose ends, and that each square must consist of four pieces. They had to say "done" at the end of each trial to the experimenter, who recorded the latencies.

Results and discussion

Figure 4 presents the mean latencies to solve the eight problems. The two blocks did not differ reliably ($z = 1.61$, n.s.), and therefore we collapsed their latencies for analysis. Participants solved problems with a salient solution reliably faster than those without a salient solution (Wilcoxon signed-rank test, $z = 3.85$, $p<.001$). In addition, they also solved problems with a symmetric initial shape reliably faster than those with an asymmetrical initial shape (Wilcoxon signed-rank test, $z = 2.09$, $p<.05$). The two variables did not interact. These results demonstrate that existing factors in the problems can constrain problem solving strategies. To investigate how people develop strategies to cope with shape problems, however, participants would need to solve a series of problems calling for the removal of different numbers of pieces, and to think aloud as they solve the problems. Experiment 2 employed this procedure.

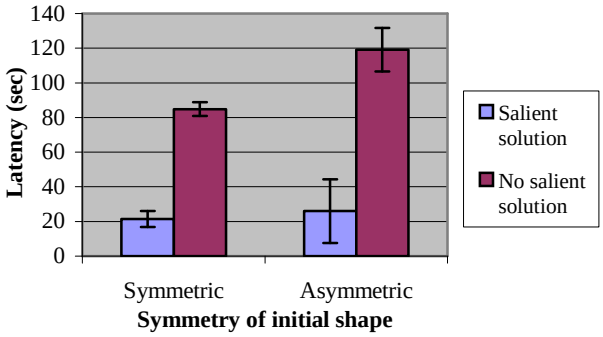


Figure 4: Experiment 1: Mean latencies as a function of symmetry and the presence of salient solution.

Experiment 2

This experiment tested the key prediction of a shift in strategy from a neo-Darwinian exploration of steps to their use in constraining the generation of steps. As a corollary, there should be a reduction in the number of steps that individuals take to solve problems.

Method and procedure

Fourteen Princeton University students carried out 12 problems presented in a random order. The problems (see Figure 5) varied in terms of symmetry, number of matches to be removed, and tactical steps, but they all called for removing two squares. The experimental procedure was the same as that in Experiment 1. However, there was an additional requirement: participants had to think aloud as they solved the problems. We video-recorded what they did and what they said.

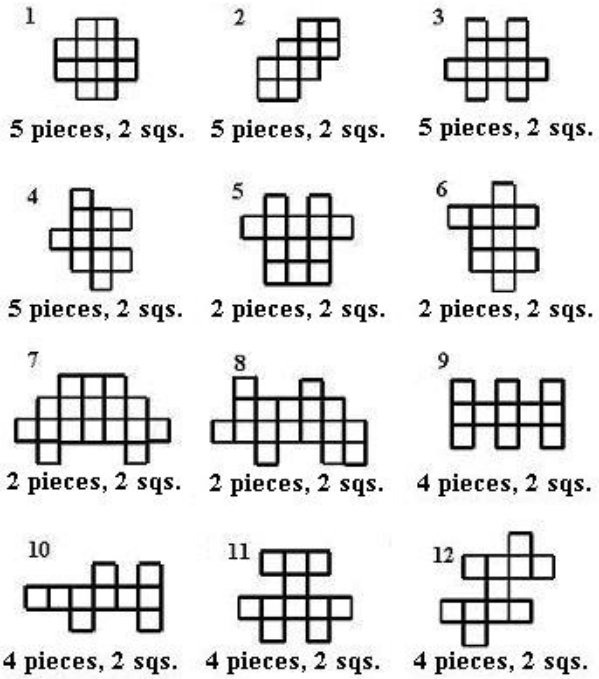


Figure 5: The 12 problem shapes used in Experiment 2.

Results and discussion

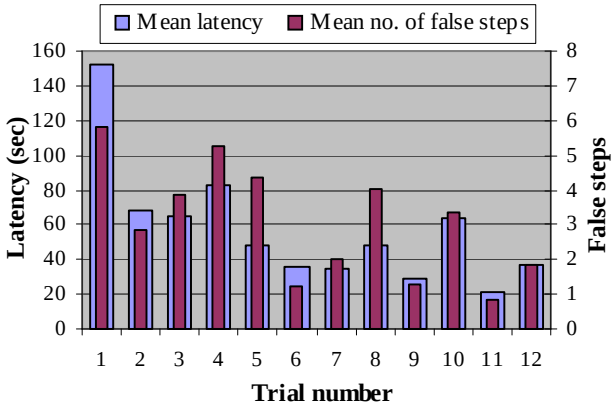


Figure 6: Experiment 2: Mean latencies and numbers of false steps across trials.

Figure 6 presents the mean latencies to solve the problems, and the mean numbers of false steps, over the 12 trials. A false step was one that the participants subsequently undid. As predicted, the participants were able, with experience, to solve the problems faster (Page's $L = 7882.5$, $z = 4.86$, $p <<$

.001), and to make fewer false steps (Page's L = 6750.0, z = 2.16, p <.05). These two variables correlated reliably for all but two of the problems (Pearson's r ranged from .65 to .95, with p <.05 to p <.001). In addition, in a post-experimental questionnaire, the participants were most likely to identify those tactical steps that they had used during the experiment: they all mentioned the *outer* and the *U-shape*, but none identified all seven tactical steps (see Figure 2).

The transcriptions of the video-recordings showed that the participants relied mainly on a single strategy, but there were other two strikingly different strategies. The main strategy has two stages. The first stage is exploratory: the participants try out various tactical steps, which they usually subsequently undo. They are acquiring knowledge of these steps, including steps irrelevant to the present problem. They are also acquiring knowledge of the *p/s ratio*, i.e., the ratio of pieces and squares to be removed (see the previous section). They grasped its relevance, but rarely in a complete way. The duration of this stage depends on the participants' experience with the problems. It accordingly shrinks in proportion over the problems as the participants acquire knowledge. The second stage of the strategy is the application of tactical knowledge. The participants consider the *p/s ratio*, often mentioning it explicitly, and use their tactical knowledge to select an appropriate tactical step. The shift has occurred from a neo-Darwinian strategy to a multi-stage strategy. Hence, the participants are able to make rapid progress to the solution. For some problems, they make no false steps. Likewise, they can combine tactical steps into a single step that solves the problem at a stroke. In other cases, they apply their knowledge recursively, removing a correct piece, re-assessing the number of pieces and the number of squares to be removed, and, as result, selecting a further tactical step, and so on, until they solve the problem. They are converging on a neo-Lamarckian strategy for solving shape problems.

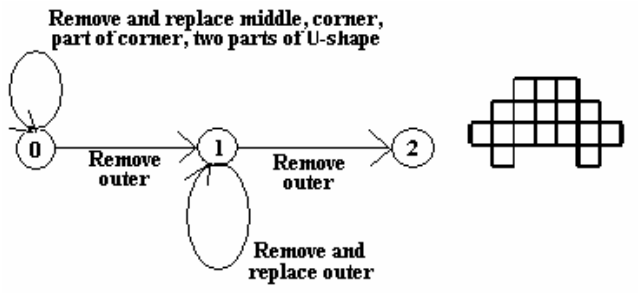


Figure 7: An example of the main strategy in a non-deterministic finite-state automaton. On the right, is the shape problem in which the goal is to remove 2 pieces and 2 squares (Problem 7).

Figure 7 shows an example of the main strategy. The participant started in an exploratory state (state 0) and tried various steps, which she immediately undid. She then correctly removed an *outer* (shifting to state 1), and then removed another outer to solve the problem. On some subsequent trials, e.g., with problem 5, she proceeded at once to the correct solution with no false steps. The protocol is typical in that it appears to reflect the use, not of a simple deterministic strategy, but one in which various steps are tried out in a way that appears to be non-deterministic.

One participant used a quite different sort of strategy. During the first stage of tackling a problem, the participant removed some pieces – often the required number – in an apparently arbitrary way, sometimes leaving several loose ends. The participant then carried out one of three actions: removing a new piece, replacing a piece removed earlier, or moving a piece from one position in the shape to fill the position of a piece that had been removed. The participant persisted in these steps until the solution emerged. The strategy was inefficient, yielding many more false steps than other participants. Yet, the participant gradually acquired some tactical knowledge, which became evident in both a more judicious initial removal of pieces and in more efficient steps in the second stage.

Another participant used a strategy that depended on the initial shape. The participant used the statement of the problem to divide the initial shape into two or three conceptual parts. For example, for Problem 2 (remove five pieces to remove two squares), the participant identified the number of squares to remain in the solution (eight), and then partitioned this number into two parts (three squares plus five squares). The participant then searched for ways to eliminate all but these configurations. Unfortunately, the attempt ignored the number of pieces to be removed. The strategy was inefficient, and yielded little tactical knowledge.

All three strategies stabilized as instances of the multi-stage algorithm outlined earlier. No-one developed a neo-Lamarckian strategy that guaranteed that they could proceed directly to the solution of a problem without any false steps.

Experiment 3

When individuals acquire a deterministic strategy, it transfers to new problems (see, e.g., Luchins, 1942). With shape problems, however, individuals do not acquire a deterministic strategy guaranteed to lead to solution, but instead acquire a tactical knowledge that constrains the generation of steps. Their resulting strategy does not appear to be deterministic (see Figure 7). Nevertheless, it should be possible to bias the development of strategies by giving participants an experience of only certain tactical steps in the initial problems. Experiment 3 tested this prediction.

The participants first encountered a series of four problems that could be solved only by using certain tactical steps. These tactics differed between two groups of participants. Both groups then tackled two "ambiguous" problems that could be solved using either set of tactics. A final unambiguous problem could be solved only with novel tactics, i.e., a problem used to train the participants in the other group. Such a problem should force the participants

back to a greater use of the exploratory stage of their strategy.

Method

Twenty Princeton undergraduates were assigned at random to one of two groups: both carried out seven problems calling for the removal of four pieces to eliminate two squares. In Group 1, participants tackled four problems that could be solved only by removing two *corners*; in Group 2, they tackled four problems that could be solved only by removing a *U-shape* and an o*uter*. Each participant carried out these trials in a different random order. Both groups then attempted two ambiguous problems, and finally a problem chosen randomly from the first four problems given to the other group. Figure 8 shows the complete set of problems. The experimental procedure was the same as that in Experiment 1.

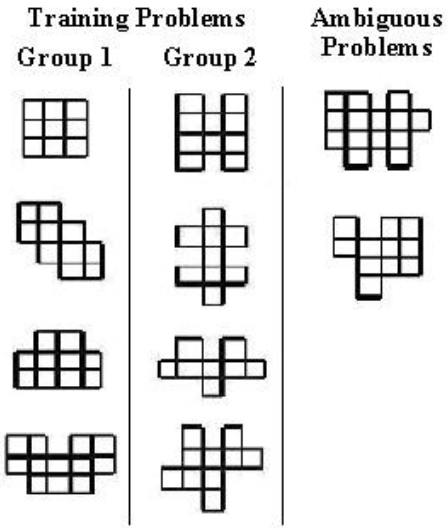


Figure 8: The problems used in Experiment 3. Each problem called for the removal of four pieces and two squares.

Results

Figure 9 presents the mean latencies of the two groups to solve the problems. The participants took progressively less time to solve the problems over the seven trials (Page's L = 1618.0, z = 4.23, p <<.001). The ambiguous problems took slightly longer than the last training problem, though the difference was only marginally significant (Wilcoxon test, z = 1.64, p >.05). However, the choice of pieces to be removed showed that both groups persevered with the same tactics that they had used in training: overall, 92% of solutions were based on the same tactics; 14 out of the 20 participants used these tactics on both ambiguous problems, and the remaining participants were ties (Binomial test, p << .001). The final control problem took significantly longer to solve than the last training problem (Wilcoxon test, z = 2.61, p < .01).

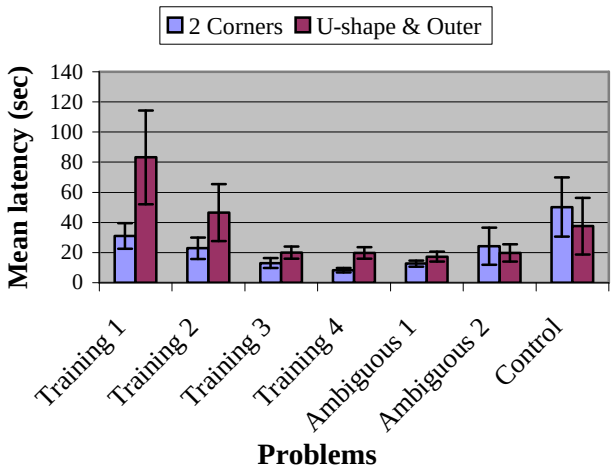


Figure 9: Mean latencies of the two groups in Experiment 3.

The results show that even when individuals have not acquired a deterministic strategy, their knowledge of tactics transfers to new problems. The training trials sufficed for the participants to develop tactical knowledge, and this knowledge constrained their search for solutions to the subsequent problems. With ambiguous problems, they readily succeeded though there was a marginal tendency for them to be slightly slower. In the case of the final problem, the tactics were inappropriate, and so they had to revert to a longer exploratory stage, which slowed them down.

General Discussion

Problem solving calls for creativity, because it calls for the generation of ideas that are novel (at least for the individual). In the case of, say, Duncker's X-ray problem (Duncker, 1945), psychologists can study only how individuals solve the problem for the first time. Hence, in order to investigate the development of strategies for solving problems, it is necessary to use problems that can be presented in a series that call for distinct solutions. In the past, such problems have been open to solution by a simple deterministic strategy of one sort or another (see, e.g., Luchins, 1942). In contrast, our goal was to examine the development of strategies for coping with problems that lie outside the bounds of a deterministic strategy – at least for our participants. We therefore studied problems that come in a series, just as many problems in daily life do – from the writing of computer programs to the search for a job.

Experiment 1 explored two constraints in shape problems, and found that symmetry of the initial problem shape, and the presence of salient solutions in the shape, facilitated problem solving. Experiment 2 showed that individuals do indeed normally begin to tackle such problems by exploring

the consequences of various tactical steps in a way akin to a neo-Darwinian procedure. They choose a step arbitrarily, and then evaluate its consequences in relation to the solution of the problem. More importantly, however, they acquire knowledge of the number of squares that the step removes. They pick up this knowledge whether or not the step is useful in the solution of the problem. And, as the think-aloud protocols also showed, they acquire some understanding of the importance of the p/s ratio in determining appropriate steps for a given problem. This ratio, between the number of pieces to be removed and the number of squares to be removed, constrains the set of useful steps at any point in solving a shape problem. As the theory postulates, a strategic shift then occurs. Individuals start to use their knowledge of tactical steps and the ratio to govern the generation of tactical steps. In this way, they are able to avoid useless false steps in the solution of problems. No participant, however, was able to converge completely on a neo-Lamarckian strategy that guaranteed a solution to any problem without false steps. Indeed, it is an open question whether such a strategy is possible for shape problems of any degree of complexity.

Experiment 3 corroborated the prediction that constraints in the form of tactical knowledge do transfer to new problems. Participants acquired tactical knowledge during training trials, and they continued to use these tactics for problems that could be solved in other ways. When the tactics were inappropriate, they were slowed down because they had to revert to a longer exploratory stage to find the right tactics. Luchins (1942) discovered that deterministic strategies transfer in this way. Our results generalize his findings to show that even when experience leads at best to a strategy that is not deterministic, the strategy nevertheless transfers.

Is the strategic shift an instance of *insight*? The answer depends on what one takes insight to be (cf. Weisberg, 1986; Kaplan & Simon, 1990; Isaak & Just, 1995; Ormerod, MacGregor, & Chronicle, 2002). When the current constraints fail to yield a solution, the shift yields new constraints on the generation of tactical steps. This change, in turn, can yield the solution of a problem. The development of strategies for shape problems accordingly reflects a series of small insights in which constraints are changed as a result of strategic shifts.

Ackowledgments

This research was supported by a grant from the National Science Foundation to the second author to study strategies in reasoning (BCS-0076287). We thank Sam Glucksberg, Geoffrey Goodwin, Uri Hasson, Cathy Haught, Sanna Reynolds, and three anonymous reviewers for helpful comments.

References

Duncker, K. (1945). On problem solving. *Psychological Monographs*, 58(5), Whole number 270.

Isaak, M.I., & Just, M.A. (1995). Constraints on thinking in insight and invention. In Sternberg, R.J., & Davidson, J.E. (Eds.), *The Nature of Insight*. (p.281-325). Cambridge MA: Bradford books, MIT Press.

Johnson-Laird, P.N. (1993). *Human and Machine Thinking*. New Jersey: Lawrence Erlbaum Associates.

Katona, G. (1940). *Organizing and Memorizing*. New York: Columbia University Press.

Kaplan, C.A., & Simon, H.A. (1990) In search of insight. *Cognitive Psychology*, 22, 374-419.

Luchins, A.S. (1942). Mechanization in problem-solving: the effect of Einstellung. *Psychological Monographs*, 54(6): 95.

Newell, A., & Simon, H.A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.

Ormerod, T.C., MacGregor, J.N., & Chronicle, E.P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 791-799.

Weisberg, R.W. (1986) *Creativity: Genius and Other Myths*. New York: Freeman.

Decision-Making on the Full Information Secretary Problem

Michael D. Lee, Tess A. O'Connor and Matthew B. Welsh
{michael.lee,tess.oconnor,matthew.welsh}@psychology.adelaide.edu.au
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

The secretary problem is a recreational mathematics problem, suited to laboratory experimentation, that nevertheless is representative of a class of real world sequential decision-making tasks. In the 'full information' version, an observer is presented with a sequence of values from a known distribution, and is required to choose the maximum value. The difficulties are that a value can only be chosen at the time it is presented, that the last value in the sequence is a forced choice if none is chosen earlier, and that any value that is not the maximum is scored as completely wrong. We report a study of human performance on full information secretary problems with 10, 20 and 50 values in the sequence, and considers three different heuristics as models of human decision-making. It is found that some people achieve near-optimal levels of accuracy, but that there are individual differences in human performance. A quantitative evaluation of the three heuristics, using the Minimum Description Length criterion, shows inter-individual differences, but intra-individual consistency, in the use of the heuristics. In particular, people seem to use the heuristics that involve choosing a value when it exceeds an internal threshold, but differ in how they set thresholds. On the basis of these findings, a more general threshold-based family of heuristic models is developed.

Introduction

Many real world decision-making problems are sequential in nature. A series of choices is made available over time, and it is often efficient (and sometimes even necessary) to make a selection without waiting to be presented with all of the alternatives. On long cross-country drives, for example, people refill their cars at one of a sequence of towns on the route, without knowing the price of fuel at subsequent towns. This type of sequential decision has a continuous utility function. People aim to choose the cheapest price, and measure their success by how much their purchase exceeded this minimum.

Other sequential decision-making tasks have binary utility functions, where any incorrect decision is equally (and completely) incorrect. For example, consider being a witness for a police line-up, where, because of the circumstances of the case, the offender is known to be in the line-up. Police line-up policy demands that suspects are presented one at a time, may only be viewed once, and that a suspect must be identified at the time they are presented (e.g., Steblay, Deisert, Fulero, & Lindsay 2001). Suppose also (unrealistically, we hope) that the police insist that a suspect be identified, and indicate that they will force the identification of the last person in the line-up if none of the previous people are chosen. The aim is to choose the offender, and any misidentification has the equally bad outcome of selecting an innocent suspect.

This decision-making scenario has the same essential features as a recreational mathematics problem known as the 'secretary problem' (see Ferguson 1989 for a historical overview). In secretary problems, an observer is presented with a sequence of possible choices, and must decide whether to accept or reject each possibility in turn. The number of choices in the complete sequence is fixed and known, and only the rank of each possibility, relative to those already seen, is presented to the observer. If the observer chooses the best possibility in the sequence, their decision is correct, and any other choice is regarded as incorrect.

Variants of the secretary problem have been considered that change or relax different parts of the problem. In particular, the full information version of the secretary problem, sometimes known as the 'Cayley' problem, presents observers with a score from a known distribution for each possibility, and the goal is to choose the maximum score in the sequence. Rank information corresponds to the assumption that witnesses keep a relative ordering of people in line-ups, whereas value information corresponds to the assumption that witnesses evaluate some continuous measure of the probability that a person is the offender. In either case, the secretary problem has the important feature of using the same binary utility function as the line-up decision. The goal is to choose the actual offender, and any incorrect decision is equally wrong.

Problem Solving and Secretary Problems

Human performance on secretary problems is an interesting topic for cognitive science, for a number of reasons. It offers a well defined task, suited to labora-

tory experimentation, that nevertheless is ecologically representative of a class of real world situations. Because of their inherent complexity, secretary problems also provide an opportunity to study the relationship between rational analysis and heuristic strategies in human problem solving.

Most laboratory research on human problem solving has relied on artificial problems that are characterized by well-defined initial and terminating states that must be linked by a systematic, finite series of steps. Typically, these problems, like the 'Towers of Hanoi' or 'Cannibals and Missionaries', are deterministic, and have state spaces with combinatorially limited possibilities. A major focus of studying people's abilities to solve these tasks involves examining under what circumstances, if any, people make rational decisions. Violations of rationality are easy to measure, because the tasks permit a complete formal analysis. This approach to studying human problem solving assesses what Simon (1976) terms 'substantive' rationality: the ability of people to produce optimal final decisions. Typically, they do not address what Simon (1976) terms 'procedural' rationality—the efficiency of the processes required to make the decision—because of the limited combinatorial complexity of the problem.

More recently, however, some research has studied human performance on difficult combinatorial optimization problems, such as visually presented Traveling Salesperson Problems (TSPs), that have very large state spaces, and resist complete formal solution (e.g., MacGregor & Ormerod 1996; Vickers, Butavicius, Lee, & Medvedev 2001). In attempting to solve these problems, subjects are constrained both by the nature of the task (e.g., limited time), and by their cognitive capabilities (e.g., limited memory). In other words, their performance is constrained not only by the need to achieve a substantively rational outcome, but also by the need to use procedurally rational heuristic processes that are sufficiently fast and accurate, and are implementable with available cognitive resources. Procedural rationality offers an important additional constraint for understanding human problems solving processes, and for the development and evaluation of cognitive models of decision-making.

Secretary problems provide an opportunity to continue and extend this line of study. Because they are not inherently perceptual, secretary problems allow consideration of whether results obtained with problems like TSPs generalize to cognitively-based problem solving. Secretary problems also introduce uncertainty, and place demands on memory. While visual problems like TSPs are combinatorially large, the basic information about distances between points is always perceptually available in a complete and certain form to subjects. In contrast, the sequences of information in secretary problems are stochastic and presented only temporarily, requiring people to deal with uncertainty and rely on their memory.

Previous Research

Gilbert and Mosteller (1966) provide a thorough and useful overview of early mathematical analysis of several versions of the secretary problem. When only rank information is provided, the optimal decision rule takes the form of observing some fixed proportion of values in the sequence, remembering the maximum presented, and then choosing the first subsequent value that is greater, if one exists. Gilbert and Mosteller (1966, Table 2) detail the optimal 'cutoff' proportion for the initial sequence of observations, which depends upon the length of the sequence, but converges to the value $1/e \approx 0.368$. They also give the associated probability of making a correct decision using the optimal decision rule.

For the full information version, where values rather than ranks are presented, the optimal decision rule requires choosing the first value that exceeds a threshold level for its position in the sequence. Gilbert and Mosteller (1966, Tables 7 and 8) detail these optimal thresholds and the associated probabilities of making a correct decision. Since Gilbert and Mosteller's (1966) seminal work, a large literature has developed on mathematical analyses of a large number of variants on the secretary problem, often with a focus on the performance of heuristic decision rules (e.g., Freeman 1983).

Relatively less attention has been given to studying human performance solving secretary problems. Seale and Rapoport (1997) consider the rank information version of the problem with lengths of 40 and 80, and focus on the evaluation of plausible heuristic models of human decision-making. In an individual subject analysis, they found a parameterized version of the optimal cutoff rule provided the best fit. Kahan, Rapoport and Jones (1967) studied human performance on full-information versions of the problem with length 200, where the values were drawn from either a positively skewed, negatively skewed, or a uniform distribution. They found no evidence for the different distributions affecting the decisions made. They also compared individual and group decision-making, and found that decisions were made earlier in the sequence by individuals. Other empirical studies (e.g., Kogut 1990), make a large methodological departure by requiring subjects to sacrifice explicitly held resources to view additional presentations, usually because they are interested in applications of the problem to economic decision-making.

In this paper, we study human performance on the full information version of the secretary problem, where values are chosen from a uniform distribution. We consider problems of length 10, 20 and 50, under the binary utility function, but without any explicit

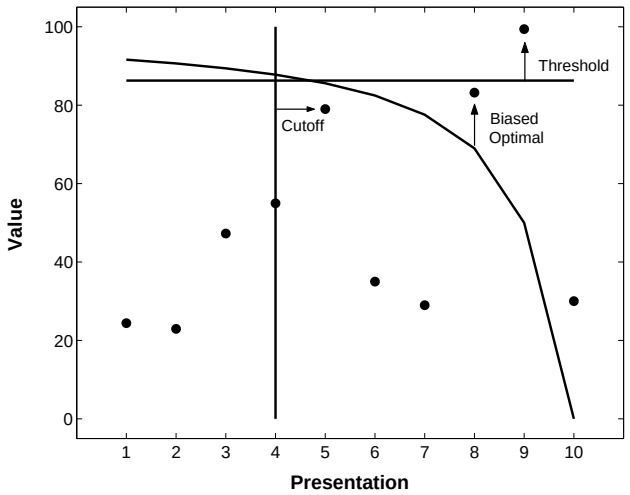


Figure 1: A sample secretary problem of length 10, with the sequence of values shown by filled circles, demonstrating the operation of the biased optimal (curved line), threshold (horizontal line) and cutoff (vertical line) heuristics.

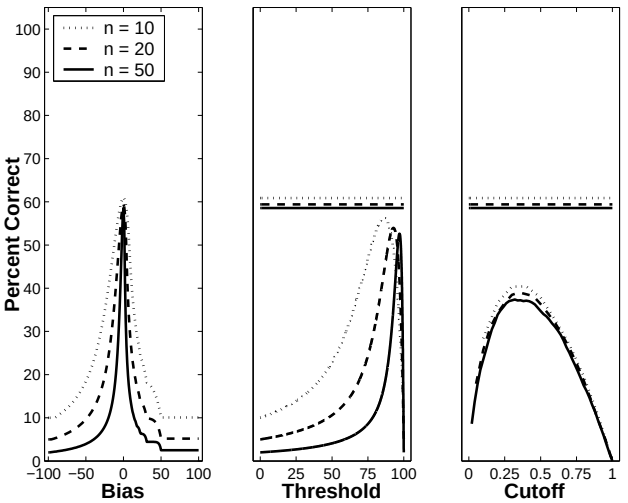


Figure 2: The accuracy of the heuristics, across their parameter spaces, for 10, 20 and 50 sequence length problems.

search cost. Our primary interest, like that of Seale and Rapoport (1997), is to develop and evaluate competing cognitive models of human decision-making.

Three Heuristics

We consider three possible heuristics as models of human decision-making. The first is a biased version of the optimal decision rule. This heuristic chooses the first value that exceeds a threshold level for its position in the sequence. The threshold levels correspond to the optimal values, for the given problem length, all shifted by the same constant. The second heuristic is inspired by Simon's (1956) notion of satisficing. It simply chooses the first value that exceeds a single fixed threshold. The third heuristic is inspired by the optimal decision rule for the rank information version of the secretary problem. It observes a fixed proportion of the values in the sequence, and remembers the maximum value up until this cutoff point. The first value that exceeds the maximum in the remainder of the sequence is chosen. For all three heuristics, if no value meets the decision criterion, the last value becomes the forced choice.

Figure 1 summarizes the functioning of the three heuristics on a problem of length 10. The sequence of values presented is shown by the filled circles. The threshold levels for the optimal heuristic (with no bias) follow the solid curve. The horizontal line shows the constant level used by the threshold heuristic. The vertical line shows the proportion used by the cutoff heuristic. Under these parameterizations, the biased optimal, threshold, and cutoff heuristics choose, respectively, the eighth, ninth, and fifth values presented.

The left panel of Figure 2 shows the accuracy of the biased optimal heuristic for bias values between -100 and 100 for problems of length 10, 20, and 50, calculated using the analytic method of Gilbert and Mosteller (1966, p. 55). At zero bias, the heuristic corresponds to the optimal decision rule, and so the maximum possible accuracy is obtained. The middle panel of Figure 2 shows the accuracy of the threshold heuristic for threshold values between 0 and 100 for problems of length 10, 20 and 50, calculated using the same analytic method. The maximum possible accuracy, corresponding to the use of the optimal decision rule, is shown for each problem length by the horizontal lines. Finally, the right panel of Figure 2 shows the accuracy of the cutoff heuristic for proportions between 0 and 1 for problems of length 10, 20 and 50, generated by simulation on a large sample of independently generated problems. Once again, the maximum possible accuracies are shown by the horizontal lines.

There are two observations worth making about the accuracy of the heuristics shown by Figure 2. It is clear that the threshold heuristic is capable of making better decisions than the cutoff heuristic. This is interesting, given that the cutoff heuristic is optimal for rank information secretary problems. It is also clear that the accuracy of both the biased optimal and threshold heuristics are very sensitive to their parameterizations, particularly for larger problem lengths.

Experiment

Participants Ten participants completed the experiment. There were 4 males and 6 females, with a mean age of 26.1 years.

Method Each participant completed the same three sets of problems. The first set contained 20 problems of length 10. The second contained 20 problems of length 20. The third set contained 20 problems of length 50. Participants always did the three sets in the same order—length 10, then 20, then 50—but the order of the 20 problems within each set was randomized across participants.

For each problem, the participants were told the length of sequence, and were instructed to choose the maximum value It was emphasized that (a) the values were uniformly and randomly distributed between 0.00 and 100.00, (b) a value could only be chosen at the time it was presented, (c) the goal was to select the maximum value, with any selection below the maximum being completely incorrect, and (d) if no choice had been made when the last value was presented, they would be forced to choose this value. As each value was presented, its position in the sequence was shown, together with 'yes' and 'no' response buttons. When a value was chosen, subjects rated their confidence in the decision on a nine point scale ranging from "completely incorrect" to "completely correct".

Results Table 1 summarizes the accuracy of the decisions made by all of the subjects for all of the problems. The average accuracy for the 20 problems in each set is given, together with averages across all problems for each subject, and across all subjects for each problem length. There are three observations worth making about these results. First, some subjects achieve levels of accuracy competitive with the optimal decision rule. Secondly, there appear to be individual differences between the subjects, with a range in average accuracy from 33% to more than 60%. Thirdly, there is some suggestion that human performance worsens as the problem length increases, even after accounting for the slightly decreased accuracy of the optimal decision rule.

Table 1: Accuracy of human decisions, showing the percentage of correct answers for each participant on each set of problems. Average accuracy for each participant, and for each problem length are also shown.

Participant	$n = 10$	$n = 20$	$n = 50$	Mean
1	65	65	55	61.37
2	45	45	20	36.67
3	55	45	50	50.00
4	40	35	25	33.33
5	55	35	55	48.33
6	65	45	20	43.33
7	45	60	50	51.67
8	55	50	45	50.00
9	70	55	55	60.00
10	50	35	55	46.67
Mean	54.50	47.00	43.00	

Model Evaluation One compelling aspect of the model evaluation undertaken by Seale and Rapoport (1997) is that it was done at the level of individual subjects, rather than by averaging decisions across subjects. As noted by Estes (1956), averaging non-linear decision processes in the presence of noise, and with significant individual differences, acts to corrupt the form of the empirical data being modeled. Because these criteria are likely met in the current problem, we also undertook individual subject evaluation of the biased optimal, threshold and cutoff heuristics.

A potential criticism of Seale and Rapoport (1997) is that the quantitative component of their model evaluation relied solely on the ability of a heuristic, at one or more parameterizations, to match the decisions made by a subject. As argued by Roberts and Pashler (2000), measures of goodness-of-fit fail to account for important quantifiable components in model selection. In particular, it is important also to assess the complexity of parameterized models, to ensure that good fit to empirical data does not merely arise because a model is so complicated that it can fit any data, including data that are never observed.

In model theoretic terms, there are clear differences in the complexity of the three heuristics being considered. For the set of 20 length 10 problems given to subjects, there are 10^{20} possible combinations of decisions. The biased optimal, threshold, and cutoff heuristics can predict, respectively, 78, 60, and 9 of these possibilities by varying their parameters. Similar differences in complexity hold for the longer problem lengths, with 88, 70 and 17 data distributions being indexed by the parameters for the length 20 problems, and 121, 90 and 30 for the length 50 problems. Accordingly, any superiority in the ability of the biased optimal heuristic over its competitors, or in the threshold over the cutoff heuristic, could possibly be due to greater complexity, rather than fundamentally capturing regularities in the empirical data.

These concerns are best addressed using advanced model selection methods (e.g., Pitt, Myung, & Zhang 2002), which provide criteria for choosing between models in ways that consider both goodness-of-fit and complexity. One interesting challenge in doing this is for the current models is that they are deterministic, and do not specify an error theory. This means that various probabilistic model selection criteria, such as Bayes Factors (e.g., Kass & Raftery 1995), Minimum Description Length (MDL: e.g., Grünwald 2000)

Table 2: Minimum Description Length (MDL) criteria values for the Biased Optimal (BO), Threshold (Th) and Cutoff (Cu) models, measured against the decision made by the ten participants on each problem length. Bold entries highlight strong evidence in favor of the preferred model.

	$n = 10$			$n = 20$			$n = 50$		
	BO	Th	Cu	BO	Th	Cu	BO	Th	Cu
1	32.7	**26.3**	40.1	34.4	**25.6**	52.6	34.6	**23.2**	60.6
2	**19.4**	32.4	33.2	38.0	40.4	45.0	34.6	32.8	60.6
3	35.4	**26.3**	35.7	38.0	29.7	35.2	29.7	32.8	47.5
4	**15.3**	35.1	42.0	26.3	25.6	47.8	**18.7**	32.8	44.1
5	32.7	32.4	40.1	30.4	29.7	50.3	52.0	48.9	60.6
6	**19.4**	29.5	38.0	**21.8**	29.7	42.1	29.7	28.1	43.8
7	26.6	22.9	40.1	26.3	21.2	41.5	18.7	17.8	35.8
8	32.7	26.3	40.1	41.5	**33.5**	50.3	**12.5**	23.2	47.8
9	26.6	32.4	38.0	26.3	25.6	52.6	24.4	23.2	36.0
10	**29.8**	37.6	40.1	**21.8**	37.0	52.6	34.6	**28.1**	43.8

or Normalized Maximum Likelihood (Rissanen 2001), are not immediately applicable. Grünwald (1999), however, develops a model selection methodology that overcomes these difficulties. He provides a principled technique for associating deterministic models with probability distributions, through a process called 'entropification', that allows MDL criteria for competing models to be calculated.

Table 2 shows the MDL values found by applying Grünwald's (1999) method to all three heuristics, taking each subject individually, and considering each problem length separately. Lower MDL values indicate better models, and differences between values can be interpreted on the log-odds scale. This means, for example, that the threshold heuristic (MDL value 26.3) provides an account that is about 600 times more likely than that provided by the biased optimal heuristic (MDL value 32.7) for the decisions made by the first subject for the length 10 problems, since $e^{32.7-26.3} \approx 600$. Kass and Raftery (1995) suggest, without being prescriptive, that a difference of six or more in the log-odds given by MDL values corresponds to 'strong' evidence in favor of the preferred model. Adopting the same standard, Table 2 highlights in bold those instances where the MDL for one of the heuristics provides strong evidence in its favor compared to both of the others.

There are several conclusions that can be drawn from this analysis. First, despite its simplicity, the cutoff heuristic does not provide a good model of the human decisions. For almost every subject and every problem length, it has the greatest MDL value, and often is so much larger as to provide strong evidence against its suitability. Secondly, there is clear evidence of inter-individual differences in the use of the biased optimal or threshold heuristics. There are approximately as many instances, for each problem length, where the biased optimal or threshold heuristic is strongly favored as an account for an individual subject. Thirdly, there is also some evidence of intra-individual consistency in using the biased optimal or threshold heuristic. This is because, in most instances, strong preferences favor the same heuristic for the same subject on different problem lengths.

Once the MDL criteria have been used to control for effects of model complexity, it is sensible to examine the goodness-of-fit of the heuristics. This was done by considering the average percentage of correct predictors made by each heuristic, for just those participants with MDL values favoring the heuristic. The biased optimal heuristic correctly predicted an average of 81%, 78% and 88% of participant decisions for, respectively, the 10, 20 and 50 length problems. The threshold heuristic correctly predicted an average of 74%, 78% and 79% of decisions. These results suggest that, while the heuristics may not provide a complete account of human performance, they do capture important regularities in the decision-making data.

Where there is strong evidence for a participant using either the biased optimal or threshold heuristic, it is also worthwhile to examine the parameter values used. For participants using the biased optimal heuristic, the bias parameter was always negative, indicating they underestimated the optimal threshold value for each position in the sequence. As the problem length increased from 10 to 20 then 50, however, the average bias changed from -5.1 to -1.8 then -1.9. This suggests that, for the longer sequences, participants were better calibrated to the optimal curve. For participants using the threshold heuristic, the average best-fitting threshold increased from 88.1 to 93.2 then 94.6. These values compare to theoretically optimal thresholds of 86.4, 92.6 and 97.2, as shown in the middle panel of Figure 2. It is clear that participants are sensitive to the need to increase the threshold as the length of the sequence increases, and do not seem either to under- or over-estimate the optimal value. It should be acknowledged, however, that these parameter values analyses are based on limited data, and additional data are required to confirm the suggested trends in this experiment, as well as to ascertain whether there are significant individual differences that also need to be considered.

Discussion

This study constitutes a first attempt to understand human decision-making on the full information version of the secretary problem. A first contribution of the study is to reject the usefulness of the cutoff heuristic, on both theoretical and empirical grounds, as an

account of human decision-making. This is a worthwhile finding, given that Seale and Rapoport (1997) found good evidence for people using this strategy on the rank information version of the secretary problem.

More importantly, it seems clear that both the biased optimal and threshold heuristics do capture something fundamental about human decision-making on the full information version. Both heuristics take the form of choosing the first value that exceeds a threshold, with the key difference being that the biased optimal heuristic uses thresholds that are sensitive to the position in the sequence, rather than being fixed.

Indeed, the biased optimal and threshold heuristics represent the two extremes of a continuum of threshold-based decision-making heuristics. Instead of using a constantly changing or a fixed threshold, it is possible for a decision process to use a small number of thresholds, and apply each to a sub-sequence of the presented values. For example, for a problem of length 10, a heuristic might apply one threshold for the first five values, decrease it for the next three values, and finally decrease it again for the penultimate value[1]. These sorts of heuristics seem likely to have complexity that lies somewhere between that of the biased optimal and threshold heuristics. It may well be the case that human performance is best explained by an account that is more sophisticated than the threshold heuristic, but does not have the full complexity of the biased optimal approach.

A final interesting problem for future research is whether the observed individual differences in accuracy are related to more traditional measures of problem solving ability and psychometric intelligence. In the everyday world, the ability to solve practical problems is generally regarded as an expression of intelligence. There is some evidence (e.g., Vickers *et al.* 2001) of a relationship between solution quality on TSPs and measures of IQ. Given that secretary problems are representative of a class of real world sequential decision-making tasks, they allow the possibility that there is a similar relationship for non-perceptual tasks to be examined.

Acknowledgments

We thank Helen Braithwaite, Marcus Butavicius, Matt Dry, and Douglas Vickers.

References

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin 53*(2), 134–140.

Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science 4*(3), 282–296.

Freeman, P. R. (1983). The secretary problem and its extensions: A review. *International Statistical Review 51*, 189–206.

Gilbert, J. P., & Mosteller, F. (1966). Recognizing the maximum of a sequence. *American Statistical Association Journal 61*, 35–73.

Grünwald, P. (1999). Viewing all models as 'probabilistic'. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT' 99)*, Santa Cruz. ACM Press.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology 44*(1), 133–152.

Kahan, J. P., Rapoport, A., & Jones, L. V. (1967). Decision making in a sequential search task. *Perception & Psychophysics 2*(8), 374–376.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kogut, C. A. (1990). Consumer search behavior and sunk costs. *Journal of Economic Behavior and Organization 14*, 381–392.

MacGregor, J. N., & Ormerod, T. C. (1996). Human performance on the traveling salesman problem. *Perception & Psychophysics 58*, 527–539.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*(5), 1712–1717.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review 107*(2), 358–367.

Seale, D. A., & Rapoport, A. (1997). Sequential decision making with relative ranks: An experimental investigation of the "Secretary Problem". *Organizational Behavior and Human Decision Processes 69*(3), 221–236.

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review 63*, 129–138.

Simon, H. A. (1976). From substantive to procedural rationality. In S. J. Latsis (Ed.), *Method and Appraisal in Economics*, pp. 129–148. London: Cambridge University Press.

Steblay, N. M., Deisert, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior 25*, 459–474.

Vickers, D., Butavicius, M. A., Lee, M. D., & Medvedev, A. (2001). Human performance on visually presented traveling salesman problems. *Psychological Research 65*, 34–45.

[1]Because the last value is always a forced choice, the threshold is always effectively zero for any heuristic.

Incremental Construction of an Associative Network from a Corpus

Benoît Lemaire (Benoit.Lemaire@upmf-grenoble.fr)
L.S.E., University of Grenoble 2, BP 47
38040 Grenoble Cedex 9, France

Guy Denhière (denhiere@up.univ-mrs.fr)
L.P.C & C.N.R.S. Université de Provence
Case 66, 3 place Victor Hugo
13331 Marseille Cedex, France

Abstract

This paper presents a computational model of the incremental construction of an associative network from a corpus. It is aimed at modeling the development of the human semantic memory. It is not based on a vector representation, which does not well reproduce the asymmetrical property of word similarity, but rather on a network representation. Compared to Latent Semantic Analysis, it is incremental which is cognitively more plausible. It is also an attempt to take into account higher-order co-occurrences in the construction of word similarities. This model was compared to children association norms. A good correlation as well as a similar gradient of similarity were found.

Introduction

A computational model of the human semantic memory may be valuable for its ability to mimic the human semantic representations, but also for its ability to mimic the *construction* of these representations over a long period of time. Not all models possess both features. For instance, symbolic formalisms like semantic networks had proven to be interesting for representing human knowledge but they do not tell us how human beings build such representations over their life. Several computational models of both the representation and construction of the human semantic memory have been proposed in the recent years. Some of them are based on a general common mechanism that rely on a huge input, composed of examples of associations between words. The statistical analysis of the occurrences of each word within well-defined units of context leads to a computational representation of association links between words. The representation of word meanings per se is not of significant interest, it is rather their association links which combined will form a model of the long-term semantic memory.

These models can be distinguished along six features:

1. the kind of input they are based on (either a corpus or word association norms);
2. the knowledge representation formalism (either vector-based or network-based);
3. the way a new context is added to the long-term semantic memory (incrementally or not);
4. the unit of context in which co-occurrence information is considered (either a paragraph or a sliding window);
5. the use or not of higher-order co-occurrences;
6. compositionality: the way the meaning of a text can be inferred from the meaning of its words.

After a description of existing models, we will discuss these features, present our model and describe an experiment that aims at comparing our model to human data.

Existing computational models of the construction of semantic representations

Latent Semantic Analysis

LSA (Landauer, 2002) takes as input a corpus of free texts. The unit of context is the paragraph. The analysis of the occurrences of each word within all paragraphs leads to a representation of the meaning of words as vectors, which is well suited for drawing semantic comparisons between words. The underlying mechanism (singular value decomposition of the word-paragraph occurrence matrix) implicitly takes into account higher-order co-occurrences (Kontostathis & Pottenger, 2002). Compositionality in this model is straightforward: the meaning of a text is a linear combination of the meaning of its words. There is however no way of updating the semantic space with a new unit of context without redoing the whole process. LSA's semantic representations have been largely tested in the literature (Foltz, 1996 ; Wolfe et al., 1998). This model can account for some mechanisms of the construction of knowledge (Landauer & Dumais, 1997).

Hyperspace Analogue to Language

HAL (Burgess, 1998) is also a model of the semantic memory. It is similar to LSA except that (1) it does not take into account higher-order co-occurrences since vectors are just direct co-occurrence vectors; (2) the unit of context is a sliding window of a few words which takes into account the lexical distance between words and (3) updating the semantic space with a new paragraph can be done easily.

Sparse Random Context Representation

SRCR (Sahlgren, 2001, 2002) is also based on the use of a sliding window applied to a large corpus. Words have an initial random vector representation (1,800 dimensions), which is updated with the vectors of the co-occurring words: they are all added to the current word, but with a multiplying factor which depends on their distance to the current word within the window. The way the initial

representation is computed is important: all 1,800 values are set to 0 except eight which are randomly selected and set to 1. This method is intrinsically incremental. It has better results than LSA on the famous TOEFL test. However, it does not take into account higher-order co-occurrences.

Word Association Space

WAS (Steyvers, Shiffrin & Nelson, in press) is not based on a corpus but on association norms providing associates for 5,000 words. The authors applied scaling methods to these data in order to assign a high-dimensional representation to each word. In particular, they relied on singular value decomposition, the mathematical procedure also used by LSA. The idea is similar to LSA: words that appear within similar contexts (i.e. words with similar associative relationships) are placed in similar regions in the space. WAS appeared to be a better predictor of memory performance than LSA.

Features

We will now discuss the six previous features in order to sketch out a model of construction and representation of the long-term semantic memory that would attempt to overcome existing limits.

Input

A corpus of free texts as input is cognitively more plausible than association norms or even a sublanguage of a few propositions (Frank et al. 2003). As humans, we do not obviously construct our semantic representations solely from written data (Glenberg & Robertson, 2000), but there is currently no formalism able to model all perceptual data such that they can be processed by a computational model. In addition, written data, although it is not perfect, seems to cover a large part of our semantic representations (Landauer, 2002).

Representation

Most models are based on a vector representation of word meaning. Dimensions of the semantic space can be the result of a statistical analysis which keeps hundreds of dimensions like in LSA or SCRC (they are therefore unlabelled), the most variant words as in HAL, the most frequent ones (Levy & Bullinaria, 2001) or even a predefined subset of words, either taken from a thesaurus (Prince & Lafourcade, 2002) or selected as being the most reliable across various sub-corpora (Lowe & McDonald, 2000).

One major interest of the vector representation is that it offers a simple way to measure the similarity between words. The angle between the corresponding vectors or its cosine are generally used.

One drawback of the vector representation however is the difficulty to determine the words that are similar to a given word or, say differently, the words that are activated in memory. It requires the scanning of all vectors in order to find the closest ones, which is both computationally and cognitively not satisfactory. A direct link between a word and its associates should exist in a plausible model of the semantic memory.

Another problem with the vector representation is that similarity is symmetrical: similarity(A,B)=similarity(B,A). This is not coherent with psycholinguistic findings showing that semantic similarity is not a symmetrical relation (Tversky, 1977). For instance, *bird* is a very close neighbor of *swallow*, but the opposite is not so obvious.

A network of words with simple numerical oriented links between nodes (what is called an oriented graph in graph theory) would be better for that purpose. Numerical links would represent semantic similarities. Such a basic network would offer a direct connection between a word and its neighbors and represent differently similarity(A,B) and similarity(B,A).

Memory updating

A model of the construction of the semantic memory should describe the way processing a new piece of written data affects the representation of the long-term memory. Some models like Latent Semantic Analysis are not incremental, which means that the whole process needs to be restarted in order to take into account a new context. Actually, a new paragraph can easily be represented by a vector in this model, by a simple linear combination of its words, but this operation does not affect at all the semantic space. Incremental models are much more cognitively plausible: processing new texts should modify, even slightly, the semantic memory.

Unit of context

The semantic relations between words are constructed from the occurrences of words within contexts. The size of such contexts plays an important role. Psychological experiments as well as computer simulations (Burgess, 1998) tend to consider that a context composed of a few words before and after the current word is reasonable. However, computational constraints have led some models to consider a whole paragraph as a unit of context, which is probably a too large unit. Latent Semantic Analysis is such a model. The use of a sliding window allows models like HAL or SRCR to take into account the distance between words within the window, whereas approaches based on paragraphs deal with bags of words.

Higher-order co-occurrences

It has been shown that higher-order co-occurrences play an important role (Kontostathis & Pottenger, 2002) in the latent structure of word usage. Two words should be considered associated although they never co-occur in context units, provided that they occur within similar contexts. A is said to be a second-order co-occurrence of B if it co-occurs with C which also co-occurs with B. If C were a second-order co-occurrence of B, A would be considered as a third-order co-occurrence of B, etc.

By means of the singular value decomposition procedure, LSA semantic similarity indeed involves higher-order co-occurrences (Lemaire & Denhière, submitted). Other approaches such as SCRC or HAL do not.

Table 1: Features of different models

	Input	Representation	Memory updating	Unit of context	higher-order co-occurrences	Compositionality
LSA	corpus	vectors	not incremental	paragraph	yes	easy
HAL	corpus	vectors	incremental	sliding window	no	easy
SCRC	corpus	vectors	incremental	sliding window	no	easy
WAIS	association norms	vectors	not incremental	N/A	no	easy
ICAN	*corpus*	*network*	*incremental*	*sliding window*	*yes*	*hard*

Compositionality

Compositionality is the ability of a representation to go from words to texts. The vector representation is very convenient for that purpose because the linear combination of vectors still produces a vector, which means that the same representation is used for both words and texts. This might be a reason why vector representations are so popular. On the contrary, symbolic representations of word meaning like semantic networks do not offer such a feature: it is not straightforward to build the representation of a group of words from the individual representations of words, especially if the representation is rich, for instance with labelled links.

Summary

Table 1 describes some of the existing models along the previous six features. We present ICAN, our proposal, at then end of the next section.

ICAN

Basic mechanisms

Like others, this model takes as input a corpus of free texts and produces a computational representation of word meanings. This model is based on a network representation, which we believe is more accurate in modeling the process of semantic activation in memory. The idea is to associate to each word a set of neighbors as well as their association weights in [0..1], exactly as in rough semantic network. The model is incremental which means that the set of connected words for each word evolves while processing new texts. In particular, new words can be added according to the co-occurrence information and other words can be ruled out if their association strengths with the current word become too low.

Links between words are updated by taking into account the results of a previous simulation on 13,637 paragraphs of a corpus (Lemaire & Denhière, submitted), which showed that:

- co-occurrence of W_1 and W_2 tends to strongly increase the W_1-W_2 similarity;
- occurrence of W_1 without W_2 or W_2 without W_1 tends to decrease the W_1-W_2 similarity;
- second and third-order co-occurrence of W_1 and W_2 tends to slightly increase the W_1-W_2 similarity.

In our model, a sliding window is used as a unit of context. Therefore, each word of the corpus is considered with respect to its preceding and following contexts. The size of the window can be modified. For the sake of simplicity, we will not use the third-order co-occurrence effect. The algorithm is the following:

For each word W, its preceding context $C_1..C_k$ and its following context $C_{k+1}..C_{2k}$ (the sliding window therefore being $[C_1\ C_2\ ...\ C_k\ \mathbf{W}\ C_{k+1}\ C_{k+2}...\ C_{2k}]$):

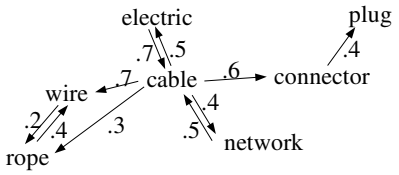
Direct co-occurrence effect: reinforce the link W-C_i (if this link does not exist, create it with a weight of 0.5, otherwise increase the weight p by setting it to $p+(1-p)/2$;

Second-order co-occurrence effect: let p be the weight of the W-C_i link. For each M linked to C_i with weight m, reinforce the link W-M (if such a link does not exist, create it with a weigth of $p.m$, otherwise, increases the weight q by setting it to $q+A(1-q)(p.m)$, A being a parameter;

Occurrence without co-occurrence effect: reduce the links between W and its other neighbors (if the weights were p, set them to a fraction of p, e.g., $0.9p$). If some of them fall under a threshold (e.g., .1), then remove these links.

Example

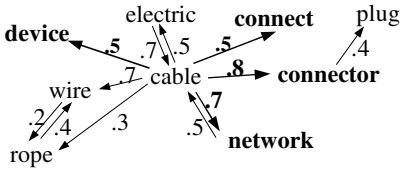
As an example, consider the following association network, which is the result of processing several texts:



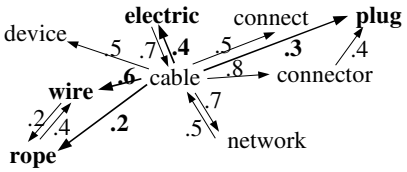
The new text being analyzed is:

... if you have such a device, connect the cable to the network connector then switch...

We now describe how this text will modify the association network, according to the previous rules. Suppose a window of size 5 (2 preceding words, 1 current word, 2 following words). Since functional words are not taken into account, the current window is then [device, connect, **cable**, network, connector], *cable* being the current word. The direct co-occurrence effect leads to reinforce the links between *cable* and the four co-occurring words. Two of them are new links, while others are existing links whose weights are simply increased. The network becomes:



The second-order co-occurrence effect reinforces the links between *cable* and all words connected to one of its four co-occurring word. In this small example, this is only the case for the word *plug*. Finally, the occurrence without co-occurrence effect leads to a decrease of the links between cable and its other neighbors. The network is then:



The next current word is *network*, the window is [connect, cable, **network**, connector, switch] and the process repeats again.

Measure of similarity

Similarity between words W_1 and W_2 is the combination (i.e. the product) of the links of the shortest path between W_1 and W_2. If W_2 is connected to W_1, it is just the weight of the link; if W_1 is connected to Z which is connected to W_2, it is the combination of the two weights. If W_2 does not belong to the neighbors of W_1's neighbors it is probably sufficient to set the semantic similarity to 0. Since the graph is oriented (the link weight between A and B might be different from the link weight between B and A), this way of measuring the similarity mimics the asymmetrical property of the human judgment of similarity better than the cosine between vectors.

Tests

Comparison to association norms

In order to test this model, we compared the association links it provides to human association norms. The corpus we relied on is a 3.2 million word French child corpus composed of texts that are supposed to reproduce the kind of texts children are exposed to: stories and tales for children (~1,6 million words), children productions (~800,000 words), reading textbooks (~400,000 words) and children encyclopedia (~400,000 words). All functional words were ruled out. Words whose frequency was less than 3 were not taken into account. The program is written in C, it is available on demand. Processing the whole corpus takes a few hours on a standard computer, depending on the window size.

Once the association network was built, we measured the similarity between 200 words and 6 of their associates (the first three and the last three), as provided by the de la Haye (2003) norms for 9 year-old children. The association value in these norms is the percentage of subjects who provided the associate. For instance, the six associates to *abeille(bee)* are:

- miel(honey): 19%
- insecte(insect): 14%
- ruche(hive): 9%
- animal(animal): 1%
- oiseau(bird): 1%
- vole(fly): 1%

Actually, 16 words were not part of the corpus. Only 1184 pairs of words were therefore used.

We then compared these values to the similarity values provided by the model. We had two hypotheses. First, the model should distinguish betwen the three first associates and the last ones and there should be a gradient of similarity from the first one to the last ones. Second, there should be a good correlation between human data and model data.

Several parameters have to be set in the model. The best correlation with the human data was obtained with the following parameters (see the algorithm presented earlier):

- window size = 11 (5 preceding and 5 following words);
- co-occurrence effect: $p \rightarrow p+(1-p)/2$;
- 2nd-order co-occurrence effect : $p \rightarrow q+.02(1-q)(p.m)$;
- occurrence without co-occurrence effect : $p \rightarrow .9p$.

Using these parameters, the average similarity values between stem words and associates, as well as the children data, are the following:

	1st associates	2nd associates	3rd associates	Last associates
ICAN	.415	.269	.236	.098
Norms	30.5	13.5	8.2	1

All model values are highly significantly different, except for the 2nd and 3rd associates which differ only at the 10% level. Our model reproduces quite well the human gradient of association.

We also calculated the coefficient of correlation between human data and model data. We found an interesting significant correlation: $r(1184)=.50$.

The exact same test from the same corpus was also applied to Latent Semantic Analysis. Results are the following:

	1st associates	2nd associates	3rd associates	Last associates
LSA	.26	.23	.19	.11

Similarities between the stem word and the first associates appear stronger in the ICAN model. LSA' scorrelation with human data is $r(1184)=.39$, which is worse than our correlation.

Similarity as direct co-occurrence

One can wonder whether the similarity could be mainly due to the direct co-occurrence effect. Similarity between words is indeed often operationalized in psycholinguistic researches by their frequency of co-occurrence in huge corpus. Experiments have indeed revealed the correlation between both factors (Spence & Owens, 1990). However, this shortcut is questionable. In particular, there are words that are strongly associated although they never co-occur. Burgess & Lund (1998) mentioned the two words *road* and *street* that almost never cooccur in their huge corpus although they are almost synonyms. In a 24-million words French corpus from the daily newspaper *Le Monde* in 1999, we found 131 occurrences of *internet*, 94 occurrences of *web*, but no co-occurrences at all. However, both words are strongly associated. Edmonds (1997) showed that selecting the best typical synonym requires that at least second-order co-occurrence is taken into account. There is clearly a debate: is the frequency of co-occurrence a good model of word similarity?

In order to test that hypothesis, we modified our model so that only direct co-occurrences are taken into account: the 2nd order co-occurrence effect as well as the occurrence without co-occurrence effect were inhibited. Results are the following:

	1st assoc.	2nd assoc.	3rd assoc.	Last assoc.
ICAN (only direct co-occurrences)	.903	.781	.731	.439

The gradient of similarity is still there but the correlation with human data is worse ($r(1184)=.39$). This is in accordance with our previous findings (Lemaire & Denhière, submitted) which show that the frequency of co-occurrence tends to overestimate semantic similarity.

Effect of second-order co-occurrence

Another test consisted in measuring the effect of second-order co-occurrences. This time, we only inhibited this effect in order to see whether the loss would be significant. Results are presented in the next table:

	1st assoc.	2nd assoc.	3rd assoc.	Last assoc.
ICAN (no 2nd-order co-occurrences)	.371	.225	.191	.056

Correlation with human data was not significantly different from the full model. It only decreased from .50 to .48. This means that second-order co-occurrences do not seem to have much effect in this simulation. One reason might be due to the mathematical formula we used to model higher-order co-occurrences. It might not be the right one. Another reason could be that we only implemented second-order co-occurrence effects. Third and higher-order co-occurrence effects might play a much more significant role than could be expected. A final reason could be that higher-order co-occurrence does not play any role. But, how then could we explain the high similarity between words that almost never co-occur? More experiments and simulations need to be carried out to investigate this issue.

Window size

We also modified the model in order to shed light on the role of the window size. Results are as follows:

Window size	Correlation with human data
3 (1+1+1)	.34
5 (2+1+2)	.38
7 (3+1+3)	.44
9 (4+1+4)	.48
11 (5+1+5)	.50
13 (6+1+6)	.49
15 (7+1+7)	.47

We found that the best window size is 11 (5 preceding words and 5 following words). This is in agreement with the literature: Burgess (1998) as well as Lowe and McDonald (2000) use a window of size 10, Levy and Bullinaria (1998) found best performance for a window size from 8 to 14, according to the similarity measure they relied on.

Conclusion

This model could be improved in many ways. However, preliminary results are encouraging: the model produces better results than the outstanding Latent Semantic Analysis model on a word association test. In addition, it adresses two major LSA drawbacks. The first one has to do with the representation itself: the fact that LSA's associations are symmetrical is not satisfactory. A network representation seems better for that purpose than a vector representation. The second limitation of LSA concerns the way the semantic space is built. LSA is not incremental: adding a new piece of text requires that the whole process is run again. Like HAL or SCRC, ICAN has the advantage of being incremental.

ICAN's main limitation is related to compositionality. The construction of a text's representation is not straightforward, given the representation of its words. Representing every text as a simple function of its words, and in the same formalism, as in the vector representation, is very convenient since text comparisons are then easy to perform. Compared to other approaches, LSA is for instance very good at simulating the human judgment of text comparisons (Foltz, 1996). However, the cognitive plausibility of such a representation can be questioned. Do we really need the exact same representation for words and texts? Is it cognitively reasonable to go directly without any effort from words to texts? Why having two ways of processing texts: one which would be computationally costly (singular

value decomposition in LSA) and another one very quick (adding its words)?

A solution could be to process each new text by the mechanism described in this paper: a text would then be represented by a subgraph, that is by a small subset of the huge semantic network, composed of the text words, their neighbors and their links. An information reduction mechanism like the integration step of the construction-integration model (Kintsch, 1998) could then be used to condense this subgraph in order to retain the main information. This smaller subgraph would constitute the text representation. This way, there would be a single mechanism used to process a text, construct its representation and update the long-term semantic memory. However, much work remains to be done in that direction.

Once a corpus was processed, it would be interesting to study the resulting network structure. In particular, this structure could be compared to existing semantic networks, in terms of connectivity or average path-lengths between words, much like Steyvers & Tenenbaum (submitted) did recently.

Acknowledgements

We would like to thank Maryse Bianco, Philippe Dessus and Sonia Mandin for their valuable comments on this model, as well as Emmanuelle Billier, Valérie Dupont and Graham Rickson for the proofreading .

References

Burgess, C. (1998). From simple associations to the building blocks of language: modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30, 188-198.

Burgess, C., Livesay, K & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25, 211-257.

de la Haye, F. (2003). Normes d' associations verbales chez des enfants de 9, 10 et 11 ans et des adultes. *L' Année Psychologique*, 103, 109-130.

Edmonds, P. (1997). Choosing the word most typical in context using a lexical co-occurrence network. *Meeting of the Association for Computational Linguistics*, 507-509.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28-2, 197-202.

Frank, S.L., Koppen, M., Noordman, L.G.M. & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science* 27(6), 875-910.

Glenberg, A. M. & Robertson, D. A., (2000). Grounding symbols and computing meaning: a supplement to Glenberg & Robertson. *Journal of Memory and Language,* 43, 379-401.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Kontostathis, A. & Pottenger, W.M. (2002). Detecting patterns in the LSI term-term matrix. *Workshop on the Foundation of Data Mining and Discovery, IEEE International Conference on Data Mining.*

Landauer T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.), *The psychology of Learning and Motivation, 41*, 43-84.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

Lemaire, B. & Denhière, G. (submitted). Effects of higher-order co-occurrences on semantic similarity of words.

Levy, J.P., Bullinaria, J.A. & Patel, M. (1998). Explorations in the derivation of semantic representations from word co-occurrence statistics. *South Pacific Journal of Psychology,* 10, 99-111.

Levy, J.P., Bullinaria, J.A. (2001). Learning lexical properties from word usage patterns: which context words should be used? In R. French & J.P. Sougne (Eds) *Connectionist Models of Learning, Development and Evolution: Proceedings* of *the Sixth Neural Computation and Psychology Workshop*, 273-282. London: Springer.

Lowe, W. & McDonald, S. (2000). The direct route: mediated priming in semantic space. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 675-680, New Jersey. Lawrence Erlbaum Associates.

Prince, V. & Lafourcade, M. (2003). Mixing semantic networks and conceptual vectors: the case of hyperonymy. In *Proc. of ICCI-2003 (2nd IEEE International Conference on Cognitive Informatics),* South Bank University, London, UK, August 18 - 20, 121-128.

Sahlgren, M. (2001). Vector-based semantic analysis: representing word meaning based on random labels. *Semantic Knowledge Acquisition and Categorisation Workshop at ESSLLI '01*, Helsinki, Finland.

Sahlgren, M. (2002). Towards a flexible model of word meaning. *AAAI Spring Symposium 2002*. March 25-27, Stanford University, Palo Alto.

Spence, D.P. & Owens K.C. (1990). Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research* 19, 317-330.

Steyvers, M., Shiffrin R.M., & Nelson, D.L. (in press). Word Association Spaces for predicting semantic similarity effects in episodic memory. In A. Healy (Ed.), *Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington DC: American Psychological Association.

Steyvers, M., & Tenenbaum, J. (submitted). Graph theoretic analyses of semantic networks: small worlds in semantic networks.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch & W., Landauer, T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25, 309-336.*

Integrating Spatial Language and Spatial Memory: A Dynamical Systems Approach

John Lipinski (john-lipinski@uiowa.edu)
John P. Spencer (john-spencer@uiowa.edu)
Larissa K. Samuelson (larissa-samuelson@uiowa.edu)
Department of Psychology, University of Iowa
Iowa City, IA 52242 USA

Abstract

The domain of spatial language is an ideal testing ground for proposals addressing the representational gap between perceptual-motor and language systems precisely because it is an unambiguous case of these systems coming together. To date, however, efforts addressing this representational gap within the domain of spatial language have generated conflicting results. Focusing on an "above" ratings task, we provide here a dynamical systems approach to spatial language performance and supporting empirical results that address this impasse. The development of a dynamical systems model linking spatial language and spatial memory is also discussed.

Representation and Spatial Language

A current focus in cognitive science is understanding how the sensory-motor and linguistic systems interact. Because spatial language brings words and physical space together so directly, it is the ideal vehicle for exploring this interaction. To date, two general approaches to representation speak to this issue of interaction in spatial language (Barsalou, 1999): amodal symbolic systems and perceptual symbol systems.

Amodal symbolic systems presume representational independence between symbolic processes like language and sensory-motor systems (Harnad, 1990; Anderson, 2000). The amodal view thus requires a transduction process that permits "communication" between linguistic and non-linguistic systems. This transduction process is best described by Jackendoff's representational interface (1992; 1996; 2002) in which communication between different types of representations (e.g. auditory and visual) is achieved through a process of schematization─ the simplifying and filtering out of information within one representational format for use in another representational system (Talmy, 1983). The representational interface approach ultimately permits abstract conceptual structures that can encode spatial representations but still capture the core characteristics of the symbolic view (e.g. pointers to sensory modalities, type-token distinctions, taxonomies).

There is significant empirical support for this perspective. Talmy (1983), for example, showed that language uses closed-class prepositions (such as "above", "below", or "near") to provide an abstracted, skeletal structure of a scene that narrows the listener's attention to a particular relationship between two objects by disregarding other available information (Talmy, 1983; Hayward & Tarr, 1995). Thus, in the sentence "The bike stood near the house", all of the specific information about the bike (e.g. size, shape, orientation) is disregarded and the bike is instead treated as a dimensionless point (Hayward & Tarr, 1995). As a result of this schematization, linguistic representations of relational states can be extended to a variety of visual scenes and objects with little regard to the individual object characteristics.

In contrast to transduction and the amodal approach, Barsalou's Perceptual Symbol Systems (1999) posits perceptual symbols: "records of neural states that underlie perception" (p.583) that are both inherently grounded in the given sensory modality and capable of replicating the flexible, productive, and hierarchical capacities of amodal symbolic systems. These perceptual symbols are implemented when top-down processes partially reactivate sensory-motor areas and organize the perceptual components around a common frame. Ultimately, perceptual components implement a simulator that captures both perceptual memories and core symbolic behaviors (e.g. type-token distinctions, hierarchies). Because these symbols are grounded in sensory-motor processes, they do not require pointers or transduction to become "meaningful".

A growing empirical literature supports Barsalou's (1999) PSS as well. For example, Stanfield and Zwaan (2001) argued that if symbolic, linguistic representations are integrated with perceptual symbol systems, people should be faster to recognize visual objects described in a sentence as the similarity between the perceived object and the description increase. Consistent with this prediction, they found that people were faster to recognize an object (e.g. a vertically oriented pencil) as part of a previous sentence when that sentence matched the orientation (e.g. He placed the pencil in the cup) than when it conflicted (e.g. He placed the pencil in the drawer). Visual information has also been shown to facilitate real-time resolution of temporarily syntactically ambiguous sentences (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), further evidence against a hard separation between linguistic and sensory systems. Finally, recent work by Richardson et al. (2003) shows that verbal stimuli interact with visual discrimination performance, additional evidence that linguistic processing can directly impact the processing of visual space.

In summary, the contrasting amodal and modal perspectives both appear to be substantially supported.

iven the clear contrast between the two theories, however, both cannot be correct. Thus, despite a vigorous debate and valuable empirical data on both sides, the fundamental question of how linguistic and non-linguistic systems relate remains unanswered.

Limits of the Current Approaches

Further consideration suggests two critical limits of the proposals and empirical support discussed above. First, they rely on descriptive, conceptual accounts of representational structure. Though critical at initial stages of theory development, the flexibility of conceptual accounts makes them ultimately difficult to critically test and falsify. Consequently, data collected in support of one view can be reinterpreted by the other view. Jackendoff (2002), for example, incorporated the resolution of syntactic ambiguity through visual processing (Tanenhaus, et al., 1995) using characteristics of a syntax-semantics interface.

The second, related limit of the current literature is treatment of representational structure in the abstract. In particular, with the exception of recent tests of the PSS theory (e.g. Richardson et al., 2003), spatial language studies have tended to focus on the nature of representational structure without considering the second-to-second processes that give rise to those structures. This can lead to an impasse because representations are not strongly grounded in task-specific performance. Consideration of an ongoing debate within spatial language illustrates this point. Because this debate is central to our empirical work, it is considered in some detail.

Evidence for Shared Representations

In order to explore the possible correspondence between the linguistic and sensory-motor representations of space, Hayward and Tarr (1995) conducted a series of experiments designed to compare how object relations are linguistically and visually encoded. In the first experiment, participants were presented with a visual scene depicting a referent object and a target object and asked to generate a preposition describing the relationship. Results suggested that the prototypical spatial positions for "above" and "below" lie along a vertical reference axis and prototypical spatial positions for "left" and "right" lie along a horizontal axis. In addition, use of these terms declines as target positions deviate from the horizontal and vertical reference axes.

Next, Hayward and Tarr built on these findings by using a preposition ratings task. In the ratings task, participants were asked to rate on a scale of 1 (least applicable) to 7 (most applicable) the applicability of a given spatial term (e.g. above) to a relationship between two objects. This ratings task is particularly valuable because it permits quantification and metric manipulation of an otherwise gross measure of linguistic output (e.g. above/not above). As such, this task provides a means of empirically bridging the continuity of sensory-motor representations with the discreteness of linguistic representations. Results from this ratings task showed strong metric effects of spatial language use around the vertical and horizontal axes. For instance, "above" ratings were highest along the vertical axis and systematically decreased as the target object's position deviated from the vertical axis. Hayward and Tarr concluded that this ratings gradient across spatial positions reflected the use of prototypical vertical and horizontal reference axes.

To compare the representational prototypes of spatial language with visual representations of space, Hayward and Tarr compared these findings with performance on location memory and same-different discrimination tasks. Importantly, the areas of highest spatial recall accuracy were vertically aligned with the reference axes used as prototypes in the ratings task. Performance in the same-different location task yielded similar findings, showing that discrimination was best along the vertical and horizontal axes. Collectively, data from these four experiments point to a shared representational spatial structure between linguistic and sensory-motor systems, a result consistent with Barsalou's PSS approach.

Evidence Against Shared Representations

Follow-up results from Crawford, Regier, & Huttenlocher (2000) present a different picture. To probe both linguistic and visual representations of space, they analyzed "above" ratings as well as spatial memory performance. Although results showed an "above" ratings gradient aligned with the vertical axis similar to that of Hayward and Tarr (1995), Crawford et. al. also found location memory bias away from the vertical axis when participants had to recall the locations of targets to the left and right of this axis. These researchers proposed that the cardinal axes that appear to function as prototypes in the linguistic task instead serve as category boundaries in the spatial memory task. Thus, while both linguistic and sensory-motor spatial representations use the same cardinal axes, these axes serve functionally distinct representational roles in the two tasks. It therefore appears that the linguistic and sensory-motor representations of space differ in critical ways, a conclusion consistent with an amodal representational interface perspective.

Considered together, these results illustrate the limits of dealing with representation in the abstract: both sets of researchers used similar experimental tasks and reported largely similar findings, yet they draw starkly different conclusions, conclusions that depend critically on abstract definitions of representational structure. Because we do not yet know the process that selects, creates, and encodes spatial prototypes nor the process used to create a spatial rating, we cannot go beyond abstract representational descriptions to make predictions about the similarities and differences across tasks. Notably, the failure to resolve this particular debate within spatial language mirrors the larger failure to resolve the modal-amodal conflict. In most general terms, the empirical support offered in both cases fails to delineate between the proposed accounts.

A Process Approach to Spatial Language

The theory and data discussed so far appear to be at an impasse, due in part to an emphasis on descriptive accounts of representational systems and a focus on representation in the abstract. To move beyond these fundamental limitations, the current proposal seeks to establish and test a process model that relates spatial memory and verbal performance. Such a process model can move beyond description and representation in the abstract and provide strong, testable predictions.

To lay the foundation of this proposed model, consider again the results of Crawford et al. (2000). The distinguishing result was the finding of spatial memory biases away from the vertical axis. They interpreted this movement away from midline to be a function of bias towards spatial categories. This interpretation is derived from Huttenlocher et al.'s (1991) Category Adjustment (CA) model. According to the CA model, people encode spatial location at two levels. The first level encodes fine-grained information about target location (e.g. angular deviation), while the second level encodes the region or category of target location. Specifically, the CA model proposes that people represent a central or prototypical value within a category that is most representative of that category. To remember a location, these two levels of detail are then combined to produce a remembered target location. Importantly, fine-grained and categorical information can be weighed differently. If, for example, the fine-grained detail is less certain, the prototype can be given more weight, resulting in a bias away from midline. Moreover, evidence from Huttenlocher et al. (1991) indicates that these spatial prototypes lie along the diagonal axes. According to Crawford et al. (2000), these spatial prototypes along the diagonals are the source of the observed spatial memory biases away from midline. Recall, however, that spatial prepositions maintained their highest applicability ratings along the vertical and horizontal axes. Thus, spatial prepositions appear to maintain prototypes along vertical and horizontal axes while spatial categories appear to maintain prototypes along the diagonal axes.

But must the drift away from midline observed in spatial memory performance result from spatial prototypes along the diagonal? A recent model suggests no. Specifically, the Dynamic Field Theory (DFT) (Spencer & Schöner, 2003; Schutte, Spencer, & Schöner, 2003) provides a formalized process account of spatial memory bias away from reference axes without positing prototypes. This model specifies how location-related activation is maintained in spatial working memory (SWM) during short-term delays and how perception and memory are integrated within this single representational system.

The DFT can be best understood within the context of a location memory task used to test predictions of the model. In this task, participants are seated at a large empty table and a spaceship-shaped object is flashed for 2 seconds on the table. After a variable delay, participants are asked to indicate the location of the ship using a computer mouse. Participants in this task show the same biases away from midline reported by Huttenlocher and colleagues. The focus of the DFT is to explain this performance through activation in the SWM field.

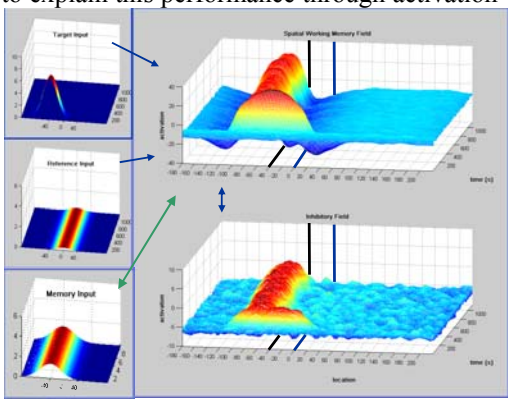


Figure 1 Dynamic Field Theory of spatial working memory

Figure 1 shows the structure of the DFT model. The large box shows the excitatory and inhibitory layers of neurons that together form the SWM field. Each layer has a collection of spatially tuned neurons that respond selectively to a specific location. Spatial location is indicated by position along the x-axis, where 0˚ is the center of the space; positive locations are rightward, and negative locations are leftward. The y-axis represents time which is moving away as a particular experimental trial proceeds from start to finish. The z-axis captures the activation of each neuron in the field.

In addition to the excitatory and inhibitory layers of the SWM field, there are input fields: target input, reference input, and memory input. The upper left portion of Figure 1 shows the target input which feeds into the excitatory layer of the SWM field. This target input turns on when the target is visible and turns off when the target is hidden. Figure 1 also shows the reference input. This reference input captures perception of the midline or vertical symmetry axis, the same axis central to the linguistic and non-linguistic representations of space discussed above. The third input is the long-term memory field which reflects the activation history of the SWM. This field also reciprocally feeds into the SWM field to impact real-time spatial memory processes.

The integration of these inputs in working memory is governed by an interaction function that determines how activation at one site in the SWM field influences activation at other sites. The DFT uses a local excitation and lateral inhibition function. Thus, activation at one site increases the activation of its neighbors and decreases the activation of sites further away. There are two main consequences of the interaction function. First, strong target input can lead to a self-sustaining peak of activation. These self-sustaining peaks of activation maintain themselves even after the target input is removed. In this way the field can maintain a memory of the target location during short-term delays.

The second consequence of the interaction function is that self-sustaining peaks can drift away from reference axes such as midline during memory delays. The process that gives rise to such delay-dependent spatial drift is illustrated in Figure 2. The short activation profile in this figure was generated by running a simulation of the model shown in

Figure 1 with only a single input—the reference input—and taking a time slice through the excitatory layer (top layer in the large panel of Figure 1) of the resultant SWM field (at time 8.00 s). Thus, this short activation profile reflects the influence the reference input has on each neuron in SWM at a particular moment in time (note that, because the reference input in Figure 1 is constant throughout the memory delay, the short activation profile actually captures the resultant influence of the reference input throughout the trial).

As can be seen in Figure 2, the resultant reference profile has stronger activation around midline; however, there are also two troughs in activation to the left and right of midline. These troughs cause systematic delay-dependent drift away from midline when targets are positioned to the left and right of this axis. This is captured schematically by the tall activation profile in Figure 2. As can be seen in Figure 2, this tall activation profile receives slightly more reference-related input on its left side than its right side. As a consequence, neurons on the left side of the activation peak will tend to join into the locally-excitatory interaction, while neurons on the right side of the peak will tend to be laterally inhibited.

The excitatory (top) layer of the SWM field in Figure 1 shows that as this interactive process propagates through time, activation peaks can spatially drift. In particular, Figure 1 shows a simulation of the model during a single trial to a -40º location. At the start of the trial, activation in the excitatory layer of SWM is relatively uniform because no strong inputs are present. At 2 s, the target appears at -40º and the strong target input associated with this event builds a peak of activation in SWM. Importantly, this activation peak sustains itself even after the target disappears at 4 s. And, during the ensuing memory delay, the peak drifts systematically away from midline (i.e., away from 0º). Note that this effect is partially counteracted by the long-term memory input at -40º.

In summary, the DFT provides a process-based alternative to Huttenlocher et al.'s (1991) category adjustment model. Critically, this model links spatial memory biases to a process that integrates remembered information in working memory with perceived reference frames, the same reference frames implicated in research by Hayward and Tarr (1995). As a result, the central argument against Hayward and Tarr's claim of shared structure between linguistic and non-linguistic representations of space—that memory is biased away from a category boundary—no longer follows obligatorily from the data. This provides the impetus to once again consider the potentially rich and direct connections between spatial memory and spatial language.

Connecting the DFT with Spatial Language

Inspired by our model of spatial working memory, we recently conducted a set of experiments designed to investigate Hayward and Tarr's (1995) claim that linguistic and non-linguistic representations overlap, not by examining representational structures in the abstract, but by considering the specific representational structures that emerge in our formalized process model. In particular, we asked whether the processes that create delay-dependent spatial drift in spatial working memory might also leave some empirical signature in a spatial language task. Toward this end, we used the ratings task from Hayward and Tarr given its capacity to reveal quantifiable metric effects and its centrality in the spatial language literature (e.g., Hayward & Tarr, 1995; Crawford et al., 2000; Logan & Sadler, 1996; Regier & Carlson, 2001).

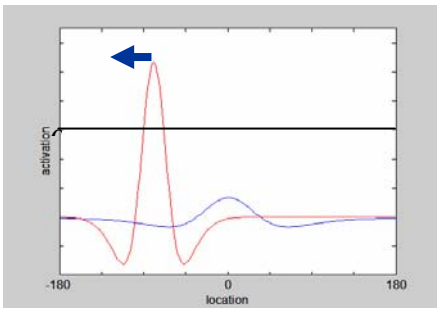


Figure 2 Local excitation/ lateral inhibition

To relate the DFT to performance in the ratings task, we borrowed an idea from Regier and Carlson's (2001) Attentional Vector Sum model and scaled verbal ratings for "above" by the angle between the representation of the target location and the representation of the reference axis, that is, by the spatial distance between the center of the activation peak in SWM and the midline axis (0º). Ratings should be highest when activation is centered at 0º and should fall off systematically as the activation peak is shifted to the left or right. Based on this proposal and the dynamic properties of the DFT, we predicted that if spatial language and spatial memory use the same representational system—spatial working memory—then ratings performance should show delay-dependent "drift", giving systematically lower "above" ratings as memory delays increase (i.e. as the distance between the activation peak in SWM and the midline axis increases).

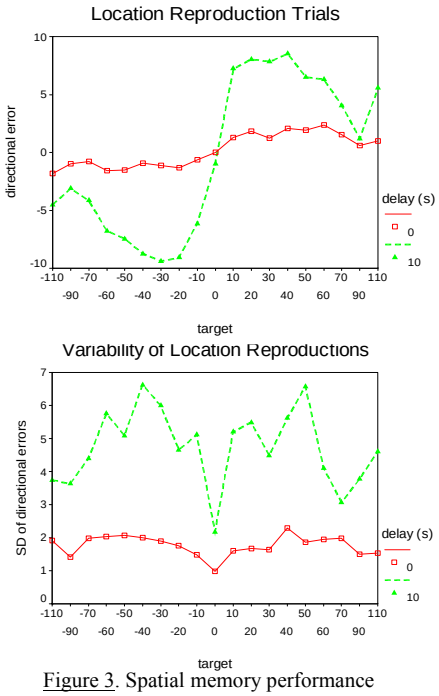


Figure 3. Spatial memory performance

Experimental Support

Subjects. 15 University of Iowa undergraduates participated in this study in exchange for class credit or payment.

Method. Experimental sessions were conducted in dim lighting in a room with black curtains covering all external landmarks. A curved border occluded the corners of the table (and therefore the diagonal symmetry axes).

A single referent disc appeared at midline 30cm in front of the participant and remained visible throughout each presentation trial. At the start of each trial the participant moved a computer mouse to this disk. A number (100-500) then appeared and participants begin counting backwards by 1s aloud until they made a response. This counting task prevented the verbal encoding and maintenance of the spaceship position or rating. A small, spaceship-shaped target then appeared on the screen for two seconds.

Trial Types For spatial memory trials, participants were instructed to move the mouse to the location corresponding the ship's location when the computer says "Ready-Set-Go". For spatial language rating trials, on the other hand, participants are instructed to rate on a scale of 1 ("definitely not above") to 9 ("definitely above") the extent to which the word "above" describes the spaceship's location relative to the reference disk and say their rating when the computer says "Please give your 'Above' rating." The spoken stimuli that indicated which response to provide were each 1500ms in duration. In No Delay conditions, completion of the spoken stimulus was timed to coincide with the offset of the spaceship target. In the 10s Delay conditions, completion of the spoken stimulus occurred exactly 10 seconds after the disappearance of the spaceship target. Spaceship targets appeared at a constant radius of 15cm at 19 different locations relative to the midline axis (0º): every 10º from -70º to +70º as well as ±90º and ±110º to map onto previous research.

Results

Participants in our modified spaceship task either gave a spatial memory response or a verbal ratings response (1 = "definitely not above", 9 = "definitely above") after a 0 s or 10 s delay. The top portion of Figure 3 shows directional errors on the memory trials across target locations and delays. Positive errors were clockwise, while negative errors were counterclockwise. Consistent with previous work (Spencer & Hund, 2002), directional error was larger in the 10 s delay condition and responses were systematically biased away from midline (responses to negative or leftward targets showed counterclockwise bias; responses to positive or rightward targets showed clockwise bias). Similar effects were found for variable error (see lower portion of Figure 3). Specifically, variability was higher in the 10s delay condition, and responses to targets to the left and right of midline were more variable than responses to the target aligned with 0º.

Critically, we also found the predicted delay-dependent drift effect in participants' ratings performance. The top portion of Figure 4 shows that "above" ratings in the spaceship task followed a gradient similar to that obtained by Hayward and Tarr (1995) and Crawford et al. (2000). However, there was a systematic and significant decrease in ratings in the 10 s delay condition. Examination of ratings variability revealed effects of delay comparable to those found on the spatial memory trials (see Figure 4). Specifically, ratings variability was higher at the long delay and lower for targets near the midline axis.

In a final analysis, we compared spatial memory and ratings responses directly by converting the ratings "drift" apparent in Figure 4 into a spatial deviation measure (e.g., deviation at 10º target = (change in 10 s delay rating between 10º and 20º) * 10º / (change in 0 s delay rating between 10º and 20º)). This analysis revealed a high degree of overlap in delay-dependent spatial drift across the two tasks (see Figure 5). These results support the prediction we generated from the DFT and suggest that a shared working memory representation underlies performance in both tasks.

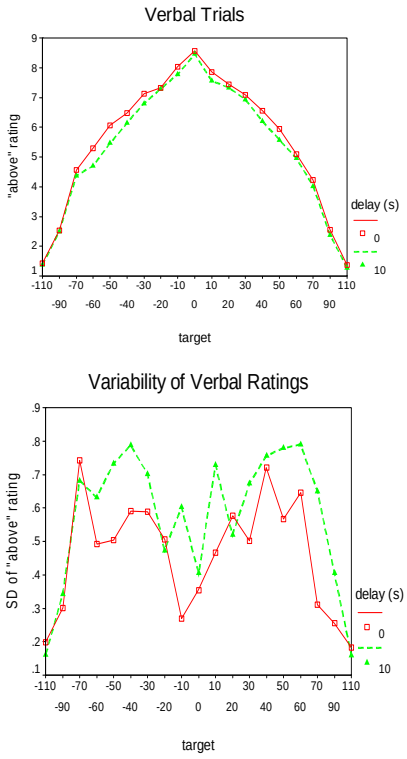


Figure 4 Ratings performance

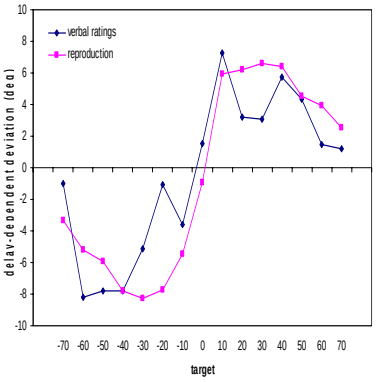


Figure 5 Ratings and drift

Towards a Dynamic Field Model of Spatial Language Performance

Given that we have a formal theory of SWM and encouraging preliminary data, we are in a unique position to develop a process model of both spatial memory and verbal behavior in spatial tasks. The starting point of such a model will be a modified dynamic field model that links two dynamic fields—the SWM field discussed previously and a new spatial prepositions field. Although this new spatial prepositions field has yet to be formalized, the current data suggest two important features. First, this field must be alignable with particular locations in SWM, in particular with perceived reference frames. We are currently developing a process within the field theory that handles the alignment of multiple fields, including the anchoring and scaling necessary in such situations. Critically, these processes must be developed in a way that allows for

generalizability to spatial prepositions beyond "above" such as "left", "right", and "below".

Second, consistent with the dynamic nature of the tasks employed here, and cognition more generally, the two fields should be dynamically coupled. This means that activation in SWM can serve as input to the spatial preposition field and vice versa. This dynamic coupling is critical given the presented evidence that verbal ratings reflect the same dynamic processes underlying spatial working memory performance. If these layers are indeed dynamically coupled as we suggest, then establishment of stable activation peaks within one layer should give rise to stable peaks in the other. Similarly, instability and drift within one layer should give rise to a instability and drift within the other layer. Experiments are currently underway to test these specific predictions.

Although this provides only a limited window onto the dynamic processes that underlie a very flexible spatial cognitive system, we contend that this is an appropriate starting point given the novelty of our general theoretical approach. Indeed, the results of our current experiments will provide the empirical foundation for a more extensive formal model that links spatial working memory with spatial language processes.

References

Anderson, J.R. (2000). *Cognitive psychology and it implications*. New York: Worth.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*,577-660.

Crawford, L.E., Regier, T., & Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization.*Cognition, 75*, 209-235.

Harnad, S. (1990). The symbol grounding problem.*Physica D, 42*, 335-346.

Hayward, W.G., & Tarr, M.J. (1995). Spatial language and spatial representation. *Cognition, 55*, 39-84.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352-376.

Jackendoff, R. (1992). *Languages of the mind*. Cambridge, MA: MIT Press.

Jackendoff, R. (1996).The architecture of the linguistic-spatial interface. In P. Bloom, M.A. Peterson, L. Nadel, & M.F. Garret (Eds) *Language and Space*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002). Foundations of language. New York: Oxford University Press.

Logan, G.D.,& Sadler, D.D. (1996). The architecture of the linguistic-spatial interface. In P. Bloom, M.A. Peterson, L. Nadel, & M.F. Garret (Eds) Language and Space. Cambridge, MA: MIT Press.

Regier, T., & Carlson, L.A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General, 130,*2373-298

Richardson, D.C., Spivey, M.J., Barsalou, L.W., & McRae, K. (in press). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science.*

Schutte, A. R., Spencer, J. P., & Schöner, G. (2003). Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. *Child Development, 74*, 1393-1417.

Spencer, J.P. & Hund, A.M. (2002). Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General, 131,* 16-37.

Spencer, J.P. & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science, 6*, 392-412.

Stanfield, R.A., & Zwaan, R.A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*, 153-156.

Talmy, L. (1983). How language structures space. In H. Pick& L. Acredelo (Eds): Spatial orientation: Theory, research, and application. Plenum. New York.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.

Simplicity in Explanation

Tania Lombrozo (lombrozo@wjh.harvard.edu)
Department of Psychology, Harvard University
33 Kirkland St., Cambridge, MA 02144

J. Jane Rutstein (jessica.rutstein@tufts.edu)
Department of Philosophy, Tufts University
Medford, MA 02155

Abstract

In this paper we explore the role of simplicity in choosing between competing explanations, and in particular how a preference for simplicity is integrated with information about the probability of particular explanations. In Experiment 1 we establish that all else being equal, people prefer explanations that are simpler in the sense of invoking fewer causes. Experiment 2 finds that people require disproportionate evidence in favor of a complex explanation before they will choose it over a simpler alternative. Experiment 3 suggests that this bias is not driven by assumptions about the probabilistic dependence of causes. Finally, Experiment 4 replicates the basic findings with a more ecologically valid computer task. We also find that participants who prefer a simpler but less probable explanation overestimate the frequency of events that would make the simpler explanation more probable. We conclude by suggesting that people believe simpler explanations are more likely to be true in virtue of being simple.

Introduction

Explaining the world around us is a fundamental part of everyday life. We wonder why objects have the properties they do, why people act in particular ways, and why things do or don't happen. But more often than not, explanations are vastly underconstrained by our knowledge and the available data. When more than one explanation is possible, how do we choose between them? A plausible constraint on competing explanations, often attributed to William of Occam, is simplicity. Here we explore whether people in fact prefer simpler explanations, and if so how they balance a preference for simplicity with the desire to maximize other virtues of explanation, like their probability of being true. We show that people do prefer simpler explanations, even when they are less probable than more complex alternatives. We also show that this preference can lead to systematic distortions in the perceived frequency of events.

A Metric for Simplicity

While simplicity is commonly invoked, it is notoriously difficult to formalize and justify. Several recently proposed approaches, like the Akaike Information Criterion (AIC) (e.g. Sober, forthcoming), Bayesian Occam's razor (e.g. Jeffreys & Berger, 1992), Minimum Description Length (MDL) and Kolmogorov Complexity (e.g. Chater & Vitanyi, 2003), nonetheless succeed in precisely specifying a metric for simplicity in the language of statistics and computer science. What's more, these metrics can be motivated on principled grounds. The AIC warrants a preference for simplicity by showing that simpler explanations are more likely to generalize. Similarly, Bayesian Occam's Razor shows that a simpler explanation will have a higher posterior probability. From a bottom-up perspective, considerations of processing constraints make measures like MDL and Kolmogorov complexity attractive.

While compelling, formal measures of simplicity are generally formulated over well-defined problems like line fitting, which bear little resemblance to the complex inductive leaps that characterize everyday explanatory judgments. For this reason we looked to the history of science for a more psychologically plausible metric. In the *Principia*, Newton wrote that "we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances" (1686). This maxim, similar to Occam's statement that entities should not be multiplied beyond necessity, suggests that explanations invoking fewer causes are to be preferred. We thus chose to quantify simplicity in terms of number of causes, where explanations involving fewer causes are simpler.

Simplicity and Probability

Newton's endorsement of explanations with fewer causes was likely grounded in metaphysical assumptions. After the quote above, he went on to suggest that "nature is pleased with simplicity and affects not the pomp of superfluous causes." If nature is in fact simple, then simple explanations are more likely to be true.

In the first experiments reported below we explore whether people prefer explanations involving fewer causes, but also if this preference is motivated by the belief that simpler explanations are more likely to be true. We examine the relationship between simplicity and probability both directly and indirectly. As a direct test, we look at people's justifications for choosing a simpler explanation. More indirectly, we look at whether people switch their preference from a simpler to a more complex explanation when provided with evidence that the more complex explanation is more likely.

Seeing how people balance the competing explanatory virtues of simplicity and probability can help distinguish

two possible hypotheses about the nature of a preference for simplicity. According to what we call the *probabilistic metric* hypothesis, people prefer simpler explanations, but represent this preference in terms of probability. That is, simpler explanations are believed to be more likely to be true in virtue of being simple. While choosing between competing explanations involves deciding which is most likely to be true, simpler explanations gain a probabilistic boost just for being simple. A second hypothesis is the *trade-off* hypothesis, according to which simplicity and probability are independent virtues of explanation that must be integrated according to some weighting function. The *probabilistic metric* hypothesis differs from the *trade-off* hypothesis in that the former claims the preference for simpler explanation is expressed in terms of probability, whereas the latter assumes that simplicity and probability trade-off in a way that is not commensurate.

In the final experiment we go on to explore the consequences of a tendency to favor simpler explanations. Specifically, does the preference for simpler explanations distort our perception of probability? If so, we would expect people's preferred explanations to influence their frequency judgments.

Experiments

All experiments we report involve a simple task adapted from Lagnado (1994) in which participants are asked to choose between one and two diseases to account for some symptoms. By varying the prevalence of the diseases we were able to manipulate the relative probability of the simpler, one-disease explanation to the more complex, two-disease explanation.

Experiment 1: Explanatory Virtues

Before examining how people integrate information about simplicity and probability in explanation, we wanted to confirm that simplicity and probability are indeed virtues of explanation.

Methods Twenty-four Boston-area undergraduate and summer school students completed a questionnaire in one of two conditions: the *simplicity* condition and the *probability* condition.

In the *simplicity* condition, participants read the following:

> There is a population of 750 aliens that lives on planet Zorg. You are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's **minttels are sore** and Treda has developed **purple spots**.
>
> **Tritchet's syndrome** always causes both **sore minttels** and **purple spots**.
>
> **Morad's disease** always causes **sore minttels**, but the disease never causes **purple spots**.
>
> When an alien has a **Humel infection**, that alien will always develop **purple spots**, but the infection will never cause **sore minttels**.
>
> Nothing else is known to cause an alien's **minttels to be sore** or the development of **purple spots**.

They were then asked to choose the *most satisfying* explanation for Treda's symptoms among a list of possibilities that included every disease individually and every pairwise combination of diseases. Choosing Tritchet's syndrome would be the simplest option; choosing Morad's and a Humel infection is a more complex alternative.

In the *probability* condition, the cover story was similar, but participants were asked to choose between two diseases that each accounted for both symptoms. However, one disease was said to be present in about 50 of the aliens on Zorg, while the other was present in about 73 aliens on Zorg, making the latter choice the more probable option.

After choosing an explanation, participants were also asked to explain their reasoning. The names of the diseases were counterbalanced and we used three different sets of symptoms.

Results and Conclusions In the *simplicity* condition, 100% of participants chose the simpler explanation. They justified this choice about equally often by appeal to simplicity and probability: 50% explicitly said they chose it because it was simpler, while 42% said they thought it was more likely for the alien to have one disease than two. In the *probability* condition, 92% of participants chose the more probable explanation. All participants justified this choice by appeal to probability.

Experiment 2: Simplicity Versus Probability

Having established that people do prefer both simpler and more probable explanations, we went on to see how these virtues of explanations are traded off. To do so we had participants choose explanations in cases where the simplest was not the most likely to be true.

Methods One-hundred-thirty-seven Boston-area summer school and undergraduate students participated by completing a questionnaire. The questionnaire was like the *simplicity* condition from Experiment 1, but participants were additionally given information about the prevalence of each disease in the population. For example, one questionnaire read:

> There is a population of 750 aliens that lives on planet Zorg. You are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's **minttels are sore** and Treda has developed **purple spots**.
>
> **Tritchet's syndrome** always causes both **sore minttels** and **purple spots**. **Tritchet's syndrome** is present in about 50 aliens on Zorg.
>
> **Morad's disease** always causes **sore minttels**, but the disease never causes **purple spots**. **Morad's disease** is present in about 225 of the aliens on Zorg.
>
> When an alien has a **Humel infection**, that alien will always develop **purple spots**, but the infection will never cause **sore minttels**. You know that Humel Infections are present in about 210 of the aliens on Zorg.

Nothing else is known to cause an alien's **minttels to be sore** or the development of **purple spots**.

As in Experiment 1, they were then asked to choose the most satisfying explanation and selected among six options, which included each disease individually and every pairwise combination. On a second page of the questionnaire they were asked to justify their choice, and also to complete a math problem. The math problem required participants to compute the joint probability of winning at two slot machines and compare this to the probability of winning at a different machine. We included this problem to see whether participants knew how to compute joint probabilities.

We varied the prevalence of the diseases to manipulate the relative probability of having the single disease causing both symptoms (D_1) to having both of the other diseases (D_2&D_3). Table 1 indicates the 8 sets of values we used, along with the corresponding probability ratios, which were computed on the assumption that the diseases are probabilistically independent. There were 14 to 18 participants per condition.

Table 1: Disease prevalence for each frequency condition.

D_1	D_2	D_3	$P(D_1):P(D_2\&D_3)$
50	50	50	15:1
50	197	190	1:1
50	195	214	9:10
50	225	210	4:5
50	250	220	2:3
50	268	280	1:2
50	330	340	1:3
50	610	620	1:10

Explanation justifications were coded into one of three categories: simplicity, probability, and other. Justifications were coded as 'simplicity' if (1) the participant explicitly mentioned simplicity, or (2) the justification emphasized that the *single* disease accounted for *both* symptoms, thus suggesting that it was unnecessary to invoke two diseases when one would do the trick. Justifications were categorized as 'probability' if the participant claimed their choice was more probable or seemed more likely to be true. Both participants who computed the joint probability of D_2&D_3 and those who went on a subjective feeling of probability were included in this category. Finally, participants whose justifications could not be classified as simplicity or probability were included in the 'other' category. Often these justifications included a restatement of the question ("it seemed best" or "it seemed most satisfying") or an appeal to general intuition ("I went with my gut feeling").

As before, the disease names were counterbalanced, and the explanation choices were presented in random order. In addition, we counterbalanced the order of the presentation of the diseases such that half the participants read about D_1 first and half read about D_1 last. We used three different sets of symptoms.

Results and Conclusions Figure 1 indicates the percentage of participants choosing the simpler explanation in each frequency condition. Nearly all participants chose the simpler explanation when the probability ratio of D_1 to D_2&D_3 was close, but this number steadily declined as it became increasingly probable that an alien had D_2&D_3. Even when it was ten times more likely for the alien to have D_2&D_3, however, over a third of participants were still choosing the simpler explanation. Nor was this preference for the simpler explanation due to participants' inability to compute joint probabilities. The correlation between explanation choice and answering the math problem correctly was small and not significantly different from zero ($r = .12$, $p > .15$).

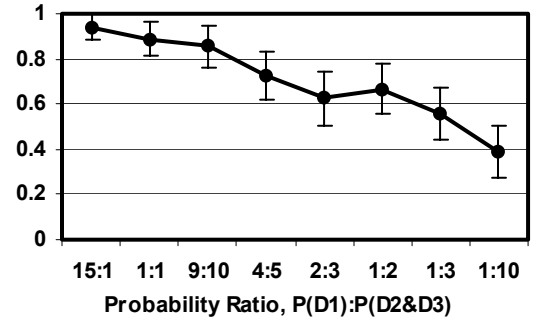


Figure 1: Percent of participants choosing simpler explanation as a function of the probability condition.

To better understand the data we conducted a logistic regression analysis. We used the natural log of the probability ratio as the predictor for the percentage of participants choosing the simpler explanation, as this choice results in a straightforward interpretation of the regression parameters. To understand why, it helps to consider how these parameters relate to the computations that would be performed by an idealized Bayesian agent. In our task, the ideal agent's data would result in a slope parameter of 1 and a constant of 0. A non-ideal agent could have a bias in favor of simplicity at either of two stages in the inference process, each corresponding to a parameter of the logistic function. A slope significantly less than 1 would suggest that the agent underweights the importance of probability: as evidence in favor of D_2&D_3 accumulates, the agent fails to reduce the probability of choosing D_1 accordingly. In contrast, a constant significantly different from zero reflects a bias at the level of the prior probability. The non-ideal agent could overweight, underweight, or appropriately weight probability information, but starts out with disproportionate confidence that D_1 is true.

The *probabilistic metric* and *trade-off hypotheses* make different predictions about the parameters of the logistic function resulting from this analysis. Specifically, the *probabilistic metric* hypothesis requires that the slope parameter be 1. If the preference for simplicity is represented in terms of probability, then probability information should be weighted appropriately. The constant, however, could be significantly different from 0. In contrast, the *trade-off* hypothesis makes no predictions about these parameters. Because simplicity and probability are

evaluated on different metrics, the bias could be reflected in either or both parameters.

The regression analysis resulted in a constant significantly different from zero, but a slope not significantly different from one. This provides some support for the *probabilistic metric* hypothesis. Specifically, the data suggest that as a group, participants think the simpler explanation is more likely than the complex alternative by a factor of about 4 (1.4 to 9, .95 confidence interval), and this belief influences what would be the prior probability in a Bayesian computation. When the probability ratio is 1:2, the percentage of subjects choosing the simpler explanation corresponds to the ideal Bayesian's posterior probability for D_1 at a frequency of 1:(2/4), and so on for the other values. As a result, participants require disproportionate evidence in favor of the complex explanation before it can rival the simpler alternative. Nonetheless, the slope of the regression suggests that participants incorporate probability information appropriately in making a decision.

We can also examine participants' beliefs about simplicity by looking at how they justified their explanation choices. When the simpler explanation was also more probable, a majority of participants justified choosing the simpler explanation by appeal to probability. However, 'simplicity' and 'other' explanations became increasingly common as the simpler explanation became less probable. These trends are illustrated in Figure 2, which indicates the percent of each justification type for the simpler explanation. Because there were few participants in some categories, the figure combines data from pairs of probability ratios.

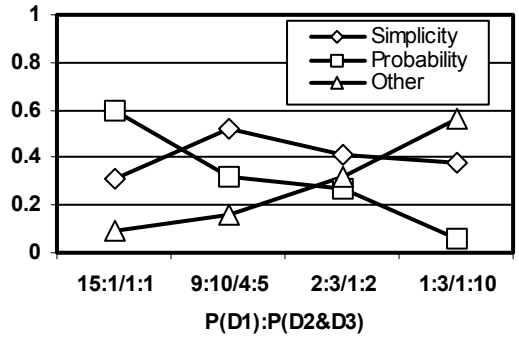


Figure 2: Distribution of justifications for choosing the simpler explanation.

Overall, the data suggest that many participants thought simpler explanations were more likely to be true. This is most apparent from the logistic regression analysis, but is also supported by the patterns of explanation justifications. In particular, many participants justified the choice of a simpler explanation by appeal to probability in conditions where the more complex explanation was as much as two times more likely. The data from the math problem suggest that this preference does not result from an inability to compute joint probabilities, but Experiment 3 considers and eliminates another alternative explanation.

Experiment 3: Independence Assumptions

In Experiment 2 we found that participants chose the simpler explanation well beyond the point at which a more complex alternative was more probable. However, the probability values against which we compared participants' choices were calculated on the assumption that diseases D_2 and D_3 are probabilistically independent—that is, that $P(D_2|D_3) = P(D_2)$ and $P(D_3|D_2) = P(D_3)$. As participants were told nothing about the dependence of the diseases, it's possible that they made a different assumption. In particular, if $P(D_2|D_3)$ is much smaller than $P(D_2)$, participants would be warranted in choosing D_1 on probabilistic grounds.

In Experiment 3 we were interested in determining participants' beliefs about the dependence of diseases. We also wanted to assess whether such beliefs influence explanatory preferences. To do so we found a domain involving dependence assumptions distinct from those for diseases, and examined whether more participants chose the complex explanation for this domain.

Methods Sixty-eight Boston-area undergraduate and summer school students participated. Twenty were in the *assumptions* condition, where we explicitly asked participants to provide a judgment of probabilistic dependence as follows:

> Suppose there are two diseases with similar symptoms, D_1 and D_2. Do you think someone who has D_1 is more or less likely to have D_2 than someone who does not have D_1?
> *Circle one*: **More** **Less**

In addition to asking about the dependence of diseases, we also wanted to find items with a different dependence assumption. We thus queried participants about books:

> Suppose there are two books on similar topics, B_1 and B_2. Do you think someone who has read B_1 is more or less likely to have read B_2 than someone who has not read B_1?
> *Circle one*: **More** **Less**

Participants in the *assumptions* condition saw both the disease and book questions, with the order counterbalanced.

The remaining 48 participants performed a task like Experiment 2 at the 2:3 probability ratio. However, half were asked about diseases, while the remaining half saw a formally identical question about books. Instead of reasoning about diseases causing symptoms, they were asked about books 'causing' knowledge of facts. For example, a passage read: "***The Zorgian Guide to Interplanetary Living*** contains the fact that **Planet Earth has an atmosphere** and the fact that **humans have two legs**. You know that about 50 aliens on Zorg have read ***The Zorgian Guide to Interplanetary Living***."

Results and Conclusions We first analyzed the data from the *assumptions* condition. Most participants (80%) claimed that having a disease makes someone *less* likely to have a similar disease, but 80% thought that having read a book makes someone *more* likely to have read a similar book.

These values were significantly different from chance, as well as being significantly different from each other ($\chi^2(1) = 14.4$, $p < .01$). Having thus established that people have different dependence assumptions about diseases and books, we went on to look at whether these assumptions affect explanatory preferences.

Replicating Experiment 2, we found that about half (46%) of participants chose the simpler explanation at the 2:3 probability ratio in the disease condition. In the book condition the results were identical, with 11 of 24 participants (46%) choosing the simpler explanation. There were also no differences between conditions in how participants justified their choice. The absence of a difference between the disease and book conditions suggests that beliefs about probabilistic dependence do not account for participants' preference for simpler explanations.

Experiment 4: Computer Replication

In the previous experiments participants were informed of the prevalence of each disease by being presented with a frequency. This method has two limitations. First, in the real world most frequency information is acquired through experience rather than a summary value, making the ecological validity of the task questionable. Second, having actual numbers allowed some participants to compute the joint probability of the diseases rather than relying on subjective judgments. For these reasons we decided to replicate the basic task in a computer format. Doing so also allowed us to examine whether explanatory preferences have consequences for perceived frequencies.

Methods One-hundred-and-eight Boston-area summer school and undergraduate students participated. The task was like Experiment 2, but on the computer. Instead of being told the prevalence of the diseases, for each disease participants saw ten screens containing a total of 75 aliens, some of which were marked as having a particular disease. In this way equivalent frequency information was communicated.

Participants were in one of four frequency conditions corresponding to probability ratios of 15:1, 9:10, 1:2 and 1:10, with 27 participants per condition. After being presented with the cover story and frequency information, participants were asked to choose the most satisfying explanation for the alien's symptoms and, as before, selected an answer among six options, which included every disease alone and each pairwise combination. They were then asked to explain their choice and to estimate the frequency of each disease in the Zorg population.

Counterbalancing and randomization was as in Experiment 2, with the additional control that the order of presentation of the disease frequencies was varied according to a Latin square.

Results and Conclusions The overall explanatory preferences in the computer task replicated those of Experiment 2, suggesting that the questionnaire format was methodologically sound (see Figure 3). Virtually all participants chose the simpler explanation when it was more likely, but nearly half continued to prefer the simpler explanation when it was as much as ten times more likely that the alien had two diseases.

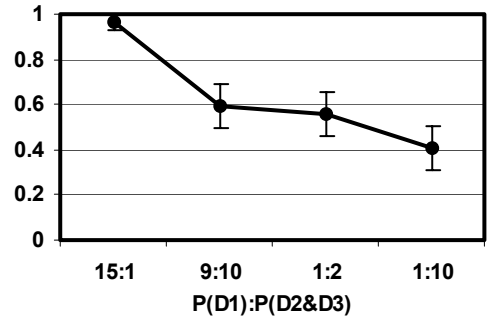


Figure 3: Percent of subjects choosing simpler explanation in computer task.

We also analyzed justifications for choosing the simpler explanation, using the coding scheme from Experiment 2. Not surprisingly, as the simpler explanation became less probable, a larger proportion of participants invoked simplicity rather than probability in their justifications (see Figure 4).

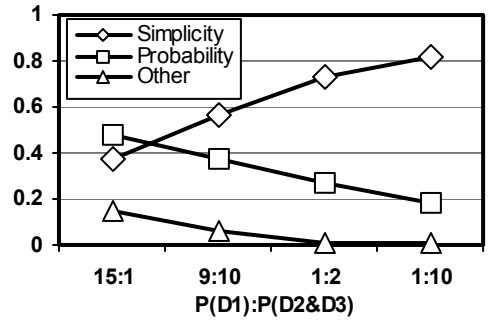


Figure 4: Distribution of justifications for choosing the simpler explanation in the computer task.

Using a computer task allowed us to examine an aspect of simplicity we couldn't address in the questionnaire format, namely how explanatory choices affect perceived frequencies. Figure 5 presents participants' estimates of the percentage of the Zorg population with each of D_1, D_2, and D_3. The average estimates are shown as a function of both frequency condition and explanation choice, with participants who chose the simpler, one-cause explanation distinguished from those who chose the more complex, two-cause explanation. Solid lines indicate the actual percentage of aliens with each disease.

While subjects were remarkably accurate overall, the data for D_1 suggest that those participants who chose the simple, one-cause explanation when it was less probable systematically overestimated the frequency of D_1. In both the 1:2 and 1:10 frequency conditions, the average estimate of participants who chose the simpler explanation were

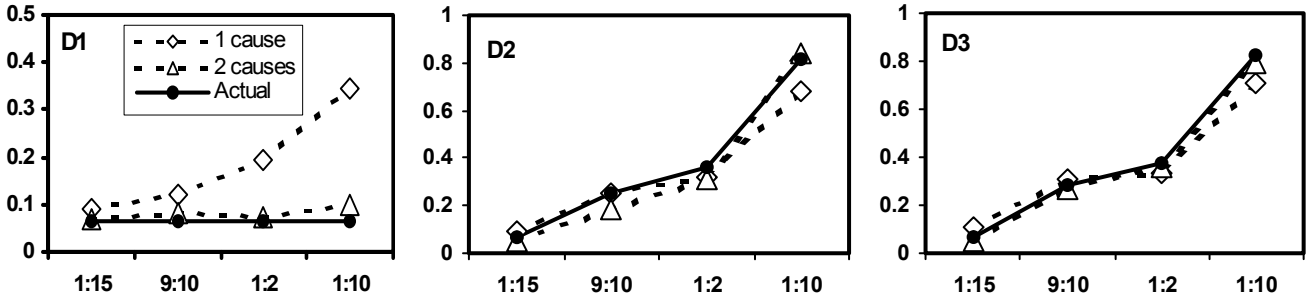


Figure 5: Average frequency estimates for each disease as a function of probability ratio and explanation choice.

significantly higher than that of participants who chose the more complex alternative ($p < .05$). One possibility, however, is that some subjects confused the frequency of D_1 with either D_2 or D_3, which would result in inflated D_1 estimates. If this were true we would expect to see systematic underestimation of D_2 and D_3. Moreover, only one subject provided a higher estimate for D_1 than either D_2 or D_3. This suggests that the overestimation of D_1 is not due to mismatching the frequencies and their corresponding diseases.

Another explanation for the D_1 overestimation is that participants who chose the simpler explanation were bad at estimating frequencies, and for this reason based their explanation choice on simplicity. This possibility is ruled out by the frequency estimation data for D_2 and D_3, where there were no differences between the estimates of participants who chose one or two cause explanations.

These data suggest that participants who chose the simpler explanation systematically overestimated the frequency of D_1 *as a result* of their explanation choice. However, it could be that some participants overestimated D_1, which in turn lead them both to choose the simpler explanation and to indicate a high prevalence of D_1. Evidence that the former interpretation is the correct one comes from the fact that participants never systematically overestimate D_1 in the 1:15 condition, when simplicity and probability converged on the same explanation.

Conclusions

We began by considering whether people prefer simpler explanations, and whether this preference is supported by a belief that simpler explanations are more likely to be true. We found overwhelming evidence for the claim that people do prefer simpler explanations, at least where simplicity is understood in terms of number of causes. Participants consistently chose a simpler explanation when provided no information about probability, and required a disproportionate amount of probability information in order to override this preference.

These findings are consistent with the idea that people believe simpler explanations are more likely to be true, albeit implicitly. Many subjects explicitly justified their choice of a simpler explanation by appeal to probability, but more telling is the fact that participants evaluated simplicity and probability as if they were commensurable quantities. The results from Experiment 2 tentatively support the *probabilistic metric* hypothesis over the *trade-off* hypothesis: people do prefer simpler explanations, but this bias manifests as a reweighing of priors rather than a failure to appropriately incorporate probability information.

The intimate relationship between simplicity and probability is most dramatically illustrated by the finding that committing to an improbable, simple explanation results in the systematic distortion of perceived frequencies. This result indicates that explanatory choices can have consequences for probabilistic judgments, and suggests that the study of explanation can provide a unique window into the mechanisms by which beliefs about the world influence decisions.

Acknowledgments

This work was supported by an NDSEG Fellowship awarded to the first author. We would like to thank Susan Carey and Tom Griffiths for helpful discussion, Liz Baraff for paper comments, and Stephanie Samuels and Greg Westin for help with data collection.

References

Chater, N. & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Science*, *7(1)*, 19-22.

Jeffreys, W. H. & Berger, J.O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64-72.

Lagnado, D. (1994). *The Psychology of Explanation: A Bayesian Approach*. Masters Thesis. Schools of Psychology and Computer Science, University of Birmingham.

Newton I. (1953/1686). *Philosophiae Naturalis Principia Mathematica*. Reprinted in H. Thayer (Ed.) *Newton's Philosophy of Nature*. New York: Hafner.

Sober, E. (Forthcoming). Parsimony. In S. Sarkar (Ed.), *The Philosophy of Science—an Encyclopedia.* New York: Routledge.

Variation in Language and Cohesion across Written and Spoken Registers

Max M. Louwerse (mlouwers@memphis.edu)
Department of Psychology / Institute for Intelligent Systems, 202 Psychology Building
Memphis. TN 38152

Philip M. McCarthy (pmmccrth@memphis.edu)
Department of English, Patterson 467
Memphis. TN 38152

Danielle S. McNamara (dsmcnamr@memphis.edu)
Department of Psychology, 202 Psychology Building
Memphis. TN 38152

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building
Memphis. TN 38152

Abstract

This paper investigates the variation in cohesion across written and spoken registers. The same method and corpora were used as in Biber's (1988) study on linguistic variation across speech and writing; however instead of focusing on 67 linguistic features that primarily operate at the word level, we compared 236 language and cohesion features at the text-level. Variations in frequencies across these features provided evidence for six dimensions: (1) speech versus writing, (2) informational versus declarative, (3) factual versus situational, (4) topic consistency versus topic variation, (5) elaborative versus constrained, (6) narrative versus non-narrative. Our cohesion and linguistic analysis showed most variation in speech and writing, whereas the linguistic feature analysis operating at the word level did not yield any difference.

Introduction

One way to investigate similarities and differences between speech and writing is by using corpus linguistic methods. The most common and largest investigation of this kind is Biber (1988). Biber used 23 spoken and written registers. These registers are language varieties mediated by social situations and are similar to genres. Biber took these registers from the Lancaster-Oslo-Bergen (LOB) corpus and the London-Lund corpus, and computed the frequency of 67 linguistic features in these registers (see Table 1 for an overview of registers).

The linguistic features used for Biber's analysis primarily operate at the word level (e.g., parts-of-speech) and can be categorized as (1) tense and aspect markers, (2) place and time adverbials, (3) pronouns and pro-verbs, (4) questions, (5) nominal forms, (6) passives, (7) stative forms, (8) subordination features, (9) prepositional phrases, adjectives and adverbs, (10) lexical specificity, (11) lexical classes, (12) modals, (13) specialized verb classes, (14) reduced forms and dispreferred structures, and (15) coordinations and negations.

Table 1. The 23 registers used in Biber (1988)

Corpus	Register
Lancaster-Oslo-Bergen corpus	Press reportage, editorials, press reviews, religion, skills and hobbies, popular lore, biographies, official documents, academic prose, general fiction, mystery fiction, science fiction, adventure fiction, romantic fiction, humor
London-Lund corpus	Face-to-face conversation, telephone conversation, public conversations, debates, and interviews, broadcast, spontaneous speeches, planned speeches
(Additional)	Personal letters, professional letters

In Biber's study the normalized frequencies of these features in each of the registers were then entered in a factor analysis, from which six factors emerged. These factors can be seen as dimensions on which registers can be placed. Biber's analysis showed that no single dimension comprised a difference between speech and writing; As such, Biber defined the sets of relations among texts as follows:

1. Involved versus informational production
2. Narrative versus non-narrative concerns
3. Explicit versus situation dependent reference
4. Overt expression of persuasion
5. Abstract versus non-abstract information
6. On-line informational elaboration.

For example, registers such as romantic fiction, mystery fiction and science fiction were positioned high on the second dimension (narrative); whereas registers such as academic prose, official documents, hobbies, and broadcasts scored low (non-narrative).

Biber's (1988) study and the multi-feature, multidimensional approach have become a standard in corpus linguistics (McEnery, 2003), leading to various extensions (Biber, Conrad & Reppen, 1998; Conrad &

Biber, 2001), as well as to assessments of the validity, stability, and meaningfulness of the approach and its findings (Lee, 2004).

Measuring cohesion

Texts obviously consist of a large variety of linguistic features, many of which can be identified at a word level (e.g. morpho-semantics, syntactic category, frequency). Biber's study has shown that these linguistic features are powerful determiners of similarities and differences between registers. But despite these impressive results, the theoretical question that remains lurking is to what extent these linguistic features fully capture the nature of a text and thereby the nature of a register.

Although linguistic features operating at the word level may identify several register characteristics, we also know that one of the key features of a text is that it is not just a concatenation of words and sentences. Instead, there is a structure in the text that glues the various text components together. In comprehending the text, the reader or listener constructs a coherent, mental representation of the situations which have been cohesively described by the text. We have used the term "coherence" for the representational relationships and "cohesion" for the textual indications through which coherent representations should be built (Louwerse & Graesser, 2004). Cohesion, it should be noted, cannot be captured only by linguistic features at the word level. Instead, cohesion stretches to the inter-clause, inter-sentence and inter-paragraph level.

But if a key component in the nature of text consists of cohesion, a practical issue related to the theoretical question needs to be addressed. Linguistic features that operate at a word level can currently be reliably identified by regular expressions, part-of-speech taggers, and syntactic parsers. However, there is the practical question of whether automated techniques can also capture the cohesion of text. Recent landmark progress in computational linguistics has indeed allowed us to go far beyond surface level components into automating deeper and global levels of text and language analysis (Jurafsky & Martin, 2001). This progress has resulted in the cohesion and coherence measurement tool Coh-Metrix.

Coh-Metrix

Coh-Metrix was initially developed in order to replace readability formulas that exclusively focus on simple and shallow metrics. Instead, Coh-Metrix is sensitive to a broader profile of language and cohesion characteristics. It analyzes texts on 236 types of cohesion relations and measures of language, text, and readability (McNaramara, Louwerse, & Graesser, 2002; Graesser, McNamara, & Louwerse, in press). For this paper, we will only focus on the textual features (cohesion) of the tool.

The modules of Coh-Metrix use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components that are widely used in computational linguistics. For example, the MRC database (Coltheart, 1981) is used for psycholinguistic information; WordNet (Miller, Beckwith, Fellbaum, Gross & Miller) for underlying lexical concepts; Latent Semantic Analysis (Landauer & Dumais) for the semantic similarities between words, sentences and paragraphs; the ApplePie parser (Sekine & Grishman, 1995) and the Brill (1995) part-of-speech tagger for a variety of syntactic categories.

Spatial restrictions do not make it possible to discuss all of the measures Coh-Metrix makes available. As such, a brief summary of key measures will have to suffice, whereas Graesser et al. (in press) has a more complete overview.

1. *Word information* includes word familiarity, word concreteness, word imageability, meaningfulness, and age of acquisition.
2. *Word frequency* includes four corpora-based standards: CELEX from the Dutch Centre for Lexical Information (Baayen, Piepenbrock & Van Rijn, 1993); the Kucera-Francis norms (Francis & Kucera, 1982); Thorndike-Lorge norms (Thorndike & Lorge, 1942) and the Brown norms (Brown, 1984).
3. *Part of speech* categories are adopted from the Penn Treebank (Marcus et al., 1993) and the Brill (1995) POS tagger.
4. *Pronoun density* is computed by taking the ratio of pronouns and nouns.
5. *Logical operators* are the incidence score of logical operators *or*, *and*, *not*, and *if-then* phrases
6. *Interclausal relationships* are the additive, temporal and causal cohesion based on connectives between clauses. These can be positive (extending) and negative (adversative), as outlined in Louwerse (2001).
7. *Type-token ratio* refers to the number of unique words divided by the number of tokens of the words.
8. *Polysemy* and *hypernym*: Polysemy is measured as the number of senses of a word in WordNet; whereas the hypernym count is defined as the number of levels in a conceptual taxonomic hierarchy that is superordinate to a word.
9. *Concept clarity* is a composite of multiple factors that measure ambiguity, vagueness, and abstractness.
10. *Syntactic complexity* refers to the noun-phrase density, the mean number of modifiers per noun phrase, the number of high-level constituents per word and the incidence of word classes that signal logical or analytical difficulty.
11. *Readability* scores are computed according to the Flesch Reading Ease formula and the Flesch-Kincaid Grade Level formula, the standard readability formulas.
12. *Coreference*: Three forms of coreference between sentences are computed, namely noun overlap between sentences, stem overlap, and stem-noun overlap.
13. *Causal cohesion* is interpreted as the ratio of causal particles to causal verbs.
14. *Latent semantic analysis*: LSA is a statistical, corpus based, technique used to represent world knowledge that computes similarity comparisons for terms and documents by taking advantage of word co-occurrences. LSA scores

can be computed for sentence to paragraph, sentence to text, paragraph to paragraph and paragraph to text. These measures can be used for measuring the local and global cohesion of the text (see Kintsch, 2002; Landauer & Dumais, 1997).

The advantage of the use of this wide range of computational linguistic tools is that Coh-Metrix is sensitive to variations in language, discourse, and cohesion. Such an analysis may not only help us to determine text difficulty, but may also help us with determining variations across registers.

Multi-dimensional study on cohesion

In our multi-feature, multi-dimensional approach we carefully followed Biber's study to allow comparison of his findings. We used the same fifteen written registers from the LOB corpus and the same six spoken registers from the London-Lund corpus. Two further non-published registers (professional and personal letters), which Biber had generated himself, were substituted with the *Compilation of Messages and Letters of the Presidents* Richardson (2003/1801) and *The Upton Letters* from Christopher Benson (1905), both downloaded from the Gutenberg text archives.

All textual coding other than alphanumeric characters and punctuation was removed. The 23 spoken and written registers were then processed through Coh-Metrix and the normalized frequencies for each of the 236 cohesion, language, and discourse features were saved. We followed Biber's approach in standardizing all frequencies to a mean of 0.0 and a SD of 1.0 and entered them in a factor analysis using the Promax rotation. Whereas Biber used a principal factor analysis to account for the shared variance, we opted for a principal component analysis to account for all the variances. Loadings with an absolute value of less than .35 were excluded from the analysis (Biber, 1988; Comrey & Lee, 1992). The scree plot of eigenvalues, illustrating the amount of variance accounted for by each factor, showed a clear break after six factors, explaining 88.3% of the total variance.

In order to translate the factor scores to the registers, we followed Biber by adding together the standardized scores of all linguistic features responsible for a factor in a particular text. This provides a measure of register's salience on a particular dimension given the presence of the linguistic features in that register. We deviated from Biber's approach in one important way: Whereas Biber removed the linguistic features from subsequent factors once it was used by a previous factor, we preferred to include these linguistic features in additional factors to account for the interactions between language, discourse, and cohesion features.

Dimensions

Space limitations dictate us to summarize the findings by presenting tables in which we have translated the ratio scale to an ordinal scale, thereby not serving full justice to the actual differences between the 23 registers.

Dimension 1: Speech versus writing. This dimension significantly accounts for 53.5% of the variance, (F (1, 22) = 35.61, $p < .001$, MSE = .721). When looking at the grouping of the registers, it immediately becomes apparent that spoken registers are distinct from written registers. In addition, the registers clearly show *the degree* to which the registers are speech-dependent. For example, fiction includes, or more closely reflects, spoken discourse, whereas this is far less likely to be the case with press reviews or professional letters.

The linguistic features with positive loadings are presented in the first data row of the table, signifying the higher presence in the register. They consist of concreteness, imageability, meaningfulness, polysemy, and frequency in the spoken discourse. Negative loadings relate to ambiguous quantification, pronoun density, argument overlap, and semantic similarity between sentences and paragraphs. Registers with a higher score on this dimension (like public conversations and face-to-face conversations) are characterized by frequent occurrences of concrete, imaginable, and meaningful language, together with higher pronoun density and ambiguous quantification. At the same time, occurrences of argument overlap and semantic similarities between text units are less prevalent. Registers with negative scores, presented in the second data row, have the opposite characteristics.

In Table 2 results are given for the Dimension 1. The first column presents the registers ranked by total scores and the second column presents the linguistic features ranked by factor loadings. Row separators mark the difference between positive and negative factor loadings. The same format is used for the remaining five tables representing the remaining five dimensions.

Table 2: Distribution registers and summary Coh-Metrix Dimension 1 (speech versus writing)

public conversations, face to face conversation, spontaneous speeches, telephone conversations, planned speeches, broadcast, mystery fiction	frequency, concreteness, imageability, meaningfulness, polysemy, Flesch Reading Ease, ambiguous quantification, pronoun density, higher level constituents per word, abstract nouns, hypernym, polysemy
Personal letters, general fiction, romantic fiction, religion, adventure fiction, skills and hobbies, official documents, humor, academic prose, editorials, popular lore, biographies, science fiction, press reportage, press reviews, professional letters	LSA sentence to sentence, ratio of causal particles to causal verbs, LSA paragraph to paragraph, paragraph to text, vague adverbs, type-token ratio for nouns, concreteness, argument overlap, average paragraph length, age of acquisition, average syllables per word, mean number of modifiers per noun-phrase, stem overlap, Flesch Kincaid Grade Level

Dimension 2: Informational versus declarative. The second dimension accounts for 16.3 % of the variance, but without significant differences between the registers (F (1, 22) = .93, p = .56, MSE = .968). This dimension shows many similarities with Biber's Dimension 6, with the majority of the registers positioned similarly along the axis in both studies. Biber tentatively labeled this "on-line informational elaboration marking stance" with registers such as planned speeches and public conversations being informational in focus and conveying the speaker's attitudes and beliefs. We come to a similar conclusion, interpreting the difference as informational and subjective versus declarative and objective. Informational registers are characterized by a higher occurrence of temporal cohesion, imageability, and concreteness, but a low occurrence of causality, whereas the opposite characterizes declarative registers.

Table 3: Distribution registers and summary Coh-Metrix Dimension 2 (informational versus declarative)

mystery fiction, religion, skills and hobbies, romantic fiction, spontaneous speeches, official documents, general fiction, popular lore, telephone conversations, adventure fiction, biographies, face to face conversation, broadcast, humor	Positive temporal connectives, polysemy (adjectives), meaningfulness, LSA paragraph to paragraph, familiarity, LSA sentence to sentence, negative temporal connectives, paragraph length, argument overlap, LSA sentence to paragraph, LSA paragraph to text, ratio of causal particles to causal verbs, LSA paragraph to paragraph, type-token ratio for nouns, LSA paragraph to text, imageability, concreteness, LSA sentence to sentence, LSA sentence to sentence, concreteness
planned speeches, public conversations, academic prose, personal letters, editorials, science fiction, professional letters, press reportage, press reviews	Negative causal connectives, frequency, (verbs), causal particles, average syllables per word, positive causal connectives, age of acquisition

Dimension 3: Factual versus situational. This dimension, explaining 7.7 % of the variance, shows similarities with Biber's Dimension 3: "explicit versus situation-dependent reference." Biber argues that the situation-dependent site of the dimension refers to places and times outside of the text (imaginary and real world), whereas the opposite side of the dimension has registers with elaborated explicit reference. Although we do not find evidence for the time or place reference, we do find a higher frequency of imageability and a lower frequency of clarification and causal connectives, with the opposite trend evident for the registers on the factual side of the dimension (F(1, 22) = 5.88, p < .001, MSE = .871). The labels "factual" and "situational" refer to the presentation, rather than the content. For instance, religion is located high on the factual dimension because this register is generally presented as factual. On the other hand, press reviews, reportages and fiction are presented in a less transparent way, often requiring the reader to imagine a situation.

Table 4: Distribution registers and summary Coh-Metrix Dimension 3 (factual versus situational)

academic prose, official documents, religion, skills and hobbies, popular lore, biographies, spontaneous speeches, personal letters, face to face conversation	Clarification connectives, causal particles, negative causal connectives, noun overlap, ratio of causal particles to causal verbs, vague adjectives, negative additive connectives, positive causal connectives, ambiguous quantification, argument overlap, vague verbs, vague nouns,
telephone conversations, humor, editorials, public conversations, press reviews, press reportage, professional letters, planned speeches, general fiction, broadcast, mystery fiction, romantic fiction, adventure fiction, science fiction	polysemy, imageability, causal verbs, mean hypernym of verbs

Dimension 4: Topic consistency versus topic variation. This dimension explains 4.6% of the variance with significant differences between the registers (F(1, 22) = 3.76, p < .001, MSE = .870). It marks the consistency of topics across and within instances of a particular register. For instance, personal and professional letters often have a similar set of topics that are used, as do biographies and spontaneous speeches. Face-to-face conversations, interviews, public debates, press reportages and editorials on the other hand, have more topics and are less predictable, often switching between different instances. In the registers located high in the topic consistency (e.g. personal letters and professional letters), semantic similarities marking global cohesion and local cohesion are higher, but noun density and type-token ratio are lower than the topic variation registers (e.g., reportages and editorials).

Table 5: Distribution registers and summary Coh-Metrix Dimension 4 (topic consistency versus topic variation)

personal letters, spontaneous speeches, professional letters, biographies, broadcast, academic prose, religion, official documents, skills and hobbies, romantic fiction, mystery fiction	frequency conditionals, frequency negations, causal verbs, positive additive connectives, polysemy, LSA paragraph to paragraph, positive causal connectives, LSA sentence to text, LSA paragraph to paragraph, LSA paragraph to text
telephone conversations, general fiction, press reviews, popular lore, planned speeches, humor, adventure fiction, science fiction, face to face conversation, public conversations, press reportage, editorials	type-token ratio, noun density

Dimension 5: Elaborative versus constrained. This dimension is harder to interpret and explains only 3.7 % of the variance. Differences between registers ($F(1, 22) = 3.55$, $p < .001$, $MSE = .866$) suggest that personal letters and press reviews for instance are more opinion-based and have a closer distance between writer and reader, whereas professional letters and press reportages, are more fact and evidence driven. It is almost as if there is more space in personal letters and press reviews to compare ideas. This conclusion is supported by the factor loadings of the linguistic features, which show a prominent role for additive cohesion, vague adjectives and adverbs, along with a high type-token ratio and an accompanying low semantic similarity in the case of the personal letters and the press reviews. It is as if many ideas are juxtaposed within these registers.

Table 6: Distribution registers and summary Coh-Metrix Dimension 5 (elaborative versus constrained)

personal letters, press reviews, biographies, skills and hobbies, religion, humor, popular lore, academic prose, official documents, editorials, general fiction	type-token ratio, negative additive connectives, vague adjectives, vague verbs, positive additive connectives
mystery fiction, science fiction, romantic fiction, telephone conversations, broadcast, adventure fiction, face to face conversation, press reportage, planned speeches, public conversations, spontaneous speeches, professional letters	LSA paragraph to text, LSA paragraph to paragraph, LSA sentence to text

Dimension 6: Narrative versus non-narrative Although significant differences were found between registers (F (1, 22) = 1.64, p = .037, MSE = .991) only 2.5 % of the variance was accounted for by this dimension. Dimension 6 is virtually identical to Biber's Dimension 2. In registers such as fiction and biographies, a narration of events is prominent, whereas narration is less obvious in press reviews and professional letters. Linguistic features like temporal connectives are primarily responsible for this dimension. Despite the similarities with Biber's dimension, there are also some important differences. For instance, in our findings, science fiction scores low on narrative but face-to-face conversations score high, whereas in Biber's analysis the opposite is the case. The clear similarities between the two studies (e.g., the clustering of the fiction texts) support this interpretation of the dimension.

Table 7: Distribution registers and summary Coh-Metrix Dimension 6 (narrative versus non-narrative)

Romantic fiction, mystery fiction, face to face conversation, general fiction, adventure fiction, biographies, religion, public conversations, telephone conversations, official documents	ambiguous temporal relation, vague nouns, positive connectives, temporal connectives
Editorials, academic prose, press reportage, skills and hobbies, humor, spontaneous speeches, popular lore, personal letters, broadcast, planned speeches, science fiction, professional letters, press reviews	LSA sentence to text, LSA paragraph to text, LSA sentence to sentence

Discussion and conclusion

The present study has investigated the multi-feature, multi-dimensional corpus linguistic approach initially outlined by Biber (1988). We have used the same corpora and the same methods as Biber, but instead of including linguistic features that primarily operate at the word level, we have included a large variety of language, discourse and cohesion features. These features ranged from the word level, to sentence, paragraph and discourse level. Six dimensions emerged from a factor analysis: (1) speech versus writing, (2) informational versus declarative, (3) factual versus situational, (4) topic consistency versus topic variation, (5) elaborative versus constrained, (6) narrative versus non-narrative. Three of these dimensions (Dimension 2, 3 and 6) show strong similarities with the distributions of registers as well as the interpretations of dimensions in Biber's study.

Results showed one crucial difference with Biber's finding. Whereas Biber was not able to find one single dimension that determined the difference between speech and writing, we found a very prominent difference in linguistic features between spoken and written discourse (Dimension 1). The most plausible explanation for this

result is the contrast between Biber's focus on the linguistic features operating at the word level and our study which included a much wider range of language and discourse characteristics that we have called cohesion.

Acknowledgments

The research was supported by the Institute for Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Baayen, R. H., R. Piepenbrock, and H. van Rijn (Eds.) (1993). *The CELEX Lexical Database* (CD-ROM). University of Pennsylvania, Philadelphia (PA): Linguistic Data Consortium.

Benson, A.C. (1905). *The Upton letters*. Retrieved January 2004 from the Project Gutenberg Text Archives.

Biber, D. (1988). Linguistic features: algorithms and functions in Variation across speech and writing. Cambridge: Cambridge University Press.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*, 543-566.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioral Research Methods Instrumentation and Computers*, 16, 502-532.

Comrey, A. L. & Lee, H. B. (1992). A first course in factor analysis. Hillsdale, NJ: Lawrence Erlbaum.

Conrad, S. & Biber, D. (2001). *Variation in English: Multi-Dimensional Studies*. Harlow: Longman

Francis, W.N., & Kucera, N. (1982). *Frequency analysis of English usage*. Houghton-Mifflin.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. (in press). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*.

International Computer Archive of Modern and Medieval English (2000). *Lancaster/Oslo/Bergen Corpus of British English* (CD-ROM).

International Computer Archive of Modern and Medieval English (2000). *The London-Lund Corpus of Spoken English* (CD-ROM).

Jurafsky, D., & Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice-Hall.

Kintsch, W. (2002) On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.) *Thematics: Interdisciplinary Studies*. Amsterdam: Benjamins.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Lee, D. Y. W. (2004). *Modeling variation in spoken and written English*. London/New York: Routledge.

Louwerse, M.M. (2002). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 291–315.

Louwerse, M.M. & Graesser, A.C. (2004). Coherence in discourse. In Strazny, P. (ed.), *Encyclopedia of linguistics*. Chicago: Fitzroy Dearborn.

McEnery, T. (2003). Corpus linguistics. In: R. Mitkov (Ed.), *The Oxford encyclopedia of computational linguistics*. Oxford: Oxford University Press.

McNamara, D.S., Louwerse, M.M. & Graesser, A.C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

Miller, G. A., Beckwith, R., Fellbaum, C. , Gross, D. & Miller, K. (1990). *Five Papers on WordNet. Special Issue of the International Journal of Lexicography, 3*.

Richardson, J.D. (2003/1801). *Compilation of the Messages and Papers of the Presidents* (Vol. 1, John Adams). Retrieved January 2004 from the Project Gutenberg Text Archives.

Sekine, S., & Grishman, R. (1995). A corpus-based probabilistic grammar with only two nonterminals. *Fourth International Workshop on Parsing Technology*.

Thorndike, E.L. and Lorge, I. (1944). The *teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

The 2004 CogSci proceedings publication

Lozano, S. C. and Tversky, B. Communicative Gestures Benefit Communicators. Pp. 849-854

has been retracted.

Both authors retract this article. Barbara Tversky believes that the research results cannot be relied upon; Sandra C. Lozano takes full responsibility for the need to retract this article.

What is Universal in Perceiving, Remembering, and Describing Event Temporal Relations?

Shulan Lu (slu@memphis.edu)
Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building, University of Memphis,
Memphis, TN 38152, USA

Abstract

What temporal relations do humans use to form dynamic mental representations of events? In the fields of artificial intelligence and computational linguistics, some have proposed an interval based representation, in which two events could be related in time by seven primitives. The seven primitives are BEFORE, MEET, OVERLAP, START, DURING, FINISH, and EQUAL. In the present study, perception, memory, and language about event temporal relations were investigated. The results showed that BEFORE, MEET, and DURING seem to be prevalent in the temporal experiences across a range of cognitive tasks, despite that there is variability with respect to different cognitive tasks.

Event Temporal Representation

Time is an inherent dimension of event representations (Freyd, 1987; Lu, 2003; Schank & Abelson, 1977; Zacks & Tversky, 2001; Zwaan, Magliano, & Graesser, 1995). In everyday life, there are many goal oriented activities that require an understanding of fine-tuned and subtle timings of events, as in the case of *making chicken soup*, *operating certain mechanical devices,* and *making a camp fire*. There are also many events that are loosely related, as in the case of *recalling some quarreling couple while dinning at a restaurant, hearing a loud sound from a house while taking a walk,* and *seeing a squirrel running on the electrical wire while walking past a fence.* How do people construct the temporal relations of events that may or may not be related in an overarching conceptual structure? For the purpose of this paper, the term event will be used as a covering term for both intentional actions (as in the case of an agent *making chicken soup*) and events that are not governed by the goals of an agent (as the in the case of *oil turning smoky*).

Previous research suggests that the plan and goal structures of everyday activities play an extremely important role in the encoding and retrieval of event temporal relations (Lichtenstein & Brewer, 1980). When people are asked to recall the sequence of events, they often place events in an order that maps onto the logical inferences derived from the goal and causal constraints instead of the actual order in time (Bauer & Mandler, 1989; Lichtenstein & Brewer, 1980). It is notable that the type of the events investigated in these studies tends to be the case where one agent manipulates one object and enacts one action at a time (Lu, 2003), whereas the examples in the previous paragraph seem to suggest that humans often experience events that have overlap in time. Additional research is needed to specify the details of how the plan and goal based theories could account for events overlapping in time.

How are events represented in temporal dimension? There are two types of primitives for temporally representing events (Allen, 1984; 1991). A point based representation captures events as being indexed as points in time. One event can be a single point in time, as in the case of *a sunrise at 4:30* or a *hiccup.* There are many singular point expressions in natural language (Moens & Steedman, 1988; TerMeulen, 1995); the events described in these point expressions appear to be conceptually instantaneous. A point based representation can also represent non-instantaneous events with a set of points in time. For example, a person's cleaning the fish tank can have a beginning at 2:15 and an end at 2:35 p.m. Its sub-event *getting the supplies* begins at 2:16 and ends at 2:19 p.m. Each of these events has points in time marking the beginning and the end.

In contrast, an interval based representation captures events as durations that may gloss over exact time points. Thus, the interval of *getting the supplies* occurred during the interval of *cleaning the fish* tank, without any specification of the exact points in time that mark the beginning and end points of events. Psychological studies have reported that people have a grasp of the range of time during which an event occurs (Golding, Magliano, & Hempill, 1992; Loftus, Schooler, Boone, & Kline, 1987). For example, John may not know exactly at which points of time he opens his car door, yet he knows it takes two or three seconds to open it. The chief theoretical challenge lies in specifying how to relate the intervals of events and how to draw inferences about the relative timing of events on the basis of interval constraints.

In the fields of artificial intelligence and computational linguistics, Allen (1984; 1991) developed a formalism that captures the various temporal relations between two events that are represented in intervals. Figure 1 provides an illustration of these seven relational structures: BEFORE, MEET, OVERLAP, START, DURING, FINISH, and EQUAL. In Figure 1, each double-headed arrow represents an event that occurs over some time interval, and each arrow-head represents either the beginning or the end of an event. The relation between each pair of events is described by one of the seven predicates. The BEFORE relation means that one event is prior to another event and that two events do not overlap in any way, whereas the MEET relation means that one event starts at the time another event ends. START means that two events share the same

beginning, but one ends before another, whereas EQUAL means that two events share the same interval and the same beginning and end. These primitives are essential for constructing a computational system of event representations (Allen, 1984). To what extent are Allen's seven relations used as conceptual primitives in event temporal representation?

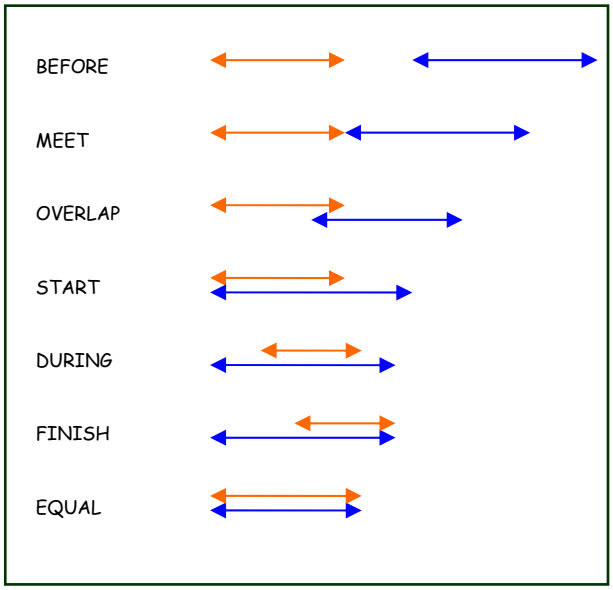


Figure 1: Temporal Representations by Allen (1991).

The formalism laid out in Figure 1 may capture some intuitive aspects of human temporal reasoning. For example, there is some evidence that endorses the distinction between BEFORE and MEET. Kate, a language found in Papua New Guinea, makes grammatical distinctions between the following two types of events: (a) events that are separated by a period of time with nothing significant, and (b) events that have successive temporal relations (Grime, 1975; but see Zwaan & Radvansky, 1998, pp. 176). In the Newtson task, participants are asked to segment a videotape of an activity into events and their parts (Newtson & Engquist, 1976). Participants are told to press the spacebar when they think one event ends and another begins. This methodology implicitly endorses MEET as the typical temporal experience encountered in the world.

Gestalt laws of perception postulate that forms are easier to process psychologically if they have more redundancy in pattern and permit fewer alternative forms. Conversely, forms are harder to process psychologically if they have less redundancy in pattern and render more alternative forms (Garner, 1974; Rock & Palmer, 1990). Examining Figure 1 based on Gestalt laws of perception, the complexity of the seven temporal relations seems to vary. For example, EQUAL may not have alternative forms, whereas START may have several alternative forms pending on the intervals of events. It is reasonable to infer that humans may capture some relations easily, but have some other relations confused (Lu, Graesser & Wolff, 2003).

In this paper, four experiments are reported to investigate the temporal relations people tend to construct. In Experiments 1 and 2, animated events of fish swimming were presented, and judgments of temporal relations were made. In Experiment 3, events were presented linguistically, and a production task was used. In Experiment 4, three separate sentence sorting tasks were conducted to see what semantic distinctions humans make when they describe events and their temporal relations.

Experiment 1: Perception of animated events

Experiment 1 investigated which temporal relations humans tend to construct out of the 7 primitives in the context of event perception. Participants were presented with animations of fish swimming events made in an animation program, called 3D Studio Max. Participants were asked to make judgments about which temporal relation out of the 7 choices best captures the animated events they saw.

Participants There were 51 college students at the University of Memphis who participated for course credit.

Materials Forty-two 3D animations were made using an animation program called 3D Studio Max release 5. Each animation depicted two fish of different colors and sizes swimming in the water. The spatial trajectory of the fish swimming was a straight line. For each relational structure in Allen's scheme, there were two sets of animations. One set of animations holds the distances of fish swimming constant but varies the speed of fish swimming, whereas the other set of animations holds the speed of fish swimming constant but varies the distances. For each set, there were three different perspective combinations: horizontal – horizontal, vertical - vertical, and horizontal - vertical.

The animation quality is near photorealistic. Each animation was 25 seconds in length and was run at approximately 30 frames / second.

Procedure Each participant was seated in front of a Pentium computer, which used MediaLab 2000 (Jarvis, 2000) to display the materials. Participants were asked to make judgments concerning how fish swimming events were related in time, as discussed below.

Participants were shown Figure 1 (without linguistic labels), and steps were taken to make sure they understood Figure 1. Participants were instructed to choose one relation out of the seven which best captured how two animated events were related in time. Before the animations were launched, participants were told that they could only have one viewing of each animation and that the screen with 7 choices would automatically pop up after each animation. Participants made choices by clicking a number that was next to the temporal relation.

Each participant received the same order of the temporal relations depicted in a diagram throughout the experiment. There were 20 sets of orders in which the temporal relations were presented. For each participant, the animations were presented in a random order.

Confusability Analysis Entropy was used to calculate the conceptual distance between each pair out of the 7 temporal

relations. The construct of entropy originated in information theory, which is a mathematical formulation of the uncertainty in a data set (Shannon, 1948). In the current study, each item may have 7 types of responses. For a given item, the following formula computes how likely one relation is confused with another one.

$$E_i = -\frac{\sum_{i=1}^{N} p_i \ln p_i}{\ln N}$$

The p_i refers to the proportion of times a given choice is selected out of the N possible choice items.

The entropy gives an index of how similar any two given structures (e.g. BEFORE and MEET) appeared to participants. A similarity matrix can thus be constructed, and then entered into the multidimensional scaling program implemented in SYSTAT version 9 with Young's S-STRESS scaling method (Wolff & Song, in press).

Results and Discussions The probability of people making the correct judgment of temporal relations was .80 on average. The error rates of the seven relations were the following: BEFORE (.20), MEET (.29), OVERLAP (.33), START (.15), DURING (.29), FINISH (.26), and EQUAL (.07). The EQUAL relation has the lowest error rates.

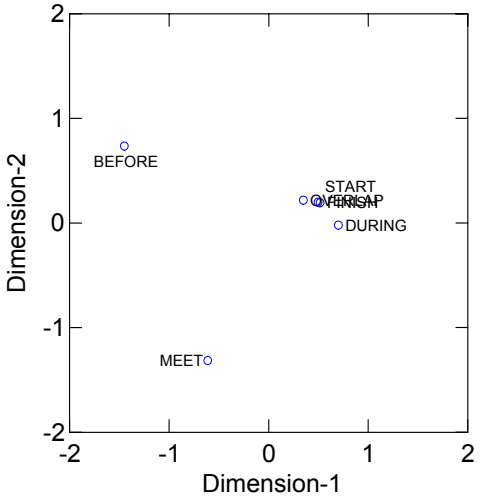


Figure 2: MDS Solution of Perception Task.

The confusability index was computed in entropy E_i. The similarity matrix was constructed using the entropy formula. This similarity matrix was submitted to the multidimensional scaling program. The MDS solutions in Figure 2 showed the following pattern of structure clustering yielded from the perception task. The similarity matrix was fit by a 2-dimensional MDS solution, with a very low stress value (.01), and a high proportion of variance accounted for (R^2 = 0.99). The seven temporal relations in Figure 1 were clustered into three main groups: BEFORE, MEET, versus (OVERLAP, FINISH, START, EQUAL, DURING).

The results showed that people tend to make distinctions whether events have overlap in time, as BEFORE and MEET stand out from the rest of the five temporal relations. People seem to make mistakes often among the five temporal relations that have overlap in time.

Experiment 2: Perception of animated events in the speeded condition

In Experiment 1, participants were given the luxury of focusing on two events. Experiment 2 investigated whether the effects observed in Experiment 1 are the result of people having enough attentional and cognitive resources during event perception. The animations used in Experiment 1 were presented at a faster rate (see Graesser & Nakamura, 1982; Reiger & Zheng, 2003, for the same method). Participants made judgments of the animated events in the same way as Experiments 1.

Participants There were 40 college students at the University of Memphis who participated for course credit.

Materials The same set of 42 3D animations used in Experiment 1 was used in Experiment 2. The animations were speeded up using an animation program called VirtualDub. The animations were displayed 30 frames / second in Experiment 1, whereas the animations was speeded to 75 frames / second in Experiment 2.

Results and Discussions The probability of people making the correct judgment of temporal relations was .66 on average. The error rates of the seven relations were the following: BEFORE (.34), MEET (.40), OVERLAP (.40), START (.23), DURING (.41), FINISH (.43), and EQUAL (.10). Compared with Experiment 1, the error frequency increased in Experiment 2.

The similarity matrix was constructed in the same way as Experiment 1, and then was submitted to the multidimensional scaling program. The MDS solutions showed the same pattern of structure clustering yielded as in Experiment 1. The seven temporal relations were clustered into three main groups: BEFORE, MEET, versus (OVERLAP, FINISH, START, EQUAL, DURING). The relational clustering in Experiment 2 did not differ from Experiment 1. Whether two events had overlap in time was again used to distinguish BEFORE and MEET from the rest of the temporal relations. Experiment 2 ruled out the possibility that the pattern observed in Experiment 1 was merely the effect of attentional allocation.

Experiment 3: Memory of everyday event time using a drawing task

Experiments 1 and 2 investigated the perception of events and their temporal relations. The question at this point is whether the relational clustering in the perception of everyday events will also be observed in the memory of everyday activities. In the perception judgment experiments, participants were provided Allen's seven relations. We were curious to find out whether the same pattern will be observed if Allen's seven relations are not provided. Participants were presented pairs of everyday events that were coded to have the temporal relations in Allen's representation. Participants read pairs of events, and then

drew which temporal relation in Allen's representation best captures the events they read.

Participants There were 34 college students at the University of Memphis who participated for course credit.

Materials A sample of events from everyday activities were collected. To ensure generality, the events were chosen from a wide range of everyday activities. Some examples include: driving a car, grocery shopping, cashing a check, and so on. Three raters were trained to understand Allen's representation, and made judgments on how each two events were related in time separately. The materials used in the experiment were the items agreed upon by all three judges.

For every pair of events, a supporting context was provided. An example is below:

Context: Imagine a passenger at an airport.
Events: She went through the security screening.
Her carry-on bags were x-rayed.

For each of the 7 temporal relations in Allen's proposal, there were 10 test items. There were 70 test items in total.

Procedure Participants were introduced how to represent an event occurring over some time with beginning and end points. Then they were given two examples: one indicating one event occurring before another, the other indicating one event occurring in the middle of another. The events were presented on A4 size papers, and space was provided for drawing. Allen's 7 temporal relations were used to code the drawings. The inter-rater agreement was 96%.

Results and Discussions The probability of people making the correct judgment of temporal relations was .29 on average. The error rates of the seven relations were the following: BEFORE (.31), MEET (.83), OVERLAP (.82), START (.95), DURING (.73), FINISH (.83), and EQUAL (.51). The probability of drawing START was significantly lower than chance, whereas the probability of drawing other six relations was higher than chance. For example, compared with the probability of drawing FINISH, the probability of drawing START are significantly lower, $t(33) = 3.021$, $p < .005$, using Bonferroni correction.

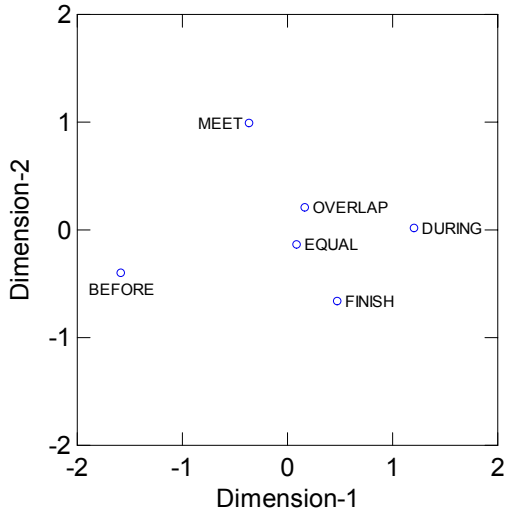


Figure 3: Drawing MDS Solution.

The similarity matrix, which was used to fit a 2-dimensional MDS solution, does not include START. The resulting MDS solution had a very low stress value (.01), and a high proportion of variance accounted for ($R^2 = 0.99$). The six temporal relations were clustered into three main groups with varying distances: BEFORE, MEET, versus (OVERLAP, FINISH, DURING, EQUAL). Compared with the MDS solution in the perception experiments, the distances among the four relations that have overlap in time are farther apart in the drawing task. Nonetheless, the clustering formed by the rest of the 6 temporal relations, are not incompatible with the previous perception tasks. Whether two events have overlap in time continued to be an important dimension in the memory of everyday events.

Experiment 4: Semantic organization of temporal lexemes

Children tend to describe an event when they answer questions about time (Nelson, 1996). Consider such an example. *When do you go to bed? When mommy comes to get me.* The child did not say something like *at night*. This example suggests that how we talk about time is correlated with how we think of time (Gentner & Boroditsky, 2001). Does the pattern of event temporal relations observed in the perception and memory of events reflect in the language about event temporal relations?

Experiment 4 investigated the semantic clustering of temporal event language when college students performed linguistic sorting tasks. Participants were presented sentences with temporal words embedded in them. They sorted the sentences into groups which shared similar meanings of the embedded words. The central question is whether the words would cluster according to theoretically interesting dimensions. Three experiments were conducted separately for verbs, adverbs, and prepositions plus conjunctions.

Experiment 4 (a): Verb sorting

Participants There were 27 college students at the University of Memphis who participated for course credit.

Materials A list of verbs and their synonyms encoding how two events are related in time was assembled from several thesauruses. Each of the 17 verbs in (4a) was printed at the top of a 4" x 6" index card. Below each verb were printed two sentences that illustrated the use of the verb. The example sentences were selected from the British National Corpus.

Procedure Participants were asked to read the sentences on the verb index card, and then sort the verb index cards into as many or as few groups as they felt appropriate. They were told that the cards in each group should have "essentially the same meaning".

Results The frequency of each pair of words co-occurring in the same group was scored and assembled in a word-pair co-occurrence matrix. The MDS solutions showed a pattern of verb clustering. The sorts were fit by a 2-dimensional MDS solution, with a very low stress value (.18), and a high proportion of variance accounted for ($R^2 = 0.87$). The verbs

in (4a) were sorted into three main groups as shown in Figure 4:
BEFORE – type verbs (anticipate, be before, foresee, go before, and precede);
AFTER – type verbs (come after, go after, follow after, succeed, and result);
DURING – type verbs (coincide, concur, co-occur, ensue, fall together, go with and overlap).

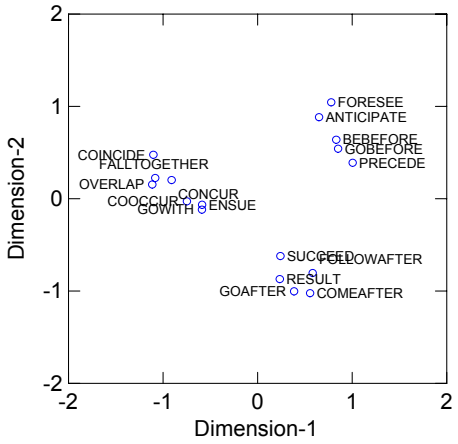


Figure 4: Verb Sorting MDS Solution.

Experiment 4 (b): Adverb sorting

There were 58 undergraduates at the University of Memphis who participated for course credit. Materials and procedure were the same as Experiment 4 (a).

Results The MDS solutions showed a pattern of adverb clustering. The sorts were fit by a 2-dimensional solution, with a very low stress value (.13), and a high proportion of variance accounted for (R^2 = 0.94). The adverbs in (4b) were sorted into three main groups:
BEFORE – type adverbs (before, beforehand, earlier, formerly, in advance, previously and sooner;
AFTER – type adverbs (after, afterwards, later, later on, next, sooner or later, and subsequently);
DURING – type adverbs (at the moment, at the same time, concomitantly, concurrently, contemporaneously, for now, in chorus, in concert, in the meantime, in the same breath, in time, in unison, instantaneously, meanwhile, on the beat, simultaneously, synchronously).

Experiment 4 (c): Preposition sorting

There were 76 undergraduates at the University of Memphis who participated for course credit.

Results The MDS solutions showed a pattern of preposition clustering. The sorts were fit by a 2-dimensional solution, with a very low stress value (.14), and a high proportion of variance accounted for (R^2 = 0.91). The prepositions in (4c) were sorted into three main groups:
BEFORE – type prepositions (before, prior to);
AFTER – type prepositions (after, soon after, as soon as, until, pending, by);
DURING – type prepositions (just as, when, as, along with, while, amid, during, in the course of, throughout, over).

Discussions of Experiment 4

Across grammatical categories, Experiment 4 showed that people are predisposed to three types of temporal relations when they talk about time. The results are compatible with the proposals by Wierzbicka (1973) in linguistics and Graesser et al. in psychology (Graesser, Wiemer-Hastings, & Wiemer-Hastings, 2001). The mapping between linguistic primitives and Allen's primitives may be more complex than the MDS solutions suggested. Nonetheless, the evidence in linguistics may indicate some alignment between the linguistic clustering and the conceptual clustering that emerged from the previous experiments.

The linguistic primitive AFTER may correspond to some aspects of MEET. Wierzbicka and others have noted that BEFORE and AFTER encode somewhat different temporal conceptions of events (Thompson & Longacre, 1985, but see Wierzbicka 2002; Wierzbicka, 2002). One difference is that the event in the BEFORE clause usually does not begin until the event in the main clause ends, whereas AFTER does not have this constraint. Consider the following two examples: (a) *John took out the water pipe before he drained the fish tank* and (b) *John washed the dishes after he cooked the dinner*. The two events in example (a) have to occur one before another, whereas *washing the dishes* in example (b) could occur before *finishing the cooking*. There is some empirical evidence in support of this conjecture. In a separate sentence rating task we conducted, two types of *after* sentences were selected from a corpus. One type refers to the situation where one occurs after another with a time interval in between, whereas a second type refers to the situation where the beginning of one event is after another event but both events have overlap in time. There were significant differences between the likelihood ratings of whether the two events described in each sentence have overlap in time.

Summary and Discussions

Four experiments investigated how people piece together events that may or may not have an overarching conceptual structure in a range of cognitive tasks. The results suggested that the distinctions made among BEFORE, MEET, and DURING seem to be prevalent when people perceive, remember, and describe how events are related in time.

However, there is apparently some variability in the MDS solution yielded from different cognitive tasks. When people had to linguistically encode and retrieve events, the event temporal representations may be harder to construct. The MDS solution in Figure 3 indicated the trend that BEFORE, MEET, and EQUAL are more likely to be constructed than other temporal relations. BEFORE, MEET, and EQUAL may be easier for humans to process because they have inherent simplicity, symmetry, and good forms, which are hallmarks of the Gestalt Law of Prägnanz (Rock & Palmer, 1990).

It is not clear at this point how the linguistic primitives map onto the primitives in perception and memory tasks. According to the relational relativity hypothesis, language

affects thought in more abstract domain (Gentner & Boroditsky, 2001; Gentner, 2003). Temporal constructs are abstract, this points to the possibility that the way we talk about time affects how we encode the temporal aspects of events. The MDS solutions in the current study indicate some alignment among the primitives across different cognitive tasks. However, the alignment may be more complicated and needs further investigations.

It appears that BEFORE, MEET, and DURING are used in the perception, memory, and linguistic experiences of temporal representations. When the representations are harder to construct, for example, in the case of linguistic retrieval of event representations, people are more likely to mistake one temporal relation with another. In further studies, it is necessary to investigate the cognitive principles that constrain the constructions of temporal event relations.

Acknowledgments

The research was supported by a grant from the National Science Foundation (REC0106965). We thank Phillip Wolff for discussions of this work, Srinivas Achunala for 3D animations, and Zhiqiang Cai for discussions on entropy. We also thank Stephanie Coe, Dorothy Presbury, and Amy Vitale for help with data collection.

References

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence, 23*, 123-154.

Allen, J. F. (1991). Time and time again: The many ways to represent time. *International Journal of Intelligent Systems, 6,* 341-355.

Bauer, P. J., & Mandler, J. M. (1989). One thing follows another: Effects of temporal structure on 1-to-2-year-olds' recall of events. *Developmental Psychology, 25*, 197-206.

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review, 94,* 427-438.

Garner, W. R. (1974). The processing of information and structure. Protomac, MD: Erlbaum Associates.

Gentner, D. (Eds.). (2003). *Language in mind: Advances in the study of language and thought.* Cambridge, MA: MIT Press.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge, England: Cambridge University Press.

Golding, J. M., Magliano, J., & Hemphill, D. (1992). When: A model for answering "when" questions about future events. In T. W. Lauer, E. Peacock, & A. C. Graesser (Eds.). *Questions and information systems.* Hillsdale, NJ: Earlbaum.

Graesser, A. C., & Nakamura, G. V. (1982). The impact of schemas on comprehension and memory. In G. H. Bower (Eds.), *The Psychology of Learning and Motivation, Vol. 16.* New York, NY: Academic Press.

Graesser, A. C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (2001). Constructing inferences and relations during text comprehension. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects.* Amsterdam: Benjamins.

Lichtenstein, E. D., & Brewer, W. F. (1980). Memory for goal-directed events. *Cognitive Psychology, 12,* 412-445.

Loftus, E. F., Schooler, J. W., Boone, S. M., & Kline, D. (1987). Time went by so slowly: Overestimation of event duration by males and females. *Applied Cognitive Psychology, 1,* 3-13.

Lu, S. (2003). Perceiving, imagining, and describing events. Unpublished manuscript, University of Memphis.

Lu, S., Graesser, A. C., & Wolff, P. (November, 2003). Perceptions and conceptions of time. Poster presented at the 44th Annual Meeting of the Psychonomic Society, Vancouver, Canada.

Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics, 14,* 15-28.

Nelson, K. (1996). *Language in cognitive development: Emergence of the mediated mind.* New York, NY: Cambridge University Press.

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12*, 436-450.

Regier, T., & Zheng, M. (2003). An attentional constraint on spatial meaning. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Rock, I., & Palmer, S. (1990). The legacy of Gestalt psychology. *Scientific American, 163,* 84-90.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379-423. Continued in following volume.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Ter Meulen, A. G. B. (1995). *Representing time in natural language: The dynamic interpretation of tense and aspect.* Cambridge, MA: MIT Press.

Wierzbicka, A. (1973). In search of a semantic model of time and space. In F. Kiefer, & N. Ruwet (Eds.), *Generative Grammar in Europe*. Dordrecht, Holland: D. Reidel Publishing.

Wierzbicka, A. (2002). Semantic primes and linguistic typology. In C. Goddard, & A. Wierzbicka (Eds.), *Meaning and Universal Grammar: Theory and Empirical Findings.* Amsterdam: John Benjamins.

Wolff, P., & Song, G. (in press). Models of causation and semantics of causal verbs. *Cognitive Psychology.*

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127,* 3-21.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 386-397.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation model in language comprehension and memory. *Psychological Bulletin, 123,* 162-185.

How Human Tutors Employ Analogy To Facilitate Understanding

Evelyn Lulis (elulis@cti.depaul.edu)
CTI, DePaul University
243 S. Wabash Avenue, Chicago, IL 60604 USA

Martha Evens (evens@iit.edu)
Department of Computer Science, Illinois Institute of Technology
10 West 31st Street, Chicago, IL 60616 USA

Joel Michael (jmichael@rush.edu)
Department of Molecular Biophysics and Physiology, Rush Medical College
1750 W. Harrison St., Chicago, IL 60612 USA

Abstract

A corpus consisting of eighty-one one-on-one tutoring sessions with first-year medical students carried on by two professors of physiology at Rush Medical College was analyzed for the use of analogies to facilitate understanding of the topics covered. Analogies were infrequently used, but had a positive effect on improving student comprehension of the topics tutored. The human tutor's goals, topics, discourse strategies, follow-up, and clarification in the presence of misunderstanding were analyzed with the long term goal of implementing analogies in an intelligent tutoring system.

Introduction

Analogies play a major role in learning. Eighty-one one-on-one tutoring sessions carried out by two professors of physiology at Rush Medical College were extensively marked for analogies using SGML. Instances of analogies were then classified in terms of the goals, targets, bases, and whether they were proposed by the student or the tutor.

Current advances in education, cognitive science, linguistics, and expert systems make it feasible to generate analogies in an intelligent tutoring system using a computational model. To date, as far as we know, no one has used full-scale natural language generation to implement analogies in an electronic tutoring system. The goal is to use computational models of memory retrieval and analog mapping to simulate the human tutor's behavior in our intelligent tutoring system, CIRCSIM-Tutor.

Analogies in Cognitive Science

Gentner defines analogies as:

> partial similarities between different situations that support further inferences. Specifically, analogy is a kind of similarity in which the same system of relations holds across different objects. Analogies thus capture parallels across different situations (Gentner, 1998, p.107).

Analogical reasoning is essential to cognitive ability (Gentner, 1998; Kurtz, Miao, & Gentner, 2001), and scientific inquiry and study (Dunbar, 1993; Goldblum, 2001; Michael & Modell, 2003; Modell, 2000; Thagard, 1997). Research studies exist that:

- analyze the way humans store and retrieve analogues from memory (Forbus, Gentner, & Law, 1995; Hofstadter, 2001; Holyoak, Gentner, & Kokinov, 2001; Holyoak & Thagard, 1995; Kokinov & Petrov, 2001; Kolodner, 1993)
- use computational models to simulate the results of human studies (Forbus, 2001; Forbus, Gentner, & Law, 1995; Holyoak, & Thagard, 1995)
- analyze the use of analogy in problem solving/reasoning (Holyoak & Thagard, 1985; Holyoak, Gentner, & Kokinov 2001; Kolodner, 1993; Thagard, 1997)
- analyze the use of analogies in education, medicine, and scientific inquiry (Dunbar, 1993, 1995; Goldblum, 2001; Thagard, 1997)

Gentner's (1983, 1998) structure mapping theory (Gentner & Markman, 1997; Holyoak, Gentner, & Kokinov, 2001; Holyoak & Thagard, 1995; Kurtz, Miao, & Gentner, 2001) seems to closely match the way our expert tutors work. New knowledge (the target) is learned by mapping its structure to existing knowledge (the base). Inferences are made from these mappings. The representation of mappings is discussed in length in Yan, Forbus, & Gentner (2003). When retrieving possible analogs from memory, the goal is to find mappings that have predictive value (Gentner, 1983).

Further studies have demonstrated that analogical encoding—the "process of comparing two examples and deriving an abstraction on the basis of their commonalities" (Loewenstein, Thompson, & Gentner, 1999, p. 586)—can be effective in facilitating the learning of similar problems. Abstractions of schemas gained through the intensive comparisons of two analogous concepts that are not fully

understood not only facilitate the understanding of the new pieces of information, but the general schemas derived can be applied to similar problems encountered later (Gentner, Loewenstein, & Thompson, 2003; Kurtz, Miao, & Gentner, 2001; Loewenstein, Thompson, & Gentner, 1999). Studies involving the learning of negotiation skills in undergraduates and graduate management students (Loewenstein, Thompson, & Gentner, 1999) and presentation of heat flow scenarios to teach the concept of heat flow (Kurtz, Miao, & Gentner, 2001) demonstrated that the intentional and intensive comparisons of two concepts that are not fully understood are as effective in knowledge transfer as structural alignment. Gentner (1983) demonstrated that this approach to teaching by analogy bypasses the common problem that humans have when trying to retrieve relevant information from memory to connect to new knowledge that one is attempting to learn. Mutual alignment is especially relevant to electronic tutoring systems that cannot always rely on the presence of existing knowledge when presenting new concepts.

Possible problems resulting from misunderstandings when reasoning analogically in a scientific domain are well-recognized (Feltovich, Spiro, & Coulson, 1989). Holyoak and Thagard (1995) have studied misconceptions and devised the multiconstraint theory that addresses the problems resulting from the use of inappropriate analogies. They recommend placing certain restrictions—of similarity, structure, and purpose—on the analogy. If all three constraints are met, only one interpretation of the analogy can be gleaned from the mapping. In cases where the three constraints are not met, misunderstandings can be identified and corrected. We have observed this behavior in our expert tutors' human sessions, as discussed below.

Analysis of Analogies Found in the Corpus

In order to understand our human tutoring session, one must first have background information on what is being tutored. The human body requires a blood pressure within a certain range to sustain life. The baroreceptor reflex is a negative feedback system that controls blood pressure in the cardiovascular system to ensure that the pressure remains within this range. When a perturbation in the system occurs, the response has three phases: direct response (DR) of the system to the perturbation, the reflex response (RR) to the new values of affected variables, and the steady state (SS), or state of the system after it has re-stabilized. CIRCSIM-Tutor asks the student to predict the qualitative changes in several important variables at all three stages. The variables are: Heart rate (HR), Cardiac Contractility (CC), Stroke volume (SV), Cardiac output (CO), Mean arterial pressure (MAP), Total peripheral resistance (TPR), Central venous pressure (CVP). Eighty-one hour-long tutoring sessions with first year medical students solving problems about the baroreceptor reflex were conducted by our experts, two professors of physiology at Rush Medical College, Joel Michael and Allen Rovick. Face-to-face sessions were recorded and transcribed. Keyboard-to-keyboard sessions were recorded using Computer Dialogue System (CDS) discussed in Li, Seu, Evens, Michael, & Rovick (1992). CDS forces each person to take turns typing. An annotation language based on SGML (Kim, Freedman, Glass, & Evens, 2002) was used to mark up the human sessions by hand. The following examples (discussed in Lulis & Evens, 2003; Lulis, Evens & Michael, 2003) were selected from the analogies found in these expert human tutoring sessions. They are representative of sessions where the tutor uses analogies: new material is explained, misconceptions are corrected, and prompts—successful and unsuccessful—are made to the student to make analogies and inferences. In each of the examples listed, the tutors used analogies after the student made an incorrect inference. The identifiers at the beginning of each sentence make it possible to find the original context at any time: initial F or K indicates whether the session was face-to-face or keyboard-to-keyboard; the session number comes next; st (student) or tu (tutor) indicates who is speaking/typing; this is followed by the turn number and the number of the sentence within the turn. A complete set of marked-up transcripts will be provided on request.

Example 1. Face-to-face session number one (F1) contains examples of the use of analogy to explain domain material and a correction by the tutor. The analogy of comparing the heart to a sink is proposed by the student (st). However, the sink is not a compliant object and the heart is. As a result, the tutor (tu) offers a better analogy—the heart is like a balloon.

F1-st-62-1: If I make an analogy of you try to fill a sink with water and you...
F1-tu-63-1: Try to fill a balloon with water, since that's what we're dealing with, a distensible object.
F1-st-64-1: OK.

After making a one-to-one mapping of the base (balloon) to the target (heart), a correct inference is made. In accordance with Holyoak and Thagard (1995) and Gentner's (1983) theory of structure mapping, the following structures underlie what the tutor does (as discussed in Lulis & Evens, 2003; Lulis, Evens, & Michael, 2003):

Structure for the balloon
- fill a balloon with water
- it will distend
- the pressure in the balloon increases as it distends

Structure for the heart
- fill the right atrium
- the right atrium will distend
- the pressure will increase as it distends

The above example demonstrated the effectiveness of the accepted structure mapping approach of connecting new knowledge to knowledge already understood by the student.

As a result, the student develops a better understanding of the new concept (Gentner, 1983, 1998; Goldblum, 2001; Holyoak & Thagard, 1995).

Example 2. We see the tutor correcting a misconception in face-to-face session #7.

F7-tu-267-1: Well, let's give it another thought, OK?
F7-tu-267-2: We can look at that central blood chamber that means the big veins and the atria together as though they were an elastic chamber.
F7-tu-267-3: Is that not correct?
F7-st-268-1: Yeah, and the heart is the pump.
F7-tu-269-1: Well, let's stick to this elastic chamber and look at it first more or less in isolation.
F7-tu-269-2: If you have an elastic chamber what are the things that determine the pressure inside that chamber.
F7-st-270-1: Size.
F7-st-270-2: No.
F7-st-270-3: I mean if you..
F7-st-270-4: I mean...
F7-st-270-5: Area is one but I gather for the heart..
F7-tu-271-1: Area of what?
F7-st-272-1: Area that..
F7-st-272-2: I mean if you want to know what the pressure is of a gas or well liquids aren't that...
F7-st-272-3: We're not talking about gas, we're talking about liquids.
F7-st-272-4: And liquids are not affected by size because you can't compress the molecules that much.
F7-tu-273-1: Oh, you mean the volume occupied by the liquid, expansion and condensation of the liquid.
F7-tu-273-2: No.
F7-tu-273-3: That's not an issue.
F7-st-274-1: No, because we're talking about liquids and liquids aren't affected.
F7-st-274-2: Like with gas, besides the container matters a lot....
F7-tu-275-1: Let's throw away this atria central venous system and take instead something inanimate elastic stretcher, say like a balloon.
F7-tu-275-2: Right?
F7-tu-275-3: What determines what the pressure is inside the balloon?

In the above example, an analogy of the atria as an elastic chamber is proposed by the tutor, but the structure of the intended base analog—the atria or central venous system as a reservoir—is not understood by the student resulting in a failure to make the connection between existing knowledge and the new knowledge. The tutor suggests a base that the student is familiar with—a balloon to compare the atria to. A correct structure mapping is made allowing the student to understand the new concept.

Example 3. In keyboard-to-keyboard session number one (K1), the tutor prompts the student to make an analogy between neurally controlled variables in the DR (direct response) period. The behavior of the variable TPR during the DR period has already been tutored. The inference is that all neurally controlled variables will behave similarly during this period. The student makes the connection between the analogs resulting in a correct inference.

K1-st-29-1:During the DR period it would remain unchanged.
K1-tu-30-1: Correct.
K1-tu-30-2: What other variables are neurally controlled?
K1-st-31-1: CC, HR
K1-tu-32-1: Again correct.

Example 4. An example of the student not getting the point of the analogy is seen in keyboard-to-keyboard session #5 (K5). After tutoring the neurally controlled variable CC, the tutor attempts to get the student to make an analogy between it and other neurally controlled variables. The student fails to make the analogy and the tutor finally tells the student the analogy that he is seeking.

K5-tu-31-1: No, it is under neural (sympathetic) control,rily.
K5-tu-31-2: We are now discussing what happens in the DR period, before any neural (reflex) effects.
K5-tu-31-3: Sowhat will happ en to CC in this pe;riod?
K5-st-32-1: Cc 0.
K5-tu-33-1: Correct.
K5-tu-33-2: While we're on the subject, what other variable is underneural contro l and how will it be affected in the DR?
K5-st-34-1: SV 0.
K5-st-34-2: It will not be affected in the DR.
K5-tu-35-1: Not true.
K5-tu-35-2: SV is only partially under neural control via CC, one of its importan t determinants.
K5-tu-35-3: We'll come back to this in a moment.
K5-tu-35-4: What othervariable is un der neural control?
K5-tu-39-3: LIKE CC!.
K5-st-40-1: TPR 0.

Example 5. More interesting analogies can be found in several sessions. Keyboard-to-keyboard session number sixty-four (K64) contains an analogy proposed by the student (st) between the heart and a traffic cop. The mapping between these analogs is not correct; the tutor (tu) proposes a more suitable analogy between the heart and a traffic jam. The structure mapping theory discussed in Gentner (1983, 1998), Goldblum (2001), Holyoak and Thagard (1995), and formalizes what the tutor is doing.

K64-st-54-1: Would it be a reasonable analogy to look at the heart like a traffic cop?
K64-st-54-2: If it slows down the rate of blood flow (lets fewer cars through) then there will be a

backup behind it (a backflow of blood prior to the heart, and therefore an increase in CVP) and fewer cars coming through (less blood coming out of the heart and therefore a decrease in MAP)

K64-tu-55-1: The analogy is OK.

K64-tu-55-2: But just as traffic jam does not occur because cars back up, the increase in CVP caused by a fall in CO is not the result of blood BACKING UP.

K64-tu-55-3: Everything soes in one direction.

K64-st-56-1: well, slowing down would be a better way to put it then

K64-tu-57-1: Yes.

K64-tu-57-2: A traffic jam caused by everybody piling into the same area at once.

Analogies in Human Tutoring Sessions

In the tutoring sessions that we have studied, we observe expert tutors taking steps to avoid misconceptions. They (Holyoak & Thagard, 1995):

- Make certain that students understand the system mapping.
- Use a variety of analogies.
- Inform students when an analogy is relevant and when it is not—point out the differences, as well as the similarities, between the known knowledge and the target.
- Correct misconceptions when they occur.

The outcomes of the analogies proposed by the tutor are shown in Table 1 (as discussed in Lulis & Evens, 2003; Lulis, Evens, & Michael, 2003). We summarize the analogies that we found in human tutoring sessions described here.

Table 1: Use of observed analogies proposed by tutors

Type	No. observed in corpus
no inference requested	**5**
successful mapping	4
failed mapping	1
inference requested	**37**
successful inference	15
failed inference	
success after repair	15
failure after repair	7
enhancement only	**9**
Total:	**51**

Out of the fifty-one analogies proposed by the tutors, nine were used after correct inferences and apparently intended to enhance the student's understanding of the material discussed and not to lead to further development. In forty-two cases, the tutor used analogies after the students made incorrect inferences. In five of the forty-two cases, the tutors did not request inferences from the students. However, students did make correct inferences four out of the five times without prompting. In the remaining thirty-seven cases, an inference was requested after the analogy was proposed resulting in correct inferences being made by students fifteen times without repair to the analogy (to correct misunderstandings) and fifteen times with repair—81% success rate. In only seven of the thirty-seven cases—19% of the time—did the tutor abandon the use of analogy and opt for a different teaching strategy. In total, the use of analogy after an incorrect prediction was followed by a correct prediction in 34 out of the 42 times—81% success rate. The empirical evidence suggests that the use of analogy had positive affects on the students' ability to understand the material.

If we examine the different bases employed while tutoring using analogies—proposed by students and tutors—we find a wide range, as shown in Table 2. The analogy that was most often proposed by the tutors was another neural variable—twenty-nine times. In five of these cases, the tutors eventually gave up on the analogy and utilized a different approach to the material, but the other twenty-four were ultimately successful. There was one successful mapping without an attempt at an inference, twelve successful mappings with correct inferences, and four successful mappings with correct inferences after repairs. Other successful mappings occurred using in a wide variety of bases such as the heart as a balloon or pump, Ohm's law, airplane wings, bootstraps, a dimmer switch, traffic jams, and a black box. These bases were not observed as often,

Table 2: Bases present in the corpus

Base	No. observed in corpus
Airplane wing	1
Another algorithm	2
Another neural variable	29
Another procedure	3
Balloon	1
Balloon as a compliant structure	2
Black box	1
Bootstrap	1
Brake & accelerator	1
Compliant structure	3
Dimmer switch	1
Elastic reservoir	1
Flight or fight	1
Gravity	1
Last problem	1
Ohm's Law	2
Physician	1
Pump	1
Reflex	2
Sugar or glucose	1
Summation	1
Traffic jam	2

but made for extremely productive and interesting structural mappings resulting in correct inferences.

Gentner's (1983; Lowenstein, Thompson, & Gentner, 1999) work suggests that information from abstract and concrete bases may be processed differently. She has observed that children find it easier to understand analogies with concrete bases than with abstract ones. We hope to investigate this phenomenon using CIRCSIM-Tutor. In our data in Table 2, we see twenty-two different bases, twelve are concrete and ten are abstract (Table 3). The use of abstract bases are observed forty-four times in the corpus, while the concrete bases are used only fifteen times. Examination of the language used suggests another potentially useful classification—into analogies that remind students of earlier experience with another neural variable or another procedure and those that depend on a base from outside the immediate domain.

Table 3: Analogies with abstract and concrete bases

No. Type of base	No. of times seen in the corpus
12 different concrete	15
10 different abstract	44

Implementation

Holyoak & Thagard (1995) identified the steps of analogical reasoning: the retrieval of possible analogs from memory, the mapping of these analogs to the new knowledge being learned, inferring something from the mapping, adjusting the new knowledge if necessary, and storing the new knowledge for future use. Computational models dealing with analogy address the first two steps—retrieval based on similarity and structural mapping. There are two dominating models for the retrieval step—case based reasoning (Birnbaum & Collins, 1989; Kass, 1990, 1994; Kolodner, 1984, 1993, 1994; Schank, 1982) and a model that emulates a document retrieval system, retrieving both relevant analogs and irrelevant ones. There are also two approaches to the mapping step. One makes inferences before the mappings—projection first—the other makes the mappings before the predictions—alignment first.

It is our goal to implement an analogy generating function in CIRCSIM-Tutor (Michael, Rovick, Glass, Zhou, & Evens, 2003). It has been decided that a document retrieval model coupled with an alignment-first mapping—MAC/FAC—(Gentner, 1998; Gentner & Markman, 1997; Forbus, Gentner, Everett, & Wu, 1997) was best suited for use when simulating human tutoring in CIRCSIM-Tutor System. MAC/FAC was chosen because we believe that it simulates how people process analogies and it the implementation is very successful.

MAC/FAC

MAC/FAC (Many Are Called/Few Are Chosen) models Gentner's (1983) theory of structure mapping and simulates the human propensity to favor relationships between bases and targets when comparisons are made and to favor superficial similarities and not retrieve the more profound analogical similarities while still, on occasion, retrieving relevant structural comparisons (Forbus, Gentner, & Law, 1995). Working memory consists of content vectors constructed from the structural representations of the bases. The MAC stage functions like a document retrieval system, searching working memory in a parallel manner seeking content vectors that are similar to the target. The dot product between each of the bases and the target is computed to determine the best and those within 10% of the best matches. Stage two, the FAC stage, utilizes the output from the MAC phase to do Gentner's (1983) structure mappings. The structure mapping engine (SME) selects the best mapping and all those within 10% of it.

Conclusion

Analogies are used by our human tutors infrequently; on the average, less than once a session. However, the human sessions have demonstrated that the use of analogies is extremely effective. We have observed tutors using analogy to tutor the topic at hand and to enhance existing knowledge. Misunderstandings were corrected and inappropriate analogies replaced with more suitable ones. The structure mappings between the analogs underlie what the tutor was doing.

Future research includes simulating the schemas observed in the corpus in our expert system CIRCSIM-Tutor (Michael et al., 2003). Many of the analogies observed can be implemented using structure mapping (Gentner, 1983, 1998; Goldblum, 2001; Holyoak & Thagard, 1995) to connect new knowledge to existing knowledge. We will attempt to simulate mutual alignment (Gentner, Loewenstein, & Thompson, 2003; Loewenstein, Thompson, & Gentner, 1999) for the most commonly found analogy in the corpus—another neurally controlled variable. The recommendations of Goldblum (2001), and Holyoak & Thagard (1995)—use more than one analog, detect and fix incorrect mappings, identify the analogical scope, and refine analogies—will also be attempted.

Acknowledgments

This work was partially supported by the Cognitive Science Program, Office of Naval Research under Grant 00014-00-1-0660 to Stanford University as well as Grants No. N00014-94-1-0338 and N00014-02-1-0442 to Illinois Institute of Technology. The content does not reflect the position or policy of the government and no official endorsement should be inferred.

References

Birnbaum, L. and Collins, G. (1989). Remindings and engineering design themes: A case study in indexing vocabulary. *Proceedings of the Workshop on Case-Based Reasoning.* Pensacola Beach, FL: 47-51.

Dunbar, K. (1993). Concept discovery in scientific domain. Cognitive Science, 17, 391-434.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real world laboratories. In R. J. Sternberg & J. Davidson (eds.), *The Nature of Insight*. Cambridge, MA: MIT Press.

Feltovich, P.J., Spiro, R., & Coulson, R. (1989). The nature of conceptual understanding in biomedicine: The deep structure of complex ideas and the development of misconceptions. In D. Evans and V. Patel (eds.), *Cognitive Science in Medicine*. Cambridge, MA: MIT Press.

Forbus, K. D. (2001). Exploring analogy in the large. In D. Gentner, K. J. Holyoak, and B. N. Kokinov, (Eds.), *The Analogical Mind*, . Cambridge, MA: MIT Press.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*(2), 141-205.

Forbus, K. D., Gentner, D., Everett, J. O., & Wu, M. (1997). Towards a computational model of evaluating and using analogical inferences. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, 229-234. Mahwah, NJ: Lawrence Erlbaum Associates.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science 7*(2):155-170.

Gentner, D. (1998). Analogy. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science*, (pp. 107-113). Oxford: Blackwell.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1): 45-56.

Gentner, D., Loewenstein, J., Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology,* 95(2):393-408.

Goldblum, N. (2001). *The brain-shaped mind.* New York: Cambridge University Press.

Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov, (eds.), *The analogical mind*, Cambridge, MA: MIT Press.

Holyoak, K. J., & Thagard, P. R. (1995). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Holyoak, K. J., Gentner, D., & Kokinov, B. N. (2001). Introduction: The place of analogy in cognition. In D. Gentner, K. J. Holyoak, and B. N. Kokinov, (eds.), *The analogical mind*. Cambridge, MA: MIT Press.

Kass, A.M. (1990). Developing creative hypotheses by adapting explanations (Unpublished doctoral dissertation). New Haven, CT: Computer Science Department, Yale University.

Kass, A. M. (1994). Tweaker: Adapting old explanations to new situations. In Roger C. Schank, Alex Kass, & Christopher K. Riesbeck (eds.), *Inside case-based explanation*. Hillsdale, NJ: Erlbaum.

Kim, J. H., Freedman, R., Glass, M., & Evens, M. W. (2002). Annotation of tutorial goals for natural language generation. Unpublished paper, Department of Computer Science, Illinois Institute of Technology.

Kokinov, B. N. & Petrov, A., (2001). Integrating memory and reasoning in analogy-making: The AMBR model. In D. Gentner, K. J. Holyoak, & B. N. Kokinov, (eds.), *The Analogical mind.* Cambridge, MA: MIT Press.

Kolodner, J. L. (1984). *Retrieval and organizational strategies in conceptual memory.* Hillsdale, NJ: Lawrence Erlbaum.

Kolodner, J. L. (1993). *Case-based reasoning,* San Mateo, CA: Morgan Kaufmann.

Kolodner, J. L. (1994). From natural language understanding to case-based reasoning and beyond: A perspective on the cognitive model that ties it all together. In E. Langer and R.C. Schank (eds.), *Reasoning and Decision Making: Psycho-Logic in Honor of Bob Abelson*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Kurtz, K., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences,* 10(4):417-446.

Li, J., Seu, J. H., Evens, M. W., Michael, J. A., & Rovick, A. A. **(**1992). Computer dialogue system: A system for capturing computer-mediated dialogues. *Behavior Research Methods, Instruments, and Computer (Journal of the Psychonomic Society), 24*(4): 535-540.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review,* 6(4):586-597.

Lulis, E. & Evens, M. (2003). The use of analogies in human tutoring dialogues. *AAAI 7: 2003 Spring Symposium Series Natural Language Generation in Spoken and Written Dialogue*, 94-96.

Lulis, E., Evens, M., & Michael, J. (2003). Representation of analogies found in human tutoring sessions. *Proceedings of the Second IASTED International Conference on Information and Knowledge Sharing*, 88-93. Anaheim, CA:ACTA Press.

Michael, J. A. & Modell, H. I. (2003). *Active learning in the college and secondary science classroom: A model for helping the learner to learn.* Mahwah, NJ: Lawrence Erlbaum Associates.

Michael, J., Rovick, A., Glass, M., Zhou, Y., & Evens, M. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments, 11*(3): 233-262.

Modell, H. I. (2000). How to help students understand physiology? Emphasize general models. *Advances in Physiology Educ.* 23: 101-1+07.

Schank, R.C. (1982). *Dynamic memory.* Cambridge, UK: Cambridge University Press.

Thagard, P. (1997). Medical analogies: why and how. In P. Langley & M. Shafto (eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, (pp. 739-744). Mahway, NJ: Erlbaum.

Yan, J., Forbus, K., and Gentner, D. (2003). A theory of representation in analogical matching. *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society.*

Social and cultural influences on causal models of illness

Elizabeth B. Lynch (bethlynch@northwestern.edu)
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208 USA

Abstract

Causal models of illness vary extensively across socio-cultural groups. The current paper describes two studies that were designed to explore the role of universal domain-specific causal knowledge in causal models of illness. The first study compares illness causal models in three American groups: registered nurses, energy healers, and college undergraduates. The second study examines illness causal models in a group of Maya in Guatemala. In all groups illness models are composed of systematic combinations of domain-specific causes. It is argued that analysis of causal models in terms of domain-specific causal types reveals similarities in illness models that would be obscured by comparison of specific, detailed causes. The analysis of illness models as patterns of domain-specific causes suggests that American energy healers have models of illness that are more similar to those of the Maya than to illness models of American undergraduates and RNs.

Introduction

An issue of interest to both anthropologists and psychologists is the extent to which conceptual representations are affected by socio-cultural factors and the mechanisms by which this influence occurs. The extent to which thinking varies across cultures depends in large part on the content of the domain. For example, the domain of folkbiology is characterized by striking similarities across cultures, while less consistency has been observed in social attribution (Choi, Nisbett & Norenzayan, 1999) moral reasoning (Miller, Bersoff, & Harwood, 1990, Haidt, Koller, & Dias, 1993), and reasoning about illness (Murdock, 1980; Kleinman, 1978). In general, theories of cultural knowledge transmission that explain diversity of knowledge are distinct from ones that explain uniformity.

Explanations for cultural diversity often assume that the mind is a "blank slate," open to any form of knowledge (Atran, 2001; Pinker, 2000; Sperber & Hirschfeld, 2004). In contrast, explanations of uniformity in knowledge across cultures appeal to a view of the mind as a highly structured, modular information-processing device. As Pinker (2000), Atran (2001), and Sperber & Hirschfeld (2004) argue, the view of the mind as a blank slate is almost certainly wrong. There is plenty of evidence that the mind is not a blank slate, but is structured in a modular way such that qualitatively distinct reasoning processes are utilized for different kinds of phenomena (Carey & Gelman, 1991; Hirschfeld & Gelman, 1994, Pinker, 2004).

The mind, like other parts of the natural world, evolved in an environment with a particular structure. The modular nature of the mind is the result of evolutionary adaptations to the objective structure of the world. The mind evolved different cognitive processes in order to effectively predict the behavior of ontologically distinct objects. This view predicts and explains cross-cultural universals in human thinking. Sperber and Hirschfeld (2004) argue that stability of cultural knowledge results from the universal structure of the human mind, in particular the fact that all minds process information in similar, highly constrained ways. For example, there is striking cross-cultural uniformity in folkbiological knowledge. Medin and Atran (in press) argue that cross-cultural uniformity in thinking and behavior with regard to plants and animals is due to the existence of a universal cognitive module, the folkbiology module, that evolved specifically to process information about plants and animals. Despite differences in experience or environmental input, minds are universally constrained to construct a particular kind of representation of plants and animals – hence, cultural uniformity and stability.

The current paper presents two descriptive studies that demonstrate how universal domain-specific knowledge is expressed in causal models of illness, which are characterized by cross-cultural diversity rather than uniformity. Domain-specificity theory implies that, just as there are different ontological kinds of objects in the world (e.g. mental and physical objects), there are also different kinds of causal mechanisms. For example, there are psychological causal mechanisms, like intentionality, which explain behavior of animate objects, and there are physical causal mechanisms which explain the behavior of inanimate objects. One role of cognitive modules is to constrain the search for causal explanations by delimiting the range of possible causes for a particular phenomenon. Thus, causes can be divided into types based on the module with which they are associated. For example, *blocked arteries* and *chemical imbalance* are physical causes of illness. *Low self-esteem* and *problems in love relationships* are psychological causes. These causes differ in specific detail but are of the same type. Causal models of illness can be analyzed in terms of the kinds of causes of which they are composed.

One important dimension of variation among cultural belief systems about illness is whether illness is attributed to psychological causes – human or spiritual agents – or to natural causes (Murdock, 1980, Foster, 1976). In a world survey of illness theories, Murdock (1980) found that every cultural group in his sample (which did not include industrialized societies) explained (at least some) illness in terms of spirit aggression. Spirit aggression is a psychological cause. In contrast, biomedical theories of illness (used by medical doctors) explain illness using physical causes. One explanation for this difference is that

different cultural groups use different cognitive modules to process information about illness. That is, some cultural groups think about illness as a psychological phenomenon, and generate psycho-social attributions, while others construe illness as a physical phenomenon, and generate physical explanations. On this account, cultural environment influences the domain in which individuals search for causes of illness. Perhaps, some cultural groups use the domain of folk psychology to explain illness where other groups use the domain of folk physics (or folk biology). An alternative possibility is that multiple domains are used to process information about illness. Sperber & Hirschfeld (2004) suggest that some belief systems, like religion, maintain stability by being "anchored" in several cognitive domains at once. Thus, different cultural groups may combine different kinds of causes in different ways in their causal models of illness (Ahn, 1998).

The current paper will look at use of psycho-social causes and physical causes in the causal models of illness of three groups of participants. The groups were chosen precisely because they have different beliefs systems of illness. The first group consisted of registered nurses (RNs) who work as medical practitioners in professional medical settings and practice standard scientific biomedicine. The second group of participants consisted of energy healers. These individuals believe that illness is caused by a disruption or imbalance of "energy" in the body and that illness can be treated by balancing that energy. Energy healers often explain illness as the result of psychological problems (Eden, 1999). RNs and energy healers were chosen because they practice healing within belief systems that vary in the kinds of causes to which illness is attributed. Variation along the dimension of psychological versus physical attribution is cross-culturally salient, as mentioned above. While findings from these groups cannot be automatically generalized to other cultural groups, current findings can generate hypotheses about the causal models of other cultural groups which can then be tested. Study Two in the current paper demonstrates that this framework can be usefully extended to very different cultural settings.

Experiment One

The question of interest is how domain-specific causes are invoked in illness causal models of different groups. To measure causal models, causal chains leading to depression and heart attack were elicited from each participant in an open-ended format. The key feature of this study is the elicitation of causal chains rather than lists of causes. Elicitation of causal chains will provide evidence for distinguishing between three possible patterns of use of domain-specific causal information. The three possible patterns are:

1. Different groups use different domains of knowledge to construct causal models of illness. This hypothesis predicts that the causal chains of energy healers will consist of psychological causes while the causal chains of RNs will consist of physical causes.

2. Illness models are anchored in several domains. Causal models of illness are composed of both psychological and physical causes which are patterned systematically within groups.

3. A third possibility is that there is no systematicity in use of domain knowledge in illness models. That is, mental and physical causes may be distributed in different ways across different individuals and/or across different illnesses.

It is expected that energy healers will cite psycho-social causes of both illnesses more often than RNs. The question that will distinguish between the first and second alternatives above is whether illness models consist of a single kind of cause or multiple kinds of cause. The question that will distinguish between the second and third alternatives is whether different kinds of causes pattern systematically within groups.

In addition to the two practitioner groups, a group of undergraduates was also interviewed. Because the nurses may have had more experience with illness than the energy healers, undergraduates were included as an independent measure of the effect of experience on concepts of illness.

Method

Participants This study included three groups of participants. The first group consisted of 13 registered nurses (RNs) with an average age of 41 and an average of 13 years of nursing experience. The second group consisted of 14 energy healers with an average age of 48 and an average of 8 years of energy healing experience. Practitioners had an average of four years of college education and there was no difference in level of education across groups. The final group consisted of 23 undergraduates (UGs) with an average age of 18 and no energy or medical experience. Some undergraduates did only a single illness and others did both (5 did both illnesses, 10 only heart attack, 8 only depression). No differences were observed among those who did one versus both illnesses.

Procedure All participants were asked about heart attack, a physiological illness, and clinical depression, a psychological illness. The order of the illnesses was counterbalanced across participants. For each illness, participants were first asked to list all the causes of the illness and then, for each cause, were asked for causal chains linking each elicited cause to the target illness. To elicit the causal chain linking cause X to the illness, the experimenter asked, for example, "How does X cause illness A" (e.g. How does high blood pressure cause a heart attack?). The participant responded with an intermediary cause, Y. The experimenter then repeated the probe with

cause Y: "How does Y cause a heart attack?" This process was continued until the participant said the causal chain was complete. All interviews were recorded and transcribed.

Results

Across illnesses causes were collapsed into four types: physiological, psycho-social (henceforth called mental), behavioral, and energy. Depression models included an additional cause type, external environmental.

Depression Table 1 lists the average proportion of each type of cause in the models of participants from each group. Anovas were used to compare the proportion of each type of cause across groups. Because these measures are not independent a Bonferroni adjustment set the p-value for significance to 0.01. With this adjustment, only the difference in proportion of energy causes was reliably different across groups [$F(2,40)=17.77$, $p<.001$]. Proportion of physical causes [$F(2,40)=.925$, $p=.41$], mental causes [[$F(2,40)=4.12$, $p=.024$], and environmental causes [$F(2,40)=2.15$, $p<.13$] did not differ across groups. Proportion of behavioral causes [$F(2,40)=5.24$, $p=.01$] was marginally different across groups. Tukey post hoc tests showed that the Energy group cited slightly more behavioral causes than the UG group.

Table 1. Proportion of depression cause types by group.

CAUSE	UG	RN	EN
Physical	0.33	0.43	0.32
Mental	0.66	0.53	0.43
Behavioral	0.00	0.02	0.11
Environmental	0.01	0.02	0.10
Energy	0.00	0.00	0.19

The next set of analyses measured the types of causal relations in the conceptual models of each group. Systematic differences in the types of causal relations across groups indicates the extent to which groups represent different patterns of causal types within their models of illness. The majority of causal relations within the depression models of all groups consist of physical and mental factors (UG=99%, RN=96%, EN=75%) so analysis of causal relations focused on physical and mental causes only. Table 2 shows the proportion of each type of causal interaction among physical and mental causes across groups.

Table 2. Proportion of relations in depression models.

Causal Relation	UG	RN	EN
Physical-Physical	0.10	0.28	0.16
Mental-Mental	0.80	0.47	0.43
Physical-Mental	0.03	0.12	0.02
Mental-Physical	0.07	0.14	0.40

These measures are not independent so with a Bonferroni adjustment the p-value for significance was adjusted to 0.013. There were no differences between groups in the proportion of causal interactions among physical causes [$F(2,40)=1.4$, $p=.257$], nor in the proportion of physical-mental causal interactions [$F(2,40)=2.1$, $p=.13$], which were quite low across groups. The key difference between groups was that energy participants were more likely than other groups to cite mental-physical interactions [$F(2,40)=5.26$, $p=.01$]. Tukey post hoc tests showed that Energy healers cited more mental-physical interactions than RNs and UGs, who did not differ from one another. Undergraduates were more likely than RNs and Energy healers to cite interactions among mental causes [$F(2,40)=5.28$, $p=.01$]. Interactions among mental causes made up the bulk of undergraduate depression concepts.

The most important finding in depression models is that all groups included equal proportions of mental and physical causes in their conceptual models of depression but showed systematic differences in the patterns of causal interaction among them. Specifically, the RNs and UGs placed mental and physical causes on separate causal chains but Energy healers included both types of causes on a single causal chain.

Heart attack Table 3 shows proportions of causes in heart attack models across groups. Energy participants cited proportionately fewer physical causes of heart attack than RNs or Undergraduates [$F(2, 42 = 14.9$, $p<.0001$]. Post hoc tests indicated that the RNs and UGs cited equal proportions of physiological causes and EN participants cited fewer than both groups. Energy participants cited a greater number of psychological causes of heart attack than did either of the other groups, who were equivalent [$F(2, 42 = 34.9$, $p<.0001$]. These differences were reliable with a Bonferroni adjustment. Not surprisingly Energy participants cited a greater proportion of energy causes.

Table 3. Proportion of causes in heart attack models.

Relation	UG	RN	EN
PHYSICAL	0.70	0.77	0.45
PSYCHOLOGICAL	0.08	0.08	0.27
BEHAVIORAL	0.16	0.15	0.17
ENERGY	0.00	0.00	0.10

The next set of analyses explores patterns of causal relations in models of each group. A majority of the causes in heart attack models of all groups were ones with physical effects (RN=0.98, UG=0.99, DUAL=0.88, ENERGY=0.77), so the following set of analyses compares the proportion of physical effects that have either physical, mental, behavioral, or energy causes. Table 4 shows the distribution of these types of relations in individual models of participants in each group.

Table 4. Proportion of causal relations across groups.

Causal Interaction	UG	RN	EN
Physical-Physical	0.54	0.61	0.21
Physical-Mental	0.11	0.11	0.42
Behavioral-Physical	0.35	0.27	0.27
Energy-Physical	0.00	0.00	0.09

These measures are not independent but all effects are reliable with a Bonferroni adjustment. An ANOVA showed that Energy participants cited proportionately fewer physical – physical causal interactions (PP relations) than RNs and UGs [$F(2,42)=20.31$, $p<.0001$]. ENERGY participants mentioned the greatest number of mental-physical causal interactions in their heart attack models [$F(2,42)=23.73$, $p<.0001$]. There were no differences across groups in the frequency of behavioral-physical relations [$F(2,42)=.68$, $p=.51$]. Finally, Energy participants mentioned a greater number of energy-physical relations [$F(2,42)=9.79$, $p<.0001$]. In all causal relation analyses, Tukey post hoc tests showed that RNs and UGs were equivalent and both were different from Energy participants.

As in models of depression, energy healers and RNs and UGs showed different patterns of causal relations among the causes in their conceptual models of heart attack. Like in depression models, Energy healers cite causal interactions among mental and physical causes whereas RNs and UGs do not.

Summary Experiment 1 provides compelling evidence for Alternative two cited above – that illness models are anchored in several domains. Alternative One predicted that illness models would consist of a single kind of cause. This was not supported by current findings. All groups used both psycho-social and physical causes in illness models. Alternative Three was also ruled out, because mental and physical causes patterned systematically, rather than randomly, across participants within a group. Causes also patterned similarly across illnesses for each group. Across both illnesses energy healers frequently cited causal relations in which mental (psycho-social) causes led to physical effects. RNs and UGs rarely mentioned causal interactions between mental and physical features. For depression, mental and physical causes were conceived as distinct causal chains. For heart attack, these participants rarely mentioned mental causes at all. While energy healers combined mental and physical causes within a single causal chain, RNs and UGs kept mental and physical causes on separate causal chains. Further, they did not conceive of heart attack as psychologically caused.

Experiment Two

Whereas Experiment One included groups from a single cultural environment, the current experiment uses the same method to measure concepts of illness in individuals from a very different cultural environment, Peten, Guatemala. The question of interest is whether Maya have systematic patterns in their conceptual models of illness, and if so, whether their concepts correspond to either of the American groups.

Method

Participants Participants were 13 illiterate Itza' Maya adults living in Peten, Guatemala. All participants spoke Spanish as their primary language. Peten is a very different cultural environment from Chicago, IL where participants from Experiment 1 reside. None of the Itza' participants was trained in medicine.

Procedure Causal models were elicited in Spanish for the Itza' illnesses which most closely resemble depression and heart attack. The illnesses were *tukul* (meaning "thought" and glossed "pensiveness" in Itza', a wasting illness) and *derrame* (glossed as the verb "to spill" in Spanish; derrame cerebral is the Spanish gloss for "stroke").

Results

Itza' explain *tukul* as the result of separation from a family member which leads to dilution of the blood and rashes on the skin. 100% of Itza' participants attributed *tukul* to social causes (85% to separation from a family member), and 77% stated that social causes led to a change in the state of the blood (62% said the blood thinned, or was diluted, by too much thinking). 85% specified physical effects that result from the thinning of the blood, usually skin rashes (69%). Every Itza' participant explained *tukul* as the result of psycho-social factors leading to physical changes in the body.

Derrame is also seen as the result of an interaction of mental and physical causes. Itza' explain *derrame* as resulting from anger, which slows the blood, causing it to "spill" into the brain or nerves. 100% of Itza' participants attributed *derrame* to strong emotions (85% to anger), and 70% claimed that strong emotions cause the blood to change state in some way (e.g. blood stops circulating or gets cold), which causes it to mix inappropriately with some other substance of the body (85%), usually the brain or nerves (70%). Every Itza' participant explained *derrame* as the result of the deleterious effect of strong emotions on the state of the blood in the body.

Summary Itza concepts of tukul and derrame were structurally similar to energy healer concepts of depression and heart attack. Specifically, their conceptual representations of both illnesses included causal interactions in which psycho-social factors led to physical ones.

General Discussion

The current experiments show that, rather than being processed from within a single domain, illness knowledge is anchored in multiple domains. Illness models can be explained as specific combinations of domain knowledge. In Experiment 1 illness concepts of energy healers, RNs and Undergraduates showed systematically different causal patterns among mental and physical features. Specifically,

RNs and undergraduates rarely mentioned causal interaction between mental and physical causes. Energy healers, on the other hand, saw both illnesses as resulting from psycho-social causes which result in physical changes. Experiment 2 showed that the Itza' also view these illnesses as resulting from the physical effects of psycho-social factors.

Results from both studies clearly distinguish between the three alternatives presented above. Alternative Three, which was that mental and physical causes would be distributed in unsystematic ways in illness models, was ruled out. Evidence from RNs and undergraduates was also inconsistent with the first alternative, which proposed that illness models would be composed of causes from a single domain. RNs and undergraduates used psycho-social causes in explanations of depression, and physical causes in explanations of heart attack and depression. Thus, these participants utilized causes from distinct domains. However, for these participants a single causal chain consisted of only one type of cause. Data suggest that depression is construed both psychologically and physically for RNs and undergraduates, and heart attack is construed physically. Stress was the only psychological cause utilized in heart attack models of RNs and undergraduates. Stress may be a cause that is flexible, and can function as either psychological or physical. When RNs and undergraduates mentioned stress in the context of heart attack, they usually discussed physiological aspects of stress, such as increased adrenalin. However, because the nature of stress was ambiguous in the current study, it was coded as a psychological cause. For RNs and undergraduates, some illness models are constructed from within the cognitive domain of folk psychology, and others are constructed from within the domain of folk physics. For these participants, cultural factors influence whether an illness should be construed in psychological or physical terms. In this sense, Alternative Two is correct. Knowledge about illness is anchored in both domains.

Energy healer models are also consistent with Alternative Two. But in the case of energy healer models, single causal chains were composed of knowledge from different domains. For these participants causal chains are composed of psycho-social and physical causes. Further, when causal chains contain both kinds of causes, psycho-social causes are the distal causes and physical causes are proximate. The fact that domain boundaries are not preserved in causal chains of energy healers might be taken to suggest that domain knowledge does not constrain models of illness for energy healers. However, the systematicity in the models of energy healers is reflected in the uniformity with which individual participants combined psycho-social and physical causes. That is, the coherence of illness models across participants and across illnesses is precisely in their systematic use of causes from different domains. Participants cited different specific causal factors for heart attack and depression, but all participants were committed to the belief that some kind of psycho-social factor was the initial, distal cause of both illnesses and that physical, mechanical factors were the proximate cause. Analysis at the level of causal types reveals more agreement across individuals and illnesses than analysis at the level of specific, detailed causes.

Similarity in the causal models of energy healers and Maya is also revealed by analysis of patterns of causal types rather than analysis of overlap in specific causes. There was virtually no overlap in the specific causes cited by energy healers and Maya. In fact, it is not even clear that the illnesses being explained were conceptualized as precisely the same (biomedically defined) conditions across groups. But when analyzed as patterns of domain-specific causal types, it is clear that Maya and energy healers have similar causal models of illness. For both groups illnesses are caused by psycho-social factors which lead to proximate causes which are physical in nature. It would be unreasonable to expect that Maya, who have little or no formal education, would independently derive the same specific causes of illness as energy healers or undergraduates who live in a completely different cultural context. Analysis of illness models as patterns of domain-specific causal types reveals uniformities in thinking across cultures that are obscured by exclusive focus on specific, detailed causes.

In sum, cultural knowledge about illness may take the form of learning the culturally appropriate heuristics for combining domain specific knowledge to construct causal models of illness. Diversity among illness models reflects differences in the ways that different cultural groups combine information from different domains. For RNs and undergraduates, illness explanations are constructed from single causal types and culture specifies which type of cause is relevant to which illness. For energy healers and Maya, all illnesses are presumed to be caused by psycho-social factors which lead to physical changes. Thus, diversity in models of illness across cultures may be analogous to diversity in language across cultures where an overlapping set of categories, for example nouns and verbs, are combined in different ways.

Acknowledgments

The work in this paper would not have been possible without the generosity of the research participants, all of whom volunteered their time and wisdom. Thanks also to Douglas Medin, Lance Rips, Scott Atran, Michael Bailey, the Medin lab, and three anonymous reviewers for comments on earlier drafts. Work in this grant was supported by a National Science Foundation grant to Douglas Medin..

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition*. *69*(2), 135-178

Atran, S. (2001). The trouble with memes: Inference versus imitation in cultural creation. *Human Nature*, 12(4), 351-381.

Atran, S., Medin, D, & Ross, N. (in press). The Cultural Mind: Environmental decision-making and cultural modeling within and across populations. *Psychological Review*.

Carey, S. & Gelman, R. (Eds.). (1991). *The epigenesis of mind: Essays on Biology and Cognition*. Hillsdale, NJ: Lawrence Erlbaum

Choi, I, Nisbett, R, Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*. 125(1), 47-63.

Foster, G. M. (1976). Disease etiologies in non-Western medical systems. *American Anthropologist*, 78, 773-782.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? Journal of Personality & Social Psychology, 65(4,) 613-628.

Hirschfeld, L. A. & Gelman, S. A. (1994). *Mapping the Mind: Domain specificity in cognition and culture*. Cambridge: Cambridge University Press.

Kleinman, A. (1978). Concepts and a model for the comparison of medical systems as cultural systems. *Social Science & Medicine*, 12, 85-93.

Medin, D. & Atran, S. (in press). The Native Mind: Biological categorization, reasoning, and decision-making in development and across cultures. *Psychological Review*.

Miller, J., Bersoff, D., & Harwood, R. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality & Social Psychology*, 58(1), 33-47.

Murdock, G. P. (1980). *Theories of illness: A world survey*. Pittsburgh: University of Pittsburgh Press.

Pinker, S. (2002). *The Blank Slate: The modern denial of Human Nature*. New York: Viking Press.

Sperber, D. & Hirschfeld, L. (2004). The cognitive foundations of cultural diversity and stability. *Trends in Cognitive Sciences*, 8(1), 40-46.

Modeling Forms of Surprise in Artificial Agents: Empirical and Theoretical Study of Surprise Functions

Luís Macedo (lmacedo@isec.pt)
Department of Informatics and Systems Engineering, Engineering Institute, Polytechnic Institute of Coimbra, Quinta da Nora
3030-199 Coimbra, Portugal
Centre for Informatics and Systems of the University of Coimbra, Department of Informatics, Polo II
3030 Coimbra, Portugal

Rainer Reisenzein (rainer.reisenzein@uni-greifswald.de)
Institute for Psychology, University of Greifswald, Department of General Psychology II, Franz-Mehringstr, 47
17487 Greifswald, Germany

Amílcar Cardoso (amilcar@dei.uc.pt)
Centre for Informatics and Systems of the University of Coimbra, Department of Informatics, Polo II
3030 Coimbra, Portugal

Abstract

This paper addresses the issue of how to compute the intensity of surprise in an artificial agent. Resolution of this issue is important for the further specification of the computational model of surprise proposed by Macedo and Cardoso (2001) that was implemented in artificial agents "living" in a multi-agent environment. This model of surprise is mainly rooted in the cognitive-psychoevolutionary model of surprise proposed by the research group of the University of Bielefeld (Meyer, Reisenzein, & Schützwohl, 1997) and in proposals by Ortony and Partridge. We propose several possible functions to compute the intensity of surprise. To assess their accuracy, they were evaluated in an experimental test that focused on the comparison of surprise intensity values generated by artificial agents with ratings by humans under similar circumstances.

Introduction

Considered by many authors a biologically fundamental emotion (e.g.: Ekman, 1992; Izard, 1991), surprise may play an important role in the cognitive activities of intelligent agents, especially in attention focusing (Izard, 1991; Meyer et al., 1997; Ortony & Partridge, 1987; Reisenzein, 2000b), learning (Schank, 1986) and creativity (Boden, 1995; Williams, 1996). Psychological experiments conducted by Meyer, Reisenzein and Schützwohl provide evidence that surprising-eliciting events initiate a series of mental processes that (a) begin with the appraisal of a cognized event as exceeding some threshold value of unexpectedness or schema discrepancy, (b) continue with the interruption of ongoing information processing and the reallocation of processing resources to the surprise-eliciting event, and (c) culminate in the analysis and evaluation of that event plus immediate reactions to it and/or schema (belief) updating/revision. According to these authors, surprise has two main functions, the one informational and the other motivational: it informs the individual about the occurrence of a schema-discrepancy, and it provides an initial impetus for the exploration of the unexpected event. Thereby, surprise promotes both immediate adaptive actions to the unexpected event and the prediction, control and effective dealings with future occurrences of the event.

Ortony and Partridge's (1987) model of surprise shares several aspects with the one proposed by Meyer, Reisenzein and Schützwohl (1997), especially in that both models assume that surprise is elicited by unexpected events. The same is also true for Peters' (1998) computational model of surprise, implemented in a computer vision system, that focuses on the detection of unexpected movements. Finally, models of surprise have also been proposed in the fields of knowledge discovery and data mining (e.g. Suzuki & Kodratoff, 1998).

Macedo and Cardoso (e.g., Macedo & Cardoso, 2001)) developed a computational model of surprise that is an adaptation (although with several simplifications) of the models proposed by Meyer, Reisenzein and Schützwohl (1997) and by Ortony and Partridge (1987). In the present article, we elaborate and evaluate this model further by discussing different possible functions for the computation of surprise and by evaluating these functions in an empirical study.

The following section describes Macedo and Cardoso's surprise model in more detail, including an overview of its theoretical background models. Subsequently, we discuss several possible functions for computing the intensity of surprise. Finally, we describe an experimental test that was carried out to evaluate the accuracy of these surprise functions.

Surprise Model

As mentioned, the surprise model developed by Macedo and Cardoso (2001) is mainly based on Ortony and Partridge's (1987) proposals and on those of Meyer, Reisenzein and Schützwohl (1997). Therefore, we first give an overview of these background theories and then explain the computational model proposed by Macedo and Cardoso, by comparing it with these two models.

Background Models

Although Ortony and Partridge agree with Meyer, Reisenzein and Schützwohl and other authors that surprise is caused by events that are commonsensically called unexpected, they proposed that unexpectedness covers two cases. First, surprise results when prior expectations regarding an event are disconfirmed. Second, however, surprise can also be caused by events for which expectations were never computed. That is, according to Ortony and Partridge, there are situations in which one is surprised although one had no explicit expectations (either conscious or unconscious) regarding the surprising event. Ortony and Partridge also proposed that surprisingness is an important variable in artificial intelligence systems, particularly in attention and learning.

In more detail, Ortony and Partridge's model of surprise assumes a system (or agent) with an episodic and semantic propositional memory whose elements may be immutable (propositions that are believed to be always true) or typical (propositions that are believed to be usually but not always true). Furthermore, they distinguish between practically deducible propositions and practically non-deducible propositions. *Practically deducible propositions* comprise all propositions that are explicitly represented in memory, as well as those that can be inferred from these by few and simple deductions. Hence, practically deducible propositions are that subset of formally deducible propositions that don't require many and complex inferences. Furthermore, practically deducible propositions may be either actively or passively deduced. In the former case, their content corresponds to *actively expected* or *predicted* events; in the latter case, to *passively expected* (*assumed*) events.

Based on these assumptions, Ortony and Partridge proposed that surprise results when the system encounters a conflict or inconsistency between an input proposition and preexisting representations or representations computed "after the fact". More precisely, surprise results in three situations (Table 1 presents the corresponding range of values): (i) *active expectation failure*: here, surprise results from a conflict or inconsistency between the input proposition and an *active prediction* or *expectation*; (ii) *passive expectation failure* (or *assumption failure*): here, surprise results from a conflict or inconsistency between the input proposition and what the agent implicitly knows or believes (*passive expectations* or *assumptions*); and (iii) *unanticipated incongruities* or deviations from norms: here, surprise results from a conflict or inconsistency between the input proposition (which in this case is a practically non-deducible proposition) and what, after the fact, is judged as normal or usual (Kahneman & Miller, 1986), that is, between the input proposition and practically deducible propositions (immutable or typical) that are suggested by the unexpected fact. Note that, in this case, prior to the unexpected event there are no explicit expectations (passive or active) with which the input proposition could conflict.

In their cognitive-psychoevolutionary model, Meyer, Reisenzein and Schützwohl also assume that surprise (considered by them as an emotion) is elicited by the appraisal of unexpectedness.

Table 1: Three different sources of surprise and corresponding value ranges (adapted from (Ortony & Partridge, 1987)).

Confronted proposition	Related Cognition	
	Active	Passive
Immutable	[1]; $S_A=1$; *Prediction*	[2]; $S_P=1$; *Assumption*
Typical	[3]; $0< S_A<1$; *Prediction*	[4]; $S_P<S_A$; *Assumption*
Immutable	[5]; ∅	[6]; $S_P=1$; *none*
Typical	[7]; ∅	[8]; $0< S_P<1$; *none*

More precisely, it is proposed that surprise-eliciting events give rise to the following series of mental processes: (i) the appraisal of a cognized event as exceeding some threshold value of unexpectedness (schema-discrepancy) - according to Reisenzein (2001), this is achieved by a specialized comparator mechanism, the unexpectedness function, that computes the degree of discrepancy between "new" and "old" beliefs or schemas; (ii) interruption of ongoing information processing and reallocation of processing resources to the investigation of the unexpected event; (iii) analysis/evaluation of that event; and (iv) possibly, immediate reactions to that event and/or updating or revision of the "old" schemas or beliefs.

Overview of the Computational Model of Surprise

Macedo and Cardoso (e.g., Macedo & Cardoso, 2001) developed a multi-agent environment in which, in addition to inanimate agents (objects such as buildings), there are two main kinds of animate, interacting agents: the "author-agents" or creators, whose main function is to create things (objects, events), and the "jury-agents" or explorers whose goal is to explore the environment by analyzing, studying and evaluating it. An agent can also show both of these activities (creation and exploration).

The computational model of surprise is integrated into the *motivations* module of the architecture of the artificial agents (see Figure 1). The other modules of this architecture are: *sensors/ perception*; *memory*; *goals/desires*; and *reasoning/decision-making*. This last module and the module *motivations* are provided with information from the world obtained through *sensors/perception*, as well as with information recorded in *memory*. The *reasoning/decision-making* module then computes the current state of the world. Afterwards, probability theory is applied to predict possible future states of the world for the available actions, and a utility function (which makes use of the intensity of the generated emotions) is applied to each of these world states. Finally, the action that maximizes the utility function is selected.

The computational model of surprise incorporated in this agent system is an adaptation (although with some simplifications) of the surprise model proposed by Meyer, Reisenzein and Schützwohl in which the above-mentioned four mental processes elicited by surprising events are

present. The suggestions by Ortony and Partridge are mainly concerned with the first of these steps, and are compatible with the Meyer, Reisenzein and Schützwohl model. Accordingly, in our model, we drew on the assumptions of Ortony and Partridge for the implementation of the appraisal of unexpectedness and the computation of the intensity of surprise, as well as for the selection of knowledge structures.

In Macedo and Cardoso's model, knowledge is exclusively of an episodic kind (for an example, see Figure 2), rather than being both semantic and episodic in nature (although this will be considered in future work), as in Ortony and Partridge's model. In this respect, the knowledge structure of our model also differs from the schema-theoretic framework of the Meyer, Reisenzein and Schützwohl model that also assumes both episodic and semantic knowledge. In our model, an input proposition (or new belief) is therefore always compared with episodic representations of objects or events (or their properties) (for instance an object with squared windows, rectangular door, etc.). Besides, the agent has in its episodic memory explicit representations of similar objects. Following Ortony and Partridge, we also distinguish between *deducible* and *non-deducible*, *active* and *passive*, *immutable* and *typical* propositions as well as between different possible sources of surprise (see Table 1). The immutability of a proposition can be extracted from the absolute frequency values associated with the cases stored in episodic memory (see Figure 2). For instance, in the example shown in Figure 2, the proposition "houses have square facades" is immutable (since all the houses in memory have squared facades), whereas "houses have square windows" is a typical proposition with a probability (immutability) value of 0.50 (as implied by Ortony and Partridge's model, in our model immutability is a continuous variable).

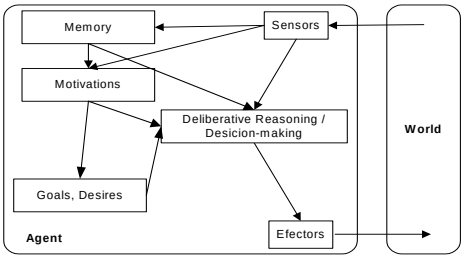


Figure 1: Architecture of an agent.

Field \ Case	C_1	C_2	C_3	C_4
Structure				
Function	House	House	Church	Hotel
Behavior	Static	Static	Static	Static
Abs. Freq.	50	40	5	5

Figure 2: Example of an episodic memory in the domain of buildings.

The usual activity of the agents consists of moving through the environment hoping to find interesting things (objects or events) that deserve to be investigated. We assume that this exploratory behavior is ultimately in the service of other (e.g., hedonic) motives, although this issue is not explicitly addressed in the present model. When one or more objects/events are perceived, the agent computes expectations for the missing information (e.g., "it is a house with 67% of probability", "it is a hotel with 45% of probability", etc.; note that the *function* of a building becomes available to the agent only when its position and that of the building are the same). On the basis of the available information (e.g., the visible *structure* of an object) and the computed expectations (e.g., predictions for the *function* of an object), the agent then determines the intensity of surprise that may be caused by the object/event (these computations, which correspond to the "appraisal of unexpectedness" in the Meyer, Reisenzein and Schützwohl model, are described in more detail below). Subsequently, the object/event with the maximum estimated surprise is selected to be visited and investigated. This corresponds to the "interruption of ongoing activity" and the "reallocation of processing resources" assumed in the Meyer, Reisenzein and Schützwohl model. The previously estimated value of surprise may subsequently be updated on the basis of the additional information acquired about the object/event. The object/event is then stored in memory and the absolute frequencies of the affected objects/events in memory are updated. This is a simplification of the fourth step of the Meyer, Reisenzein and Schützwohl model (for alternative approaches to belief revision, see, for instance, (Gärdenfors, 1988)).

The different surprise-eliciting situations distinguished by Ortony and Partridge are dealt with in our model in the following way. As said above, when an agent perceives an object, it first computes expectations (*deducible*, *active expectations*) for missing information (e.g., "it is a hotel with 45% of probability"). If, after having visited that object, the agent detects that the object is different from what was expected (e.g., if it is a post office), the agent is surprised because its *active expectations* conflict with the input proposition (note that, in our model, belief conflicts may be partial as well as total). This is thus an example of the first source of surprise distinguished by Ortony and Partridge. In contrast, when an agent perceives an aspect or part of an object with particular properties (e.g., a building with a window of a circular shape) that were not actively predicted, it may still be able to infer that it expected something (e.g., a rectangular-shaped window with, 45% probability, a square-shaped window with 67%, etc.). This is an example of a *deducible*, *passive expectation*: although the expectation was not present before the agent perceived the object, it was inferred after the object had been perceived. This case is therefore an example of the second source of surprise distinguished by Ortony and Partridge, where an input proposition conflicts with an agent's *passive expectations*. Finally, when an agent perceives an object with a completely new part (e.g., a building with no facade), it has neither an active nor a passive expectation available.

The reason is that, because there are no objects of this kind (e.g., buildings with no facade) stored in the agent's memory, the agent cannot predict that such objects might be encountered. The perception of an object with a completely new part is thus an example of a *non-deducible proposition*. This is an example of the third source of surprise distinguished by Ortony and Partridge: there is a conflict between the input proposition (e.g., "the house has no facade") and what *after the fact* is judged to be normal or usual (e.g., "buildings have a facade").

The Computation of Surprise Intensity

We now address the question of how the intensity of surprise should be computed in the model. In humans, this problem has already been successfully solved by evolution; therefore, a reasonable approach is to model the agent's surprise function according to that of humans. Experimental evidence from human participants summarized in (Reisenzein, 2000b) suggests that the intensity of felt surprise increases monotonically, and is closely correlated with, the degree of unexpectedness. On the basis of this evidence, we propose that the surprise "felt" by an agent elicited by an object/event X is proportional to the degree of unexpectedness of X (which in the model is based on the frequencies of objects/events present in the memory of the agent). According to probability theory, the degree of expecting an event X to occur is its subjective probability P(X). Accordingly, the improbability of X, denoted by 1-P(X), defines the degree of not expecting X, or for short its unexpectedness. The intensity of surprise elicited by X should therefore be an (at least weakly) monotonically increasing function of 1-P(X). As a first approach, this function (S1) could simply be taken to be the identity function, that is, the intensity of surprise could simply be equated with the degree of unexpectedness:

$$S1(Agt, X) = 1 - P(X)$$

However, on second thought, S1 does not seem to faithfully capture the relation between unexpectedness and surprise. For example, consider a political election with three candidates A, B and C, where the probability of being elected is P(A) = P(B) = P(C) = 0.333. In this case, one would not be surprised if either A, B or C is elected. Therefore, in this situation at least, S1 fails.

To arrive at a more adequate surprise function, consider the case where there are only two mutually exclusive and exhaustive alternative events, X and Y (i.e., not X). Here, intuition suggests that X is not surprising as long as P(X) ≥ 0.5, whereas X is surprising for P(X) < 0.5, and increasingly more so the more P(X) approaches 0. This intuition is captured by the following surprise function (S2):

$$S2(Agt, X) = \begin{cases} 1 - P(X) \Leftarrow P(X) < 0.5 \\ 0 \Leftarrow P(X) \geq 0.5 \end{cases}$$

To deal with sets of more than two mutually exclusive events, S2 could be generalized as follows (S3, where *n* denotes the number of events in the set):

$$S3(Agt, X) = \begin{cases} 1 - P(X) \Leftarrow P(X) < \frac{1}{n} \\ 0 \Leftarrow P(X) \geq \frac{1}{n} \end{cases}$$

However, it may be more adequate to set the upper limit of surprise not to 1, but to $\frac{1}{n}$ (see S4):

$$S4(Agt, X) = \begin{cases} \frac{1}{n} - P(X) \Leftarrow P(X) < \frac{1}{n} \\ 0 \Leftarrow P(X) \geq \frac{1}{n} \end{cases}$$

Yet another possible surprise function, suggested by further reflection on the above election example, is the following (S5):

$$S5(Agt, X) = P(Y) - P(X)$$

In this formula, Y is the event with the *highest* probability of a set of mutually exclusive events. S5 implies that, within each set of mutually exclusive events, there is always one (Y) whose occurrence is entirely unsurprising, namely the event with the maximum probability in the set (P(Y)). For the other events X in the set, the surprise intensity caused by their occurrence is the difference between P(Y) and their probability P(X). This difference can be interpreted as the amount by which P(X) has to be increased for X to become unsurprising. For instance, in the election example considered earlier, where P(A) = P(B)= P(C) = 0.333, S5 correctly predicts that one would not be surprised if either A, B or C is elected. By contrast, if P(A) = 0.55, P(B) = 0.40 and P(C) = 0.05, S5 predicts that the surprise caused by B is 0.15 and for C is 0.50, whereas for A it is 0. S5 also implies that maximum surprise, that is, S(X) = 1, occurs only if P(Y) = 1 and hence, by implication, P(X) = 0. (In the Ortony and Partridge model, this corresponds to situations [1], [2], [5] and [6], where the disconfirmed event Y is immutable, i.e., its probability is 1). Therefore, S5 seems to correctly describe surprise in the election example. Confirming this impression, S5 also acknowledges the intuition behind S2: if there are only two alternative events X and Y (= not X), S5 predicts, like S2, that X should be unsurprising for P(X) ≥ 0.5, for in this case X is also the event with the highest probability in the set. By contrast, for P(X) < 0.5, S5 predicts that X should be surprising and increasingly so the more P(X) approaches 0, with maximum possible surprise (S(X) = 1) being experienced for P(X) = 0.

Yet another possible surprise function (S6) is suggested by Information Theory (Shannon, 1948):

$$S6(Agt, X) = \log_2 \frac{1}{P(X)}$$

According to S6, surprise about X is 0 when P(X) = 1 and increases monotonically with decreasing P(X). In these respects, then, S6 is similar to S1. However, in contrast to S1, S6 is a nonlinear function of P(X), and it is not normalized. For instance, for P(X) = 0.3, *S6(X)* = 1.7 (bits), for P(X) = 0.01, *S6(X)* = 6.6, and for P(X) = 0.001, *S6(X)* = 9.9. In fact, there is no upper limit of S(X): for P(X)=0, *S6(X)* = $+\infty$. To overcome this problem, we propose the following normalized function S7 (stipulating the upper limit to be 10):

$$S7(Agt, X) = \frac{\log_2 \frac{1}{P(X)}}{10}$$

Finally, yet another surprise function (S8), a nonlinear modification of S5, is suggested by the results of the experiment, reported below, performed with humans in the domain of elections and sport games:

$$S8(Agt, X) = \log_2 (1 + P(Y) - P(X))$$

This function retains the essential features of S5: when X is the most expected event (X = Y), then S8(X) = 0; when X is different from Y, S8(X) > 0 and increases monotonically with the difference between P(Y) and P(X); and S8(X) is maximal (= 1) if P(Y) = 1 and P(X) = 0. In addition, however, S8 also captures the nonlinearity of the surprise function suggested by the experiments with humans reported below.

Experiment

To test the validity of the proposed surprise functions, we conducted an experiment that involved two steps. In step 1, we collected ratings of probability and surprise intensity from humans in two domains, political elections and sports games. In step 2, artificial agents that implemented the different surprise functions were provided with the probability judgments obtained from the humans and, on this basis, computed surprise intensity values. These predicted surprise values were then compared with the actual surprise ratings provided by the human participants.

Step 1 was conducted with ten participants (mean age, 29 years). They were presented with 20 brief scenarios, 10 of which described political elections with 2-4 candidates (see Figure 3), whereas the other 10 scenarios described sports games with 2-4 teams or players (see (Reisenzein, 2000a) for a conceptually similar experiment using knowledge questions). Political elections and sports games were chosen because we thought that these domains are familiar to most people and that the participants would have no problems to state their probabilities and their surprise about outcomes. In addition, in contrast to the domain of buildings used in a previous study reported in (Macedo & Cardoso, 2001), elections and sport games allow for an easier matching of the knowledge of artificial agents with that of humans. Part of the scenarios did not include information about the actual outcome of the election or game, whereas the remaining scenarios included this information. For scenarios without outcome information, the participants were asked to first state their expectations for all possible outcomes and to rate their probability on a 1-100 scale. Subsequently, they were informed about the outcome of the election or game and rated their surprise about the outcome first on a qualitative intensity scale and then again on a quantitative intensity scale within the chosen qualitative level. By contrast, for the scenarios that included outcome information, participants first rated the intensity of surprise about the outcome and subsequently their (passive) expectations regarding the outcome. An example of a scenario is shown in Figure 3.

> Given the following prognosis for the election of candidate A, B and C for a political position:
>
> Victory of A=45%; Victory of B=45%; Victory of C=10%
>
> a) What are your personal expectations regarding the victory of candidates A, B and C?
> b) Assume that candidate A won the election and rate the intensity of surprise that you would feel.

Figure 3: Example of a test item.

In step 2 of the study, the probability ratings obtained from each participant in step 1 were delivered to eight artificial agents, each of which implemented one of the eight surprise functions S1-S8 described earlier. Using these functions, the agents computed surprise intensity values from the probabilities. These predicted surprise values were then compared with the surprise ratings of the humans obtained in step 1.

The data obtained in the first step of the experiment suggested two qualitative conclusions. First, the occurrence of the most expected event of the set of mutually exclusive and exhaustive events did not elicit surprise in humans. For example, when the expectations for the election of three political candidates A, B and C were P(A) = 0.55, P(B) = 0.40, and P(C) = 0.05, the participants felt no surprise about the election of candidate A. This was also true when two or more candidates had equal maximal probabilities. For example, when P(A) = 0.40, P(B) = 0.40 and P(C) = 0.20, participants were not surprised when either A or B was elected. Second, beyond the point of zero surprise, the surprise function appeared to be nonlinear. For example, relatively high surprise was indicated when candidate C won the elections in both of the above situations, although it was still higher for P(C) = 0.05 than for P(C) = 0.20.

To compare the surprise values generated by the artificial agents and the surprise ratings provided by the human judges, the following fit indices were used: the root mean squared difference, the mean absolute difference, and the Pearson correlation. The results of these comparisons are shown in Table 2, separately for the 10 participants (H1, …, H10) and for six of the eight artificial agents (A1,…,A8) (the surprise functions S6 and S7 were not included because

they have a different range than the human ratings and therefore computation of the absolute and squared differences is not meaningful). It can be seen from Table 2 that, regardless of which fit index is used, agent A8 (which implemented surprise function S8) was the one with the best fit to the human ratings: it had on average, the lowest root mean squared differences (*Ms*= 0.10), the lowest absolute differences (*Md*= 0.06), and the highest correlation to these ratings (*Mr*= 0.98). A8 was closely followed by A5 (*Ms* = 0.21; *Md* = 0.08; *Mr* = 0.97), whereas agents A1 and A2 had the comparatively worst fit values (for instance, A1 had *Ms* = 0.35; *Md* = 0.26; *Mr* = 0.81). A main reason for the bad performance of A1 was apparently that it failed in the case of the occurrence of the most expected event of the set: A1 still predicts a positive surprise value (1-P(X)) for this case, whereas humans do not feel surprised by the occurrence of this event. However, in other situations, A1 performed well.

Table 2: Statistical comparison of the surprise values computed by the artificial agents and those provided by the humans (*s* = root mean squared difference, *d* = mean absolute difference, and *r* = Pearson correlation).

		H1	*H2*	*H3*	*H4*	*H5*	*H6*	*H7*	*H8*	*H9*	*H10*	*M*
A1	*s*	.35	.36	.34	.35	.35	.34	.35	.36	.35	.36	**.35**
	d	.25	.26	.25	.25	.26	.24	.27	.27	.26	.27	**.26**
	r	.82	.80	.82	.82	.80	.82	.81	.80	.82	.82	**.81**
A2	*s*	.30	.33	.29	.32	.32	.30	.33	.32	.31	.31	**.31**
	d	.18	.21	.16	.20	.21	.18	.22	.19	.19	.19	**.19**
	r	.82	.79	.82	.81	.79	.83	.80	.80	.81	.81	**.81**
A3	*s*	.22	.30	.24	.21	.30	.22	.18	.19	.19	.16	**.22**
	d	.07	.15	.09	.07	.17	.09	.09	.09	.08	.08	**.10**
	r	.95	.85	.89	.94	.81	.92	.93	.92	.92	.94	**.91**
A4	*s*	.43	.41	.45	.43	.43	.43	.44	.46	.46	.45	**.44**
	d	.29	.28	.30	.29	.29	.28	.28	.28	.29	.27	**.28**
	r	.93	.92	.88	.96	.90	.95	.91	.91	.93	.94	**.92**
A5	*s*	.22	.16	.19	.16	.23	.20	.21	.24	.24	.24	**.21**
	d	.07	.06	.11	.06	.09	.05	.08	.10	.09	.09	**.08**
	r	.97	.98	.96	.98	.95	.99	.97	.96	.96	.96	**.97**
A8	*s*	.09	.07	.13	.08	.12	.06	.11	.13	.12	.12	**.10**
	d	.05	.05	.09	.05	.08	.04	.06	.08	.07	.07	**.06**
	r	.98	.99	.98	.99	.97	.99	.98	.07	.07	.97	**.98**

Conclusions

The empirical study of the surprise functions suggests S8(X) = $\log_2(1+P(Y)-P(X))$ as the most appropriate surprise function for the domains of political elections and sport games, although S5 (the linear counterpart of S8) is a very close contender. However, before more definitive conclusions can be drawn, additional tests need to be performed in other domains, as well as with yet other possible surprise functions (e.g., Shackle, 1969).

Acknowledgments

We would like to thank A. Ortony for his helpful comments. The PhD of Luís Macedo is financially supported by PRODEP III.

References

Boden, M. (1995). Creativity and unpredictability. SEHR, 4(2).

Ekman, P. (1992). An Argument for Basic Emotions. In N. L. Stein & K. Oatley (Eds.), Basic Emotions (pp. 169-200). Hove, UK: Lawrence Erlbaum.

Gärdenfors, P. (1988). Knowledge in flux: Modeling the dynamics of epistemic states. Cambridge, MA: Bradford Books.

Izard, C. (1991). The Psychology of Emotions. NY: Plenum Press.

Kahneman, D., & Miller, D. (1986). Norm theory: comparing reality to its alternatives. Psychological Review, 93, 136-153.

Macedo, L., & Cardoso, A. (2001). Modelling Forms of Surprise in an Artificial Agent. In J. Moore & K. Stenning (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science Society (pp. 588-593). Mahwah, NJ: Erlbaum.

Meyer, W., Reisenzein, R., & Schützwohl, A. (1997). Towards a process analysis of emotions: The case of surprise. Motivation and Emotion, 21, 251-274.

Ortony, A., & Partridge, D. (1987). Surprisingness and Expectation Failure: What's the Difference?, Proceedings of the 10th International Joint Conference on Artificial Intelligence (pp. 106-108). Los Altos, CA: Morgan Kaufmann.

Peters, M. (1998). Towards Artificial Forms of Intelligence, Creativity, and Surprise, Proceedings of the Twentieth Annual Conference of the Cognitive Science Society (pp. 836-841). Mahwah, NJ: Erlbaum.

Reisenzein, R. (2000a). Exploring the strength of association between the components of emotion syndromes: The case of surprise. Cognition and Emotion, 14, 1-38.

Reisenzein, R. (2000b). The subjective experience of surprise. In H. Bless & J. Forgas (Eds.), The message within: The role of subjective experience in social cognition and behavior. Philadelphia, PA: Psychology Press.

Reisenzein, R. (2001). Appraisal processes conceptualized from a schema-theoretic perspective: Contributions to a process analysis of emotions. In K. Scherer & A. Schorr & T. Johnstone (Eds.), Appraisal processes in emotion: Theory, Methods, Research (pp. 187-201). Oxford: Oxford University Press.

Schank, R. (1986). Explanation Patterns: Understanding Mechanicaly and Creatively. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shackle, G. (1969). Decision, Order and Time in Human Affairs (2 ed.). Cambridge, UK: Cambridge University Press.

Shannon, C. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379-423;623-656.

Suzuki, E., & Kodratoff, Y. (1998). Discovery of Surprising Exception Rules Based on Intensity of Implication. In J. Zytkow & M. Quafafou (Eds.), Proceedings of Second European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD '98 (pp. 10-18). Berlin: Springer.

Williams, M. (1996). Aesthetics and the explication of surprise. Languages of Design, 3, 145-157.

Creative Abduction as Active Shaping of Knowledge. Epistemic and Ethical Mediators

Lorenzo Magnani (lmagnani@unipv.it)
Department of Philosophy and Computational Philosophy Laboratory, Piazza Botta 6
27100 Pavia, Italy, and
Department of Philosophy, Baruch College, The City University of New York,
New York, NY, 10010 USA

Abstract

The concept of *manipulative abduction* is devoted to capturing the role of action in many interesting situations: action provides otherwise unavailable information that enables the agent to solve problems by starting and performing a suitable abductive process of generation or selection of hypotheses. Many *external representations*, even if in some cases inert from an epistemological point of view, can be transformed into what is called *epistemic mediators*, active in creative abductive reasoning. An often neglected side of human creativity is related to emotional, artistic, and ethical aspects, and concerns the active shaping of values in an esthetical and ethical world. I will present some aspects of this kind of reasoning in the case of scientific and ethical thinking; moreover, I will illustrate some aspects of what I call "ethical mediators" in their activity of shaping and reshaping ethical worth of human beings and collectives.

The Inexplicability of Creativity

Creativity is certainly an important aspect of our definition of "intelligence" but the literature associates many different notions to creativity. This ambiguity has brought to a lack of consensus in the research community. The common views associate to creativity unusual and mysterious qualities that drive the concept of creativity to a confused verbosity. Statements like "to break the rules", "to think different", "to destroy one *Gestalt* in favor of a better one", and "to arrange old elements into a new form", present in the field of psychological research on creativity since 1950s, certainly do not clarify the topic, and seem to lead to the Freudian conclusion that creativity cannot be understood. This conclusion has also been supported by many philosophers who studied conceptual change in science during the second half of the last century. They distinguished between a logic of discovery and a logic of justification (i.e. between the psychological side of creation and the logic argument of proving new discovered ideas by facts). The consequent conclusion was that a logic of discovery (and a *rational* model of discovery) could not exist: scientific conceptual change is cataclysmic and irrational, dramatic, incomprehensible and discontinuous. Many other studies already argued that creativity can be understood (Boden, 1991, Sternberg, Kaufman, and Pretz, 2002), but paid attention mainly to the psychological and experimental aspects, disregarding the philosophical, logical, and computation ones.

In AI research, however, since Simon, two characteristics seem to be associated to creativity: the *novelty* of the product and the *unconventionality* of the process that leads to the new product. Hence, in a strictly *pragmatic* sense, when we can clarify what behavior we are looking for, we could implement it in a machine: a methodological criterion enables us to define and consider just those practical effects we conceive to be associated with novelty and unconventionality (cf. Buchanan, 2001).

I maintain we can overcome many of the difficulties of creativity studies developing a theory of abduction, in the light of Charles Sanders Peirce's first insights.

Abduction and Epistemic Mediators

If we decide to adopt this kind of methodology it is necessary to develop a cognitive model of creativity able to represent not only "novelty" and "unconventionality", but also some features commonly referred to as the entire creative process, such as the expert use of background knowledge and ontology (defining new concepts and searching heuristically among the old ones) and the modeling activity developed in the so called "incubation time" (generating and testing, transformations in the space of the hypotheses). The philosophical concept of *abduction* may be a candidate to solve this problem, and offers an approach to model creative processes of hypotheses generation in a completely explicit and formal way, which can fruitfully integrate the narrowness proper of a merely psychological approach, too experimentally human-oriented.

Theoretical and Manipulative Abduction

A hundred years ago, C. S. Peirce (*CP*, 1931-1958) coined the concept of abduction in order to illustrate that the process of scientific discovery is not irrational and that a methodology of discovery is possible. Peirce interpreted abduction essentially as an "inferential" *creative process* of generating a new hypothesis. Abduction has a logical form (fallacious, if we model abduction by using classical logic) distinct from deduction and induction. Reasoning which starts from reasons and looks for consequences is called *deduction*; that which starts from consequences and looks for reasons is called *abduction*.

Abduction is the process of *inferring* certain facts and/or laws and hypotheses that render some sentences plausible, that *explain* or *discover* some (eventually new) phenomenon or observation; it is the process of reasoning in which explanatory hypotheses are formed and evaluated. There are two main epistemological meanings of the word abduction (Magnani, 2001): 1) abduction that only generates "plausible" hypotheses ("selective" or "creative") and 2) abduction considered as inference "to the best explanation", which

also evaluates hypotheses. To illustrate from the field of medical knowledge, the discovery of a new disease and the manifestations it causes can be considered as the result of a creative abductive inference. Therefore, "creative" abduction deals with the whole field of the growth of scientific knowledge. This is irrelevant in medical *diagnosis* where instead the task is to "select" from an encyclopedia of prestored diagnostic entities. We can call both inferences ampliative, selective and creative, because in both cases the reasoning involved amplifies, or goes beyond, the information incorporated in the premises.

Theoretical abduction[1] certainly illustrates much of what is important in creative abductive reasoning, in humans and in computational programs, but fails to account for many cases of explanations occurring in science when the exploitation of environment is crucial. It fails to account for those cases in which there is a kind of "discovering through doing", cases in which new and still unexpressed information is codified by means of manipulations of some external objects (*epistemic mediators*). The concept of *manipulative abduction*[2] captures a large part of scientists' and physicians' thinking where the role of action is central, and where the features of this action are implicit and hard to be elicited

Peirce uses the terms "inference" and "inferential process" to refer to abduction. It is useful to try to clarify the meaning of the term "inference" as considered by Peirce's thought. Peirce stated that all thinking is in signs, and signs can be icons, indices or symbols. Moreover, all *inference* is a form of sign activity, where the word sign includes "feeling, image, conception and other representation" (*CP* 5.283), and, in Kantian words, all synthetic forms of cognition. Feelings, images, simulations, etc., are currently characterized as forms of model-based reasoning (Magnani & Nersessian, 2002). Consequently, following Peirce, we can say that a considerable part of thinking activity is *model-based* (cf. footnote 1), that most of the forms of constitution of phenomena are characterized in a model-based way. I use the term "model-based reasoning" following Nersessian (1995), that is, to indicate the construction and manipulation of various kinds of representations, not necessarily sentential and/or formal. Scientific concept formation, scientific discovery, and – as we will see – diagnostic reasoning are often related to *heuristic* procedures that resort to mental/internal but also to external "models" and representations.

Peirce gives an interesting example of model-based abduction related to sense activity: "A man can distinguish different textures of cloth by feeling: but not immediately, for he requires to move fingers over the cloth, which shows that he is obliged to compare sensations of one instant with those of another" (*CP* 5.221); this idea surely suggests that abductive movements also have interesting extra-theoretical characteristics and that there is a role in abductive reasoning for various kinds of manipulations of external objects (cf. below, the problem of "action-based, manipulative abduction"). One more example is given by the fact that the perception of tone arises from the activity of the mind only after having noted the rapidity of the vibrations of the sound waves, but the possibility of individuating a tone happens only after having heard several of the sound impulses and after having judged their frequency. Consequently the sensation of pitch is made possible by previous experiences and cognitions stored in memory, so that one oscillation of the air would not produce a tone.

Model-based thinking activity also exploits *external models*. We have seen that the concept of manipulative abduction is devoted to capturing the role of action on external models in hypothetical and creative reasoning. This kind of manipulation provides otherwise unavailable information that enables the agent to solve a problem by performing abductive processes of generation or selection of hypotheses. An expert manipulation of objects directed by abductive movements that implicates the strategic application of old and new *templates* of behavior mainly connected with extra-theoretical components also esthetical, ethical, and emotional.

Manipulative abduction happens when we are thinking *through* doing and not only, in a pragmatic sense, about doing. It refers to an extra-theoretical behavior that aims at creating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices. Gooding (1990) refers to this kind of concrete manipulative reasoning when he illustrates the role in science of the so-called "construals" that embody tacit inferences in procedures that are often apparatus and machine based. The embodiment is of course an expert manipulation of objects in a highly constrained experimental environment, and is directed by abductive movements that imply the strategic application of old and new *templates* of behavior mainly connected with extra-theoretical components, for instance emotional, esthetical, ethical, and economic.

Epistemic Mediators

Recent research, taking an ecological approach to the analysis and design of human-machine systems, has shown how expert performers use action in everyday life to create an "external" model of task dynamics that can be used in lieu of an internal model (Kirlik, 1998). Not only a way for moving the world to desirable states, action performs an *epistemic* and not merely performatory role that is very relevant to abductive reasoning.

The whole activity of manipulation is devoted to build various external *epistemic mediators* that function as an enormous new source of information and knowledge. I derive this expression from the cognitive anthropologist Hutchins (1995), that coins the expression "mediating structure" to refer to various external tools that can be built to cognitively help the activity of navigating in modern but also in "primitive" settings. Any written procedure is a simple example of a cognitive "mediating structure" with possible cognitive aims: "Language, cultural knowledge, mental

[1] Magnani (2001) introduces the concept of theoretical abduction. He maintains that there are two kinds of theoretical abduction, "sentential", related to logic and to verbal/symbolic inferences, and "model-based", related to the exploitation of internalized models of diagrams, pictures, etc., cf. below in this paper.

[2] Manipulative abduction and epistemic mediators are introduced and illustrated in Magnani (2001).

models, arithmetic procedures, and rules of logic are all mediating structures too. So are traffic lights, supermarkets layouts, and the contexts we arrange for one another's behavior. Mediating structures can be embodied in artifacts, in ideas, in systems of social interactions […]" (pp. 290-291).

In this light manipulative abduction in science represents a kind of redistribution of the epistemic and cognitive effort to manage objects and information that cannot be immediately represented or found internally (for example exploiting the resources of visual imagery).[3]

The *hypothetical* character of manipulations in creativity is clear: they are a sort of test, they can be developed to examine further chances, they are a provisional creative organization of experience and some of them become in their turn hypothetical "interpretations" of experience, suggesting new worldviews. Step by step the new interpretation – that at the beginning is completely "practice-laden" – relates to more "theoretical" modes of understanding (narrative, visual, diagrammatic, symbolic, conceptual, simulative).

A Cognitive Theory of the Abductive Modeling Activity

We can say abduction is a complex *process* that works through *imagination*: it suggests a new direction in reasoning by *shaping* new possible ways for explaining object and hypotheses (cf. the templates mentioned above). In this sense imagination should not be confused with an act of intuition. Peirce describes abduction as a dynamic modeling process that fluctuates between states of doubt and states of belief. To solve the doubt, and some eventually linked anomalies, the agent implements a process of information gathering which at the same time relates to the "problem", to the agent's evolving understanding of the situation and to its changing requirements. By imagination here I mean this process of knowledge gathering and shaping. A process, that Kant considered "blind", that leads to *see* things *as* we would not otherwise have seen them: "a blind but indispensable function of the soul, without which we should not have no knowledge whatsoever" (Kant, 1929, A78-B103, p. 112). Scientific creativity, it is pretty obvious, involves seeing the world in a particular new way: scientific understanding permits us to see some aspects of reality in a particular way and creativity relates to this capacity to shed new light. Suggestions which make us able to further analyze this process come from a theory developed in the area of computer vision: the *active perception* approach (see Thomas, 1999).

This approach aims at understanding cognitive systems in terms of their environmental *situatedness*: instead of being used to build a comprehensive inner model of its surroundings, the agent's perceptual capacities are seen as simply used to obtain "whatever" specific pieces of information are necessary for its behavior in the world. The agent constantly "adjusts" its vantage point, updating and refining its procedures, in order to uncover a piece of information. This resorts to the need of specifying how to efficiently examine and explore and to the need of "interpreting" an object of a certain type. It is a process of attentive and controlled perceptual exploration through which the agent is able to collect the necessary information: a purposefully moving through what is being examined, actively picking up information rather than passively transducing (cf. Gibson, 1979).

As suggested for instance by Lederman and Klatzky (1990), this view of perception may be applied to all sense modes: for example, it can be easily extended to the haptic mode**.** Mere passive touch, in fact, tells us little, but by actively exploring an object with our hands we can find out a great deal. Our hands incorporate not only sensory transducers, but musculature which, under central control, moves them in appropriate ways: lifting something tells about its weight, running fingers around the contours provides shape information, rubbing it reveals texture. As already stressed by Peirce in the quotation I already reported above, when dealing with the hypothesizing activity of what I call manipulative abduction, "A man can distinguish different textures of cloth by feeling: but not immediately, for he requires to move fingers over the cloth, which shows that he is obliged to compare sensations of one instant with those of another" (*CP* 5.221).

Thomas (1999) suggests we can think of the fingers together with the neural structures that control, for example, running them so that we can consider the afferent signals that they generate as a sort of (perceptual) *instrument* to gather knowledge: a complex of physiological structures capable of active testing for some environmental property. The study of manipulative abduction that I outlined above, can gain from this approach. To give an example, the role of particular epistemic mediators (*optical diagrams*) in non-standard analysis has been studied, and so their function in grasping and teaching abstract and difficult mathematical concepts (see Magnani and Dossena, 2002). In this case the external models (mathematical diagrams) do not give full available knowledge, but, on the contrary, compel the agent to engage a continuous epistemic dialogue between the diagrams and its internal knowledge to the aim of understanding an already existing information or at "creating" a new one (cf. also the geometrical example in the following section).

It is clear that humans and other animals make a great use of perceptual reasoning and kinesthetic abilities. We can catch a thrown ball, cross a busy street, read a musical score, go through a passage by imaging if we can contort out bodies to the way required, evaluate shape by touch, recognize that an obscurely seen face belongs to a friend of ours, etc. Usually the "computations" required to achieve these tasks are not accessible to a conscious description. Mathematical reasoning uses language explanations, but also non-linguistic notational devices and models. Geometrical constructions represent a relatively simple example of this kind of extra-linguistic machinery we know as characterized in a model-based and manipulative - abductive - way.

Creativity and Ethical Mediators

The active process of information gathering through *mediators*, to shape knowledge, should not be restricted to the

[3] For example it is difficult to preserve precise spatial relationships using mental imagery, especially when one set of them has to be moved relative to another.

scientific activity and so to the "epistemic" side of them. An often neglected side of human creativity is, in fact, related to emotional, artistic, and ethical aspects, and concerns the active shaping of values in an esthetic or in an ethical world. In the case of morality the role of hypotheses and manipulation of the world is clear in Kant's moral doctrine. When Kant considers pure moral rules, in fact, he says that they could be applied to the concrete experience through a kind of "typification", a figurative envisioning of a non existing world, based on a metaphoric mapping, as a means for judging a given moral situation (Johnson, 1993).

As already outlined above, for Peirce all knowing is *inferring* and inferring is not instantaneous, it happens in a process that needs an activity of comparisons involving many kinds of models (signs) in a more or less considerable lapse of time. All sensations or perceptions participate in the nature of a unifying hypothesis, in the case of "emotions" too. In Peircian sense *emotions* too express a kind of model-based reasoning and have and "inferential" character. In decision making emotions play a distinguished role: they make the velocity of the decision process, surely related to what we care about, and lead directly to actions. But they are also usually considered irrational because of the serious disadvantages they present: failure to consider other options, lack of consideration of accurate and relevant information, not sharability in group situation, when the decisions have be adopted collectively. It is important to understand that emotions are not inherently irrational, for example they can be usefully intertwined with cultural aspects.

In general we can say that moral deliberations relate to a sort of selection or creation of principles (rules, prototypes) and to their application to concrete cases. We can both just select (or create, if we do not have any) moral principles (rules, prototypes) and apply them to concrete cases or looking for the best ones among them according to some ethical meta-criteria. When we create new ethics, we provide new knowledge and new rules about problems and situations not yet clearly covered from the moral point of view. In this last case we certainly are in front of a particular case, but the problem is not only the one of ethically solving the case at hand by applying already available ethical concerns – indeed we lack a satisfactory moral knowledge to handle the puzzling situation. Instead we need to create something new, for example new good reasons first which can provide an acceptable intelligibility of the problem. Once created, it will be possible to see the new principle and the new moral knowledge as a crystallization of the various insights emerging from peoples' and/or experts' experience and thinking.

The role of cognitive delegations to external objects and structures has to be extended to the case of human actions and organizations, so viewed as cognitive "mediating" mechanisms endowed with moral aspects. In this light it is possible to introduce the concept of *ethical* (or *moral*) *mediator*. Moral mediators play an important role in reshaping ethical worth of human beings and collectives. They especially involve a continuous reconfiguration of social orders aimed at rebuilding new moral perspectives and chances. These mediators represent a kind of redistribution of the moral effort through managing objects and information in such a way that we can overcome the poverty and the unsatisfactory character of the moral options immediately represented or found internally. *Moral mediators* are also used to exploit latent constraints in the human-environment system. These new constraints grant additional and precious ethical information. When we spontaneously act in a way so that we spend more quality time with our partner to save our marriage, for example, then our actions automatically can cause variables relating to "unexpected" and "positive" contents of the relationship to covary with perceptible new released informative, sentimental, sexual, and in general bodily variables. Prior to the adoption of the new reconfigured "social" order of the couple, there is no active constraint between these hidden and overt variables causing them to carry information about each other. It is also well-known that also "trained" emotions[4] play an important creative role in moral deliberations.

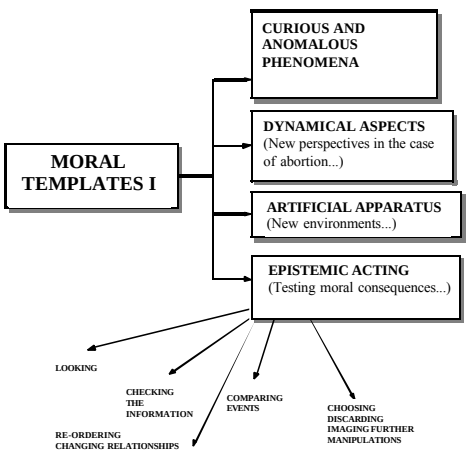


Figure 1. Conjectural moral templates I.

Templates of Moral Doing

It is difficult to establish a list of invariant behaviors that are able to illustrate manipulative reasoning in ethics. As illustrated above, certainly the expert manipulation of non-human objects in real or artificial environments implies the application of old and new *templates* of behavior that exhibit some regularities. As I have said it is important to remember they are embodied and implicit, as tacit forms of acting: I are not referring here to the moral actions and manipulations that simply follow previous explicit and devised plans. Anyway, this moral activity is still conjectural: these templates are embedded hypotheses of moral behavior (creative or already cognitively present in the people's mind-body system, and ordinarily applied) that enable a kind of moral "doing". Hence, some templates of action and manipulation can be *selected* in the set of those available and pre-stored, others have to be *created* for the first time to

[4] That is not just shaped by biological evolution but also by cultural aspects.

perform the most interesting accomplishments of manipulative moral inference.

Some common features of these "tacit" templates that enable us to manipulate external human and non-human things and structures to achieve moral effects are related to (Figure 1): 1. sensibility to the aspects of the moral situation which can be regarded as *curious* or *anomalous*; manipulations can also be performed to be able to introduce potential inconsistencies in the received knowledge (we suddenly adopt a different attitude with respect to our wife/husband to get some reactions we can regard as interesting – or "unexpected" – to confirm or discard hypotheses about her/feelings or to develop further hypotheses about them; in an investigation about a crime we spontaneously engage further manipulations of the evidence to get more interesting data to morally shape the suspect); 2. preliminary sensibility to the *dynamical* character of the situation at hands, and not only to entities and their properties, common aim of manipulations is to practically reorder the dynamic sequence of the events correlated to the main problem to promote the subsequent possibility of new possibilities and options for action (a women in front of decision in favor of abortion spontaneously tries to modify the dynamical aspects of her behavior and the structure of her human relationships to try to establish new perspectives able to make her able to envisage a possible decision different from the first one first envisaged); 3. referral to manipulations that exploit *artificial* created feelings and environments to free new possibly stable and repeatable sources of information about hidden moral knowledge and constraints (when dealing with the moral problem of capital punishment we can spontaneously handle people, for example with statistics, interviews, scientific research, associations, to artificially reconfigure social orders in a way suitable to get real and not hypocritical information, for example about the real relief generated in the victim's relatives by killing the criminal); 4. various contingent ways of spontaneous moral acting: *looking* from different perspectives, *checking* the different information available, *comparing* subsequent events, *choosing*, *discarding*, *imaging* further manipulations, *re-ordering* and *changing relationships* in the world by implicitly *evaluating* the usefulness of a new order (for instance, to help memory) (in the ethical case they certainly are all useful ways for getting suitable evidence and for stimulating the derivation of further consequences to test our previously established moral judgments; analogous of all these manipulative templates are active in epistemic settings, as illustrated in Magnani, 2001).

More features of our tacit templates and ethical mediators are related to the following additional issues (Figure 2): 5) moral spontaneous action that can be useful in presence of *incomplete* or *inconsistent* information – not only from the "perceptual" point of view – or of a diminished capacity to morally act upon the world: it is used to get more data to restore coherence and/or to improve deficient knowledge; 6) action as a *control of sense data* illustrates how we can change the position of our body (and/or of the external objects) to reconfigure social orders, collective relationships, and how to exploit various kinds of artificially created events to get various new kinds of stimulation: action provides some tactile, visual, kinesthetic, sentimental, emotional, and bodily information (e.g, in taking care of people, cf. below in the following subsection), otherwise unavailable; 7) action enables us to build new *external artifactual models* of ethical mechanisms and structures (for example through "institutions") instead of the corresponding "real" and "natural" ones.[5] For instance, we can substitute to the "natural" structure "family" an environment more adequate to agent's moral needs. In this case we aim at reconfiguring relationships for instance when we exploit the social reshaping role of the "houses" were children molested inside family are recovered, to rebuild in a whole artificial framework their moral perception for example of the sexual molestation received and of the related bad feelings. Something similar occurs in the case of the addicted people. We also establish structures to implicitly favor good manners, for example fences, barriers in the lines, etc.

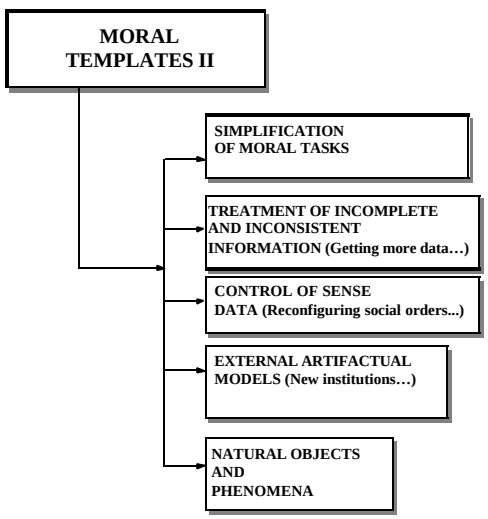


Figure 2. Conjectural moral templates II.

Moral Mediators

The whole activity of manipulation is also devoted to build various external *moral mediators*[6] that function as an enormous new source of information and knowledge. Therefore, these mediators represent a kind of redistribution of the moral effort through managing objects and information in such a way that we can overcome the poverty and the unsatisfactory character of the moral options immediately represented or found internally (for example exploiting the resources in terms of merely internal/mental moral principles, utilitarian envisaging, and model-based moral reasoning).

[5] Of course these "real" and "natural" structures are also artificial, because we can think of a "family" as a kind of not merely natural institution.

[6] I derive this expression from the one "epistemic mediators" I introduced in Magnani (2001, chapter 3): these consist of external representations, objects, and artifacts that are relevant in scientific discovery and reasoning processes.

Not only a way for moving the world to desirable states, action performs a moral and not just merely performatory role: people structure their worlds to simplify and solve moral tasks when they are in presence of incomplete information or possess a diminished capacity to morally act upon the world when they have insufficient opportunities to know. *Moral mediators* are also used to exploit latent constraints in the human-environment system. These elicited new constraints grant us additional and precious ethical information: when we spontaneously act in a way in which we spend more quality time with our partner to save our marriage, then our actions automatically cause variables relating to "unexpected" and "positive" contents of the relationship to covary with perceptible new released informative, sentimental, sexual, and, in general, bodily variables. Prior to the adoption of the new reconfigured "social" order of the couple, there is no active constraint between these hidden and overt variables causing them of carry information about each other

Conclusion

What I call *theoretical abduction* (sentential and manipulative) certainly illustrates much of what is important in creative abductive reasoning both in humans and computational programs, especially the objective of selecting and creating a set of hypotheses that are able to dispense good (preferred) explanations of data, but fails to account for many cases of explanations occurring in science or in everyday reasoning when the exploitation of the environment is crucial. The concept of *manipulative abduction* is devoted to capture the role of action in many interesting situations: action provides otherwise unavailable information that enables the agent to solve problems by starting and performing a suitable abductive process of generation or selection of hypotheses. Many external things, even if usually inert from the epistemological point of view, can be transformed into what is called *epistemic mediators*, which are illustrated in the second part of this paper, together with an analysis of the related notion of "external representation steps in a way that discharges the "internal" mind of a computational load. To define a cognitive system it seems we can no longer identify it only with internal processing devices.

By exploiting the concept of "thinking through doing" and of manipulative abduction I have tried to shed new light on some of the most interesting cognitive aspects of creative ethical reasoning of what I call "ethical mediators". Indeed, I contend that the whole activity of manipulation can be seen as an activity for building various external "ethical mediators" that function as an enormous new source of information and knowledge. Furthermore, while describing morality "through doing" a list of "moral templates" as forms of invariant behaviors that are able to illustrate manipulative ethical reasoning is furnished. These templates are forms of behavior which are inclined towards providing ethical outcomes. The application of old and new (creative) moral templates of behavior exhibits some regularities and expresses expert manipulation of human and non-human objects in real or artificial environments. These templates are embodied and implicit as tacit forms of acting. They are embedded hypotheses of moral behavior (creative or already cognitively present in the people's mind-body system, and ordinarily applied) that enable a kind of moral "doing". Hence, some templates of action and manipulation can be selected in the set of those available and pre-stored, while others have to be created for the first time in order to perform the most interesting accomplishments of manipulative moral inferences. These "tacit" templates enable us to manipulate external human and non-human things and structures to achieve moral effects.

References

Boden, M.A. (1991). *The Creative Mind: Myths and Mechanisms*. New York: Basic Books, New York.

Buchanan, B.G. (2001). Creativity at the metalevel. AAAI-2000 presidential address. *AI Magazine, fall 2001*, 13-28.

Gibson, J.J.. *The Ecological Approach to Visual Perception.* Houghton Mifflin, Boston, MA, 1979.

Gooding, D. (1990). *Experiment and the Making of Meaning*. Dordrecht: Kluwer.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Johnson, M. (1993). *Moral Imagination. Implications of Cognitive Science in Ethics.* Chicago: The University of Chicago Press.

Kant, I. (1929). *Critique of Pure Reason.* Translated by N. Kemp Smith. London: MacMillan. Reprint 1998. Originally published 1787.

Kirlik, A. (1998). The ecological expert: acting to create information to guide action. In *Proceedings of the 1998 Conference on Human Interaction with Complex Systems* (HICS'98). Piscataway, NJ: IEEE Press.

Lederman, S.J. and Klatzky, R. (1990). Haptic exploration and object representation. In M.A. Goodale (Ed.) *Vision and Action: The Control of Grasping* (pp. 98-109) Norwood, NJ: Ablex.

Magnani, L. (2001). *Abduction, Reason, and Science. Processes of Discovery and Explanation.* New York: Kluwer Academic/Plenum Publishers.

Magnani, L. and Dossena, R. (forthcoming). Perceiving the infinite and the infinitesimal world: unveiling and optical diagrams and the construction of mathematical concepts. To appear in *Foundations of Science*, 2002.

Magnani, L. and Nersessian, N.J. (Eds.) (2002). *Model-Based Reasoning: Science, Technology, Values.* New York: Kluwer Academic/Plenum Publishers.

Nersessian, N.J. (1995). Should physicists preach what they practice? Constructive modeling in doing and learning physics. *Science and Education, 4*, 87-120.

Peirce, C.S. (1931-1958). *Collected Papers* 1-6 (*CP*). Edited by C. Hartshorne and P. Weiss. *Collected Papers* 7-8. Edited by A. Burks. Harvard University Press, Cambridge, 1931-35, 1958.

Sternberg, R.J., Kaufman, J.C., and Pretz, J.E. (Eds.). *The Creativity Conundrum : A Propulsion Model of Kinds of Creative Contributions*. New York: Psychology Press.

Thomas, N.J. (1999). Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science, 23(2),* 207-245.

Event categorization: A cross-linguistic perspective

Asifa Majid (asifa.majid@mpi.nl)
Max Planck Institute for Psycholinguistics, Postbus 310 Nijmegen, 6500 AH The Netherlands

Miriam van Staden (m.vanstaden@uva.nl)
Department of Theoretical Linguistics, Spuistraat 210, 1012 VT Amsterdam, The Netherlands

James S. Boster (james.boster@uconn.edu)
Department of Anthropology, 354 Mansfield Road Storrs, CT 06269-2176 USA

Melissa Bowerman (melissa.bowerman@mpi.nl)
Max Planck Institute for Psycholinguistics, Postbus 310 Nijmegen, 6500 AH The Netherlands

Abstract

Many studies in cognitive science address how people categorize objects, but there has been comparatively little research on event categorization. This study investigated the categorization of events involving material destruction, such as "cutting" and "breaking". Speakers of 28 typologically, genetically, and areally diverse languages described events shown in a set of video-clips. There was considerable cross-linguistic agreement in the dimensions along which the events were distinguished, but there was variation in the number of categories and the placement of their boundaries.

Introduction

Categorization research in cognitive science has focused overwhelmingly on the mental representation of **objects**. Behavioral studies with adults, neuropsychological studies with patient populations, cross-cultural comparisons, and acquisition evidence provide converging evidence about how objects are represented. For example, objects are stored according to semantic domains, with natural kinds represented distinctly from artifacts. Within these categories there are subdivisions: animals are stored separately from fruits, while musical instruments are stored separately from furniture (Shallice, 1988). Objects are organized not only by semantic domain but also hierarchically, with categories at the superordinate, basic, and subordinate levels (Rosch, 1978). Basic level categories are cognitively privileged, in the sense that they are labeled with shorter words, they constitute the preferred level of naming, they can be verified faster than superordinate and subordinate categories in judgment tasks, and they are acquired earlier by children (Brown, 1958; Rosch et al., 1976). There also appears to be considerable cross-cultural consensus in the organization of object representations (Berlin, 1992; Malt, 1995).

In contrast to all the work on objects, relatively little has been done on the mental representation of **events**. One line of research, with roots in social psychology, has investigated how people segment events (Newtson & Engquist, 1976; Newtson, Engquist, & Bois, 1977; Zacks et al., 2001). Another important line of work on event representation, originating in cognitive psychology and artificial intelligence, has examined the organization of event knowledge in scripts, frames, and schemas (Minsky, 1975; Schank & Abelson, 1977).

Neither of these approaches to event representation has examined how everyday activity types are categorized. Studies of event segmentation do not ask which event segments are regarded as being "of the same kind". Script and frame research concentrates on scenarios like "going to the movies", "going to a restaurant", "sports", or "housework" (Morris & Murphy, 1990; Rifkin, 1985). These scenarios are often culture-specific, and so do not lend themselves to cross-cultural research. They are also complex, consisting of sequences of finer-grained events such as "walking into the restaurant", "sitting down", "ordering", "eating", and "paying the bill". Little is known about how uniformly people categorize such finer-grained units, but it has been widely assumed – certainly by developmentalists – that there is a universal core set of everyday event types and that children learn basic verbs such as *have, hit, move, put*, and *give* by linking them directly to these concepts (Gleitman, 1990; Pinker, 1989).

In the present study, we focus on the linguistic categorization of a set of everyday events of "cutting and breaking" – more formally known as events involving a "separation in the material integrity" of objects (Hale & Keyser, 1987).[1] This domain was chosen because such events are universal and do not rely on specialized knowledge; they are accessible to everyone. The manufacture and use of tools for purposes of cutting and breaking has been dated back to at least 2.5 million years ago in the East African Rift area. Modern humans (*homo sapiens sapiens*) appear to be distinctive for making and using particular tools for "cutting", such as pressure-flaked knives (Toth & Schick, 1993). "Cutting" and "breaking" can, then, be taken as human activities that are central to human language and cognition.

We examine the categorization of "cutting and breaking" events by looking at how speakers of

[1] The terms "cutting" and "breaking", with quotes, designate actions of the type that speakers of English typically label with verbs like *cut* and *break*; other languages may or may not have words with closely similar meanings. Throughout this paper, words in quotation marks point to actions of a certain general type, and words in italics designate linguistic forms.

genetically, typologically, and areally diverse languages describe a set of actions shown in video-clips. Do speakers of all languages make the same distinctions when they are talking about such events?

The verbs c*ut* and *break* have been widely discussed in the linguistics literature. One influential approach has suggested that "cutting"-type verbs and "breaking"-type verbs can be universally distinguished on the basis of their semantic and syntactic behavior (Guerssel et al., 1985). This suggests that speakers of different languages should recognize similar distinctions.

Other work, however, suggests that there may be significant differences in the way languages categorize "cutting" and "breaking" events; for example, English speakers use *break* for actions on a wide range of objects (e.g., a plate, a stick, a rope), while speakers of K'iche' Maya must choose from among a set of "breaking" verbs on the basis of properties of the object; e.g., *-paxi:j* 'break a rock, glass, or clay thing' (e.g., a plate); *-q'upi:j* 'break (other kinds of) hard thing' (e.g., a stick); *-tóqopi'j* 'break a long flexible thing' (e.g., a rope) (Pye, 1996; Pye, Loeb, & Pao, 1995). Differences in the categorization of "cutting and breaking" events might also be expected due to variation in cultural tools and techniques; for example, Americans and Europeans chop vegetables by holding them still and bringing a knife down on them from above, whereas Punjabi speakers in rural Pakistan and India often move the vegetables against a stationary curved knife.

In studying the categorization of "cutting and breaking" events, it is not obvious a priori what the domain of investigation should be taken to encompass. Whereas speakers of English do not use *cut* and *break* for actions like peeling a banana or pulling paper cups apart, and they do not use *open* for events like breaking the stem off an apple, perhaps such categorizations occur in other languages. Children learning English in fact make such overextensions (Bowerman, in press; Schaefer, 1979), which suggests that the boundaries of the "cutting and breaking" domain may not be cognitively obvious, and therefore not universally shared. One important goal for the present study, then, is not only to examine the categorization of "cutting and breaking" events by speakers of different languages, but also to discover the extent to which "cutting and breaking" events hang together in the first place as a relatively coherent semantic domain, as distinct from events involving other kinds of separations.

Method

Participants

Event descriptions were collected from speakers of 28 typologically, genetically and areally diverse languages. For each language there were between one and seven consultants. Twenty researchers collaborated in this effort, all of them experts on the language they worked on – a critical point for the validity of the coding of the data (see Results section). Data collection was carried out in the language being studied, not a contact language. Details of the languages, language affiliations, and researchers responsible for the collection and coding of the data are given in Table 1.

Materials

The data were collected using a set of 61 video-clips that depicted a wide range of events (Bohnemeyer, Bowerman, & Brown 2001). The majority of these clips showed an event in which an actor brought about a change of state in an object – specifically, some kind of destruction of the object's material integrity. Some clips depicted state-change events that involved separation but not material destruction, such as opening a pot or pulling paper cups apart. Still others depicted "peeling" events, which share properties with events of both material destruction and simple separation. Stimuli were constructed by varying the agent, the instrument used, the object acted upon, the manner of the destruction, and the prototypicality of the event (see Figure 1).

Figure 1: Example stills from video clips

Procedure

Consultants saw one video-clip at a time on a laptop. The clips were presented in a fixed order. The consultants' task was to describe what the agent did. After free description they were asked what other descriptions could be applied felicitously to each clip. They were also asked whether other descriptions would be infelicitous.

Results

Coding

We defined the target event we were interested in as the change in an object from a state of integrity to a state of separation or material destruction. For each of the languages, the researcher who collected the data identified those constituent(s) of a speaker's description which

Table 1: Language details and associated researchers

Language	Language affiliation	Country	Researcher
Biak	Austronesian	Indonesia	W. van de Heuvel
Chontal	Isolate	Mexico	L. O'Connor
Dutch	Indo-European	Netherlands	M. van Staden
English	Indo-European	UK, USA	M. Bowerman, A. Majid, C. Wortmann
Ewe	Niger-Congo	Ghana	F. Ameka
German	Indo-European	Germany	M. van Staden
Hindi	Indo-European	India	B. Narasimhan
Jalonke	Niger-Congo	Guinea	F. Lüpke
Japanese	Isolate	Japan	S. Kita
Kilivila	Austronesian	Papua New Guinea	G. Senft
Lao	Tai	Laos	N. Enfield
Likpe	Niger-Congo	Ghana	F. Ameka
Mandarin	Sino-Tibetan	China	J. Chen
Miraña	Witotoan	Colombia	F. Seifart
Otomi	Otomanguean	Mexico	E. Palancar
Punjabi	Indo-European	Pakistan	A. Majid
Spanish	Indo-European	Spain, Mexico	M. Bowerman, E. Palancar
Sranan	Creole	Surinam	J. Essegbey
Swedish	Indo-European	Sweden	M. Gullberg
Tamil	Dravidian	India	B. Narasimhan
Kuuk Thaayorre	Pama-Nyungan	Australia	A. Gaby
Tidore	West Papuan Phylum	Indonesia	M. van Staden
Tiriyó	Cariban	Brazil	S. Meira
Touo	Papuan Isolate	Solomon Islands	M. Dunn, A. Terrill
Turkish	Altaic	Turkey	A. Özyürek
Tzeltal	Mayan	Mexico	P. Brown
Yélî Dyne	Papuan Isolate	Rossel Island	S. Levinson
Yukatek	Mayan	Mexico	J. Bohnemeyer

encoded the event. For example, the event of "a boy cutting a carrot", at the top left of Figure 1, can be expressed in English as *The boy cut the carrot*. Here the caused state-change event is expressed solely by the transitive verb *cut*.

Languages differ in whether information about the state change is typically located in a single verb or is spread out across a number of constituents, such as additional verbs or particles. For example, speakers of Mandarin use verb compounds to describe many of the events; e.g., *qie1-duan4* 'cut-break.long.thin.object' for the scene of someone karate-chopping a carrot shown in the lower left corner of Figure 1. For purposes of the present study, we concentrated on how the stimuli were categorized by the **verbs** of a language. Every verb in the data that described the target event was input to the analysis.

Analysis

Speakers' event descriptions can be treated as analogous to the data obtained in sorting tasks designed to study categorization. In a typical sorting task, a subject might receive a set of cards, each depicting a different stimulus, and be asked to sort them into piles of objects that are similar. Speakers in the present study received no metalinguistic instructions; they were simply asked to describe what they saw in the video-clips. But each different verb they applied to the target events was taken to define a category ("pile"). Across languages (and of course also within individuals or across individuals within the same language), stimuli that are often described with the same verb ("are sorted into the same pile") can be taken to be more similar to each other than stimuli that typically fall under different verbs (Bowerman, 1996). Multivariate statistics can then be used to explore the similarity structure of the data set as a whole.

To extract the most important dimensions organizing the similarity space of our stimuli, we used correspondence analysis (Greenacre, 1984). Correspondence analysis provides a dual factoring of a rectangular matrix in which the column scores and row scores are projected into the same low dimensional space. To perform the correspondence analysis, we first transformed the linguistic data for each language into a similarity matrix. This was done by determining, for all scenes taken pairwise, whether each member of the pair was ever described by the same verb. If so, the pair was assigned a similarity score of one; if not, zero.[2]

[2] This technique was adopted rather than a more graded approach to similarity based on the number of speakers within each language who used the same description, so as not to bias the results toward the categorizations favored by languages for which we happened to have more speakers.

The similarity matrices from all the languages were then stacked one on top of another to build a matrix with 61 columns (the stimuli) and 28*61 (language*stimuli) rows. This matrix was submitted to correspondence analysis to find the dimensions that are cross-linguistically the most important in structuring the similarity space of the stimulus set. The analysis extracts first the dimension that accounts for the most variance, then the dimension that accounts for the next most variance, and so on. Each stimulus scene is positioned in this multidimensional space in such a way that the distance between any two scenes reflects the degree to which, across languages, people described them with the same verbs. Scenes often described with the same verb are positioned close together, while scenes that are rarely or never described with the same verb are positioned far apart.

The major dimensions

The first and most important dimensions extracted in our analysis distinguished between events of material destruction and other events involving separation. There was widespread consensus across languages that events of "taking apart" (e.g., separating paper cups), "opening" (e.g., opening a box) and "peeling" (fruit) should be described with different verbs than events of "cutting and breaking". "Cutting and breaking" events are distinguished as a group from other kinds of separation, and so form a coherent semantic domain.

Leaving aside the events of "taking apart", "opening", and "peeling", we next focused specifically on the similarity structure of the remaining 46 events. These stimuli were analyzed with the same procedure outlined in the previous section.

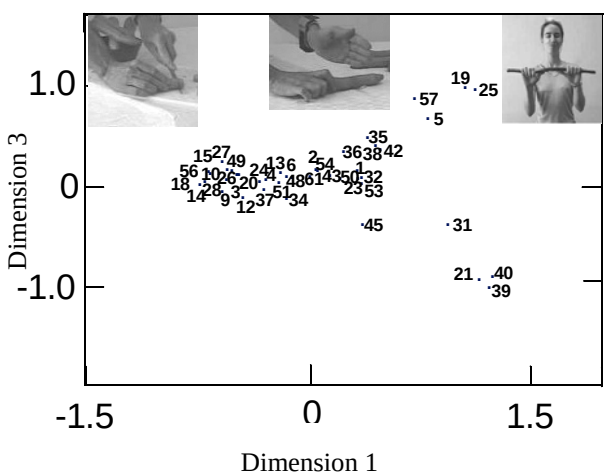


Figure 1: Plot of scenes, based on all languages, along Dimensions 1 and 3. Dimension 1 distinguishes events with precise control over the locus of separation (cutting a carrot with a knife) from scenes with intermediate control (karate-chopping a carrot) and imprecise control (breaking a stick with the hands).

The first and most important dimension of this analysis distinguishes among events on the basis of how precisely the agent controls the locus of the separation in the object. The events are distributed continuously on this dimension. (See Figure 1 for the placement of the scenes along Dimension 1. Each scene is represented by a number.) Events involving relatively precise control (e.g., cutting a carrot with a knife, scene 10) is positioned to the left, events with imprecise control (e.g., breaking a stick with the hands, scene 19) to the right, and events with intermediate degrees of control (e.g., karate-chopping a carrot, scene 32) in between. Events intermediate on this dimension are treated variably across languages, with some languages grouping them with the "precise control" events positioned to the left, others with the "imprecise control" events positioned to the right, and still others assigning them to categories of their own.

Dimension 2 distinguishes just two scenes from the rest – those showing an agent tearing a piece of cloth (a two-dimensional flexible object) partially (scene 36) or completely (scene 1) with the hands. These events were labeled *tear* in English, as distinct from *cut* and *break*. Nineteen out of the 28 languages have a verb that was used to categorise these and only these scenes. The remaining 9 languages did not distinguish these scenes, but grouped them in various ways with other scenes.

Within the group of scenes pulled out on Dimension 1 as lacking precise control over the locus of separation, Dimension 3 makes a further distinction between "snapping" and "smashing" events (see Figure 2a). The "snapping" cluster comprises events in which a one-dimensional rigid object is separated into two pieces by applying pressure to both ends (scenes 25, 19, 57, 5), while the "smashing" cluster is made up of events in which a rigid object is fragmented into many pieces by applying a blow, e.g., with a hammer (40, 39, 21, 31). The Dimension 3 distinction between "snapping" and "smashing", like the Dimension 2 distinction between "tearing" and separations of other kinds, is respected by speakers of many languages – cf. the distinction in Likpe between events described with *f3s3* (the snapping scenes) and those described with *ba* (the smashing scenes) (see Figure 2b). But this distinction is not made in all languages; colloquial Tamil, for example, collapses these two categories (along with a few additional scenes) into a single event type, denoted by the verb *oDai* (see Figure 2c).

Discussion

Speakers of a variety of typologically, genetically and areally diverse languages agree to a surprising extent in their linguistic categorization of events of material destruction of objects ("cutting and breaking" events). First, they agree on treating such events as a relatively coherent semantic domain. A priori, it is not obvious that languages will distinguish "cutting and breaking" events as a group from events involving other kinds of separations of objects or object parts, such as "taking apart", "opening", and "peeling"; after all, learners of English make a number of errors suggesting that the

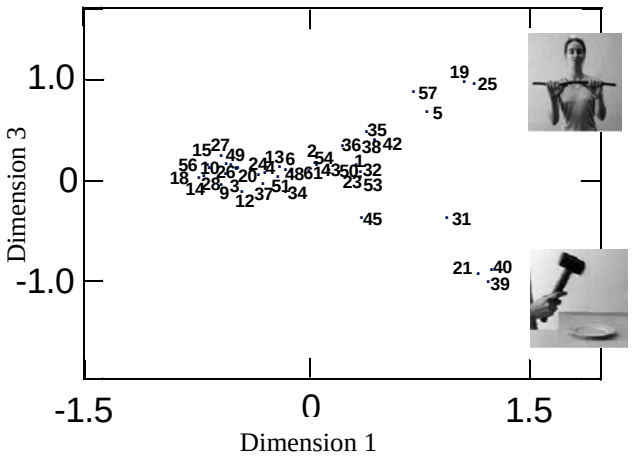


Figure 2a: Plot of scenes, based on all languages, along Dimensions 1 and 3, showing the distinction between "snapping" and "smashing" events.

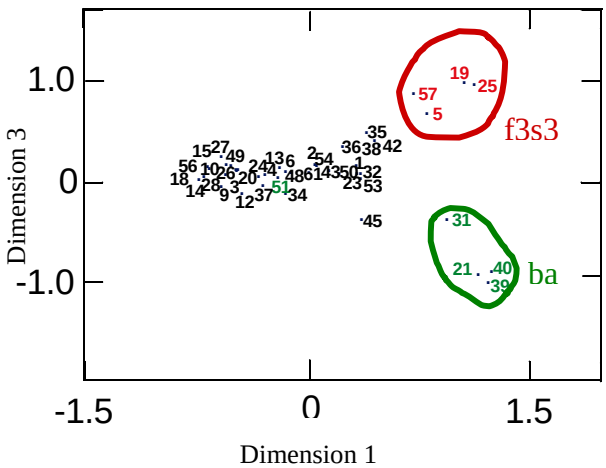


Figure 2b: Likpe is a good example of a language which distinguishes "snapping" from "smashing" events.

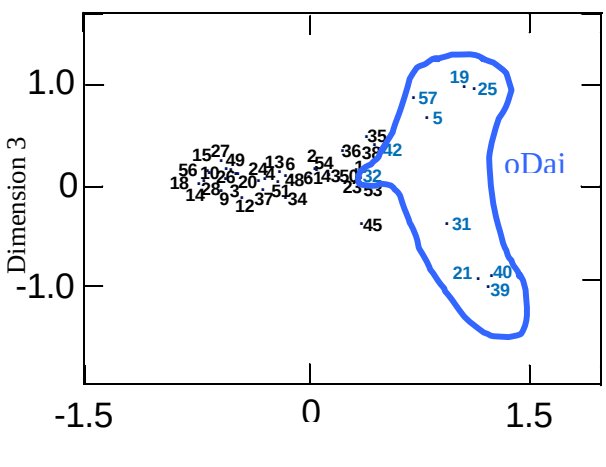


Figure 2c: Tamil collapses the "snap-smash" distinction.

boundaries of these event types are not obvious. For this reason our set of events to be described included not only scenes of "cutting and breaking", but also of various other kinds of separations. But these other separations were rarely described with the same verbs that were applied to the core set of "cutting and breaking" events. The "cutting and breaking" events were treated as far more similar to each other than they were to the other kinds of separations, in the sense that they were much more often described by the same verbs.

Second, speakers of different languages also showed considerable agreement in the kinds of distinctions they drew **within** the domain of "cutting and breaking" events. Although their societies ranged from industrial urban-dwelling to rainforest-dwelling swidden agriculturist, and they varied in their tools and techniques for cutting and breaking things in their daily lives, they converged on a shared similarity space for events of "cutting and breaking". The most important dimension for the set of 28 languages taken as a group distinguishes events featuring precise control over the locus of separation from those with imprecise control (roughly, "cutting" events vs. "breaking" events). Further, "tearing" events are very often distinguished from among other events with an intermediate degree of control (Dimension 2), while "snapping" and "smashing" events are often distinguished among the events involving imprecise control (Dimension 3).

Despite this cross-linguistic agreement there were also many differences – language-learners clearly have something to learn. Speakers of different languages varied in the number of categories of "cutting and breaking" they recognized and in where they placed the category boundaries. For example, speakers of most of the languages respected the distinction between "tearing" and other actions of material destruction, but some did not; speakers of many languages rigorously distinguished between actions of "snapping" and "smashing", but some did not (see Figures 2a-c); and languages differed in where they placed the boundary between "precisely" and "imprecisely" controlled acts of separation. These differences respected the overall structure of the semantic space; for example, no speakers described events at the far left of Dimension 1 with the same verb(s) as events at the far right, while describing the events falling between them with different verbs.

One topic we have not yet mentioned is how a language's semantic categories of "cutting and breaking" are related to one another. For instance, English clearly organizes its "cutting and breaking" terms hierarchically, with the high-frequency verbs *break* and *cut* each encompassing a number of more specific subtypes, such as *snapping* and *smashing* for *break*, and *slicing* and *chopping* for *cut*. This kind of organization is less apparent in many of the other languages in our sample. For example, Dutch has no verbs for "cutting and breaking" with as wide an application as English *cut* and *break*. "Cutting" events are obligatorily subdivided according to whether they involve a single-bladed tool like a knife or a double-bladed tool like scissors (*snijden* vs. *knippen*), and there is also no cover term for a wide range of "breaking" events; e.g., *breken* – cognate with English *break* – is used only for "snapping" events. It is unclear, then, whether the hierarchical organization found

across languages in words for **objects** will also be characteristic of words for **events**.

A final topic that we also leave to future work is the intriguing question of how the categorization of events imposed by language is related to categorization as studied with nonlinguistic techniques such as similarity ratings. For the object domain of "containers", speakers of different languages classified nonlinguistically more similarly than they classified linguistically (Malt et al., 1999). Whether the same will be true for event categories remains to be seen.

Acknowledgments

This study of "cutting and breaking" took place in the Event Representation project at the Max Planck Institute for Psycholinguistics. We thank all our colleagues who contributed their insights, data, and analysis to the study. The research was supported by the Max Planck Gesellschaft, as well as by a European Union Marie Curie Fellowship awarded to the first author, and a NWO grant to the second author. The authors are solely responsible for information communicated and the European Commission is not responsible for any views or results expressed.

References

Berlin, B. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton, NJ: Princeton University Press.

Bohnemeyer, J., Bowerman, M., & Brown, P. (2001). Cut and break clips, version 3. In S. C. Levinson & N. Enfield (Eds.), *Field Manual 2001*. Language & Cognition Group, Max Planck Institute for Psycholinguistics.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65,* 14-21.

Bowerman, M. (1996). Learning how to structure space for language: A crosslinguistic perspective. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.) *Language and Space*. Cambridge MA: MIT Press.

Bowerman, M. (in press). Why can't you 'open' a nut or 'break' a cooked noodle? Learning covert object categories in action word meanings. In L. Gershkoff-Stowe & D. Rakison (Eds.), *Building object categories in developmental time*. Mahwah, NJ: Lawrence Erlbaum.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition 1,* 3-55.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press: London.

Guerssel, M., Hale, K., Laughren, M., Levin, B., White Eagle, J. (1985). A cross-linguistic study of transitivity alternations. In W. H. Eilfort, P. D. Kroeber, & K. L. Peterson (Eds.), *Papers from the Parasession on Causatives and Agentivity at the Twenty-First Regional Meeting*. Chicago, IL: Chicago Linguistics Society.

Hale, K., & Keyser, S. J. (1987). *A view from the middle*. Lexicon Project Working Papers 10. Cambridge, MA: MIT, Center for Cognitive Science.

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive psychology, 29,* 85-148.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M. Y., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40,* 230-262.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision.* New York: McGraw-Hill.

Morris, M. W., & Murphy, G. L. (1990). Converging operations on a basic level in event taxonomies. *Memory & Cognition, 18,* 407-418.

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12,* 436-450.

Newtson, D. & Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35,* 847-862.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pye, C. (1996). K'iche' Maya verbs of breaking and cutting. *Kansas Working Papers in Linguistics, 21* (part II).

Pye, C., Loeb, D. F., & Pao, Y.-Y. (1995). The acquisition of breaking and cutting. In E. Clark (ed.), *Proceedings of the Twenty-seventh Annual Child Language Research Forum*. Stanford: CSLI Publications.

Rifkin, A. (1985). Evidence for basic level in event taxonomies. *Memory & Cognition, 13,* 538-556.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B.B. Lloyd (Eds.) *Semantic factors in cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 7,* 573-605.

Schaefer, R. (1979). Child and adult verb categories. *Kansas Working Papers in Linguistics, 4,* 61-76.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Schank, R. C. & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Lawrence Erlbaum.

Toth, N. & Schick, K. (1993). Early stone industries and inferences regarding language and cognition. In K. R. Gibson & T. Ingold (Eds.) *Tools, language and cognition in human evolution*. Cambridge: Cambridge University Press.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L. & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4,* 651-655.

Mapping Written Input onto Orthographic Representations: The Case of Bilinguals With Partially Overlapping Orthographies

Viorica Marian (v-marian@northwestern.edu)
Department of Communication Sciences and Disorders, Northwestern University
Evanston, IL 60208 USA

Margarita Kaushanskaya (m-kaushanskaya@northwestern.edu)
Department of Communication Sciences and Disorders, Northwestern University
Evanston, IL 60208 USA

Abstract

Mapping of written input onto orthographic representations was examined in bilingual speakers whose two languages have partially overlapping orthographies. Russian-English bilinguals and English monolinguals were tested with a modified version of the picture-word interference paradigm, adapted for use with eye-tracking. Compared to English monolinguals, Russian-English bilinguals (tested in English) made more eye movements to written stimuli that, if mapped onto two orthographic systems simultaneously, constituted Russian words. Results suggest parallel activation of both languages during visual processing of written input, even when the orthography is associated with different phonological representations in the two languages. We suggest that decoding of written input in languages with partial orthographic overlap is not limited to one language only, but that the mapping of visual stimuli takes place onto the orthographic systems of both languages and that lexical representations in the non-target language become activated.

Introduction

Recent studies of bilingual language processing challenge earlier accounts of the language switch hypothesis (e.g., MacNamara & Kushnir, 1971), according to which bilinguals are able to selectively activate and deactivate their two languages. Instead, data support interactive parallel processing accounts, according to which linguistic input activates both languages simultaneously. For *spoken* word recognition, evidence supporting activation of both lexicons comes from research investigating spoken language processing in bilinguals using eye-tracking (Marian & Spivey, 2003a,b; Spivey & Marian, 1999). In the eye-tracking paradigm, participants are given spoken instructions to move objects on a table while their eye movements are recorded. Although participants rarely pick up incorrect objects, it is often observed that they fixate objects that have similar phonology to the spoken word (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The eye-tracking technique, merging input from both the visual and auditory modalities, was adapted for use with bilinguals to index activation of a second language non-linguistically. For example, when Russian-English bilinguals were presented with a visual display containing four objects (actual objects or toy replicas, as applicable), such as a shark, a balloon (*sharik* in Russian), a horse, and a napkin, and were instructed in English to "pick up the shark," they frequently made eye movements to the cross-linguistic phonological competitor *sharik*. In this case, the Russian word *sharik* was a cross-linguistic cohort (cf. Marslen-Wilson, 1987; see also Cutler, 1995; Marslen-Wilson & Welsch, 1978) of the English target word *shark*, i.e., the beginning portion of the name of the target object carried phonetic similarity to the name of one of the other objects in the other language. Eye movements to the cross-linguistic cohort, even when the other language is not being used overtly, supports the hypothesis that phonemic input initially activates both languages during bilingual spoken language processing.

For *written* word recognition, studies examining whether or not both languages are activated in parallel used code switching (e.g., Doctor & Klein, 1992; Grainger, 1993; Grainger & Dijkstra, 1992; Li, 1996; Nas, 1983; Soares & Grosjean, 1984), phoneme monitoring (e.g., Colome, 2001), lexical decision (e.g., Brysbaert, Van Dyck, & Van de Poel, 1999; DeGroot, Delmaar, & Lupker, 2000; Dijkstra, Grainger, & van Heuven, 1999), and priming tasks (e.g., Beauvillain & Grainger, 1987). Results indicate that orthographic input simultaneously activates lexical items across the two lexicons in the very early stages of processing, that bilingual visual word recognition is based on a stimulus-driven analysis indifferent to language, and that lexical representation in bilingual visual word recognition is governed by orthography rather than by language. For example, Bijeljac-Babic, Biardeau, and Grainger (1997) investigated activation of orthographic representations in bilingual visual word recognition by using a masked priming paradigm. Orthographic priming was observed in both monolingual and bilingual conditions, suggesting that printed strings of letters can simultaneously activate lexical representations in both languages, insofar as these share the same alphabet.

In another study on visual word recognition, Van Heuven, Dijkstra, and Grainger (1998) used the interlingual neighbors paradigm (an orthographic neighbor is any word differing by a single letter from the target word). Cross-

language interference on target word recognition was examined with a comprehensive corpus of Dutch and English words by varying the number of orthographic neighbors of the target word in the non-target language. The results showed that words in the non-target language with a greater number of orthographic neighbors in the target language had slower response times than words that had fewer orthographic neighbors in the target language. An increase in orthographic neighbors within the same language consistently produced inhibitory effects for the non-target language and facilitatory effects for the target language.

Because this work is based on bilinguals whose languages share orthography and where orthography-to-phonology mappings largely overlap across the two languages, the extent of parallel activation for languages that do not share orthography, or share it only partially, remains unclear. Studies with bilinguals that speak languages that do not overlap orthographically are limited (e.g., Tzelgov, Henik, Sneg, & Baruch, 1996). Languages that do not share visual representation make it possible to examine phonological and semantic activation of the non-target language during bilingual reading (Besner & Hildebrandt, 1987; Bowers, Mimouni, & Arguin, 2000; Brown, Sharma, & Kirsner, 1984; Chen & Tsoi, 1990; Smith & Kirsner, 1982), but not activation of the written form of both languages. Testing Russian-English bilinguals whose two languages share some, but not all, orthographic and phonological forms, provides precisely this advantage--it becomes possible to dissociate the activation of phonology and orthography during language processing by manipulating stimulus make-up.

The two languages of a Russian-English bilingual include some graphemes that share both visual and auditory form (e.g., *K*), other graphemes that share visual, but not auditory form (e.g., *P*, which in Russian reads *R*), yet others that share auditory, but not visual form (letters specific to the Latin vs. Cyrillic alphabets). Of particular interest in designing the present study are the 12 letters that overlap orthographically across English and Russian. Of these, 6 share both orthography and phonology—A, E, K, M, O, T. The remaining six, although identical orthographically, carry no phonological overlap—B, C, H, P, Y, X (the corresponding phonological representations in Russian are, following the Library of Congress Transliteration Schemes for Non-Roman Scripts (1991): B-v, C-s, H-n, P-r, Y-u, X-h). Testing Russian-English bilinguals makes it possible to examine the mapping of the visual stimulus onto orthographic and phonological representations in the two languages during bilingual lexical access in circumstances where phonemic overlap across two languages is possible without orthographic overlap and where orthographic overlap between the two languages is possible without phonemic overlap. By manipulating orthographic form and the associated phonological representations, the present experiment tests activation of the other language when the written input shares orthographic, but does not share phonological, representation across languages. The stated relationship between Russian and English in terms of phonological and orthographic structure of the two languages can provide valuable insights into orthographic and phonological processing and contribute to understanding the extent to which constraints imposed by language structure modulate cross-language interactions.

The only other similar work exploring processing of Latin and Cyrillic alphabets comes from studies of monolingual speakers of Serbo-Croatian (e.g., Feldman & Turvey, 1983; Lukatela, Savic, Gligorijevic, Ognjenovic, & Turvey, 1978). Serbo-Croatian as a language is unique in that it uses two alphabets, Latin and Cyrillic. Serbo-Croatian speakers are slower in lexical decision tasks when two phonological interpretations could be assigned to the same letter string, an effect sensitive to the number and distribution of ambiguous characters. The major difference between studying Russian-English bilinguals and studying Serbo-Croatian speakers is that Serbo-Croatian speakers are monolingual and therefore, by definition, have an integrated lexicon.

In the present experiment, a modified version of the Picture-Word Interference (PWI) paradigm, adapted for use with an eye-tracker, was used. The PWI paradigm consists of presenting participants with a picture that also contains a written word. Participants have to name the picture while ignoring the word; reaction times are recorded. Multiple studies suggest that picture naming latencies vary as a function of the relation between a picture and a distracter word (e.g., Caramazza & Costa, 2000, 2001; Deschneiak & Schriefers, 2001; LaHeij & van den Hof, 1995; Rayner & Springer, 1986; Schriefers & Meyer, 1990). For example, semantically related words interfere more than semantically unrelated words (e.g., Levelt, Schriefers, Vorberg, Meyer, Pechmann, & Havinga, 1991; Starreveld & LaHeij, 1996, 1995). The surface form of the distractor word also influences picture naming (e.g., Meyer & Schriefers, 1991), with phonological similarities facilitating picture naming (e.g., Descheniak & Schriefers, 2001). Performance on the bilingual PWI task has been examined both for semantically and phonologically related items (e.g., Costa & Caramazza, 1999; Costa, Miozzo, & Caramazza, 1999) and was found to be vulnerable to semantic interference from a non-target language (e.g., Ehri & Buchard-Ryan, 1980; Hermans, Bongaerts, de Bot, & Schreuder, 1998), but the effect was mediated by degree of proficiency (e.g., Goodman, Haith, Guttentag, & Rao, 1985), by similarity of the two languages, and by response language (for a review, see Smith, 1997). The modification of the PWI task for use with eye-tracking consists of presenting the written word in a different quadrant of the visual display (as opposed to within the picture). The technique was piloted with monolingual English and bilingual Russian-English speakers and confirmed that interference effects persisted.

In sum, the present experiment examined parallel activation of both languages during bilingual written word recognition in monolingual settings and extended the study of parallel activation during bilingual written word recognition to languages with partial overlap in orthography

and orthography-to-phonology mappings. We predicted that, compared to English monolinguals, Russian-English bilinguals naming pictures in English would make more eye movements to written stimuli that are semantically unrelated to the picture in English, but are related to it in Russian, thus suggesting that processing of written input is not limited to the target language, but that the visual stimulus is also mapped onto non-target language orthography, even when orthographic representations are associated with different phonological forms in the two languages.

Methods

Design

The study followed a 2 x 2 mixed factorial design, with group (bilingual vs. monolingual) as the between-subject factor and condition (control vs. Russian words) as the within-subject factor. All participants were tested in English only. In the first condition, the picture and the distracter word were semantically unrelated if the written stimulus was mapped onto either language (e.g., picture of a *palmtree*, written stimulus HOCTA). In the second condition, the picture and the distracter were unrelated if the written word was mapped onto English orthography only, but semantically related if it was also mapped onto Russian orthography (e.g., picture of a palmtree, word stimulus is COCHA. In English COCHA is a non-word, while in Russian it is the orthographic representation of the word *pinetree* and is pronounced *sasna*.) In the second condition, semantic interference is present if the written stimulus is mapped in parallel not only onto English orthography, but also onto Russian orthography, therefore activating the other language in a bottom-up manner. Proportion of eye movements to distracter words and reaction times for picture naming were measured.

Participants

Fifteen Russian-English bilinguals (mean age=25 years) and 15 English monolinguals (mean age=21 years) were tested. Russian-English bilinguals and monolingual English speakers were recruited among undergraduate and graduate students at Northwestern University and via personal contacts. All participants were paid for their participation.

In addition to picture naming, all participants were administered the Language Experience and Bilingual Status (LEABS) Questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2003) for self–reported measures of language preference, proficiency, acquisition history, and current exposure. All bilinguals were fluent in both languages and were not enrolled in ESL classes. Measures of language experience collected via self-reports were included in analyses of covariance.

Stimuli

Twenty-four stimulus sets were generated for the two conditions. Across conditions, stimuli were controlled for length and bigram frequency. For English frequencies, the CELEX database was used (Baayen, Piepenbrock, & Van Rijn, 1993). For Russian frequencies, the new Frequency Dictionary at the Russian Research Institute of Artificial Intelligence (www.artint.ru/projects/frqlist/frqlist-en.asp; Sharoff, 2002) was used. Stimuli consisted of black line drawings and were generated using the IMSI Masterclips database and original artwork, and were altered in Adobe Photoshop. In order to meaningfully monitor eye movements, the locations of the target picture and the written word on the display were varied across four possible quadrants (top left, top right, bottom left, bottom right).

Procedure

All participants were tested in English. The bilingual speakers were tested in one language only so as to prevent overt activation of the other language during the experiment. Participants were asked to label pictures presented on a Mac computer display (G4 dual-processor computer) using Superlab software. Their verbal responses were recorded using a microphone. Eye movements during the experiment were recorded using an ISCAN head-mounted eye-tracker. The head-mounted eye-tracker consisted of a baseball-like cap, with two small cameras attached to the visor. One camera recorded the participant's field of view, and the other camera recorded the participant's eye movements; the outputs from the two cameras were superimposed and recorded using a digital recorder and were later analyzed using FinalCutPro software with frame-by-frame audio/video playback.

Results

The proportions of eye movements to distracter words during picture naming were analyzed with a 2 x 2 Analysis of Covariance, with group (monolingual or bilingual) and condition (control or Russian word) as the two independent variables and language preference when reading as covariate. Results revealed a main effect of group, $F(1, 27)=5.06$, $p<0.05$ and a significant interaction between condition and group, $F(1, 27)=4.78$, $p<0.05$. Overall, bilinguals made more eye movements to distracter words than monolinguals, 56% vs 34%. However, post-hoc analyses suggest that this difference was larger in the Russian words condition (61% for bilinguals vs 32% for monolinguals), where the two groups were significantly different from each other ($F(1,27)=7.77$, $p<0.01$), than in the control condition (50% for bilinguals vs 37% for monolinguals), where the two groups did not differ significantly ($F(1,27)=1.92$, $p>0.1$). A similar 2x2 Ancova on reaction time data did not reveal any significant effect of condition ($F=1.16$, $p>0.1$), group ($F=1.33$, $p>0.1$), or interaction between the two ($F=0.21$, $p>0.1$).

Based on self-reports collected with the LEABS Questionnaire, Russian-English bilinguals were grouped into 2 types, one consisting of bilinguals who preferred to read in English and one consisting of bilinguals who preferred to read in Russian. The proportion of looks made

by the two types of bilinguals to Russian competitor words and to control stimuli were compared in two ways. First, we examined whether or not the two types of bilinguals looked at the written stimuli at all (individual trials were coded with a 0 if participants did not look at the written stimulus and with a 1 if they did). Next, we examined the number of times the participants looked at a written stimulus (individual trials were coded with a 0 if participants did not look at the written stimulus, with a 1 if they looked at it once, with a 2 if they looked at it twice, with a 3 if they looked at it three times, and so on). Results of the first comparison did not reveal any significant differences, suggesting that the two types of bilinguals were just as likely to fixate distracter Russian words. Results of the second comparison revealed a marginally significant interaction between condition and type of bilinguals, $F(1, 13)=4.43$, $p<0.06$. Bilinguals who preferred to read in English (N=9) looked at Russian words more often (60%) than at control stimuli (48%); this pattern was not observed for bilinguals (N=6) who preferred to read in Russian (47% to Russian words and 52% to control stimuli).

Discussion

The present experiment examined mapping of visual input onto orthographic representations and lexical forms in bilinguals whose two languages share partial orthographic overlap. Results suggest that Russian-English bilinguals, when presented with written language input in a monolingual English setting, map the visual stimulus onto orthographic representations in both languages. As a result, lexical representations in the non-target language become activated, as evidenced by more eye movements to Russian distracter words relative to bigram-matched control stimuli in bilinguals, but not in monolinguals. These findings contribute to the existing body of literature suggesting simultaneous activation of a bilingual's two languages during written language processing and extend them to speakers whose two languages overlap orthographically only partially.

Absence of differences in picture-naming reaction times suggests that the simultaneous activation of non-target language orthography and lexicon (as demonstrated by eye movement data) did not minimize the efficiency with which bilinguals accomplished a language production task in the target language. This finding is consistent with recent accounts of optimal speed/accuracy outcomes achieved by parallel interactive models of language processing.

The finding that preferred reading language influenced the degree of activation of a bilingual's other language (as indicated by number of times bilinguals looked at Russian words) points to the importance of carefully assessing bilinguals' experience and proficiency in the two languages. Proficiency understanding, speaking, reading, and writing in the two languages, as well as factors pertaining to acquisition and current use of the two languages are just some of the variables to be taken into account when testing language processing in bilinguals (Marian, to appear). In our study, bilinguals who preferred reading in their second language (English) were just as likely to fixate Russian words as their peers who preferred reading in Russian, but once a written stimulus drew their eye movements, bilinguals who preferred to read in English were more likely to look at the Russian word again, fixating it repeatedly. These results suggest that, while the non-target language was activated in both types of bilinguals, processing the written input and possibly accessing its lexical representation was more effortful for those bilinguals whose preferred reading language was English.

It is notable that activation of the non-target language occurred in spite of the differential orthography-to-phonology mappings associated with the two languages. Although these results suggest simultaneous mapping onto the orthography and lexicon of the non-target language, they do not provide information about activation of the non-target language phonology, since lexical activation of the non-target language during orthographic processing may or may not have included phonological activation (i.e., mapping to the lexicon may have been directly from orthographic representations, bypassing phonology). To examine activation of non-target language phonology during written language processing, a mirror image of the present experiment is required. Namely, while the present experiment tested activation of non-target language *orthography* by examining processing of input that overlapped *orthographically, but not phonologically*, a separate experiment tested activation of non-target language *phonology* by examining processing of input that overlapped *phonologically, but not orthographically* (Kaushanskaya & Marian, 2004).

Moreover, for a more comprehensive understanding of parallel activation of both languages in Russian-English bilinguals during written language processing, a closer look at semantic processing is necessary. In the present study, stimuli consisted of English non-words, so as to avoid word frequency confounds across languages. Future work needs to expand this paradigm to processing written stimuli that constitute words in both languages. Whether the written stimuli are English words or nonsense strings is likely to influence the strength of competition from Russian semantically-related words. Furthermore, when the written stimulus is a word in each language, word frequencies in the two languages are likely to influence the effect, with higher-frequency mappings resulting in faster activation. Finally, future efforts will also focus on examining input properties that are likely to influence the relative activation of the non-target language during written word recognition, such as amount of orthographic and phonemic overlap across the two languages.

The proposed project has implications for understanding language development and processing in bilinguals, and reading and acquisition of literacy in bilinguals, with potential implications for bilingual education and for assessment of bilinguals. For instance, understanding written word recognition in bilinguals who are in

monolingual contexts may have implications for bilingual children entering mainstream (not ESOL) classrooms. The results of the 2000 Census indicate that 18% of American households speak a language other than English at home and that this proportion is increasing. Understanding how bilingual status influences cognitive and linguistic functioning may have direct implications for this linguistically diverse and severely under-served segment of the population. Beyond bilingual language processing, this research will contribute to advancing the understanding of language processing in general, including written word recognition and spoken word production.

Acknowledgments

This study was supported by Northwestern University junior faculty start-up funds to the first author and by a Graduate Research Grant to the second author. We thank Karla McGregor, Doris Johnson, Henrike Blumenfeld, Caitlin Fausey, and the members of the Bilingualism and Psycholinguistics Research lab for helpful discussions during the course of this work.

References

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419-439.

American Library Association and the Library of Congress. (1991). *ALA-LC Romanization tables: Transliteration schemes for non-Roman scripts*. Library of Congress, Washington D. C.

Baayen, H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

Beauvillain, C., & Grainger, J. (1987). Accessing interlexical homographs: Some limitations of a language-selective access. *Journal of Memory and Language, 26*, 658-672.

Besner, D., & Hildebrandt, N. (1987). Othographic and phonological codes in the oral reading of Japanese Kana. *Journal of Expiremental Psychology: Learning, Memory, and Cognition, 13,* 335-343.

Bijeljac-Babic, R., Biardeau, A., & Grainger, J. (1997). Masked orthographic priming in bilingual word recognition. *Memory & Cognition, 25*, 447-457.

Bowers, J. S., Mimouni, Z., & Arguin, M. (2000). Orthography plays a critical role in cognate priming: Evidence from French/English and Arabic/French cognates. *Memory and Cognition, 28,* 1289-1296.

Brown, H., Sharma, N. K., & Kirsner, K. (1984). The role f script and phonology in lexical representation. *The Quarterly Journal of Expiremental Psychology. 36A,* 491-505.

Brysbaert, M., Van Dyck, G., & Van de Poel, M. (1999). Visual word recognition in bilinguals: Evidence from masked phonological priming. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1, 137-148.

Caramazza, A., & Costa, A. (2000). The semantic interference effect in the picture-word interference paradigm: Does response set matter? *Cognition, 75*, B51-B64.

Caramazza, A., & Costa, A. (2001). Set size and repetition in the picture-word interference paradigm: Implications for models of naming. *Cognition, 80*, 291-298.

Chen, H. C., & Tsoi, K. C. (1990). Symbol-word interference in Chinese and English. *Acta Psycholigica, 75,* 123-138.

Colome, A. (2001). Lexical activation in bilinguals' speech production: Language-specific or language independent? *Journal of Memory and Language, 45,* 4, 721-736.

Costa, A., & Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? Further evidence from Spanish-English and English-Spanish bilinguals. *Bilingualism: Language and Cognition, 2,* 231-244.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection. *Journal of Memory and Language, 41*, 365-397.

Cutler, A. (1995). Spoken word recognition and production. In J. Miller & P. Eimas (eds.), *Handbook of cognition and perception* (pp. 97-136). New York: Academic Press.

DeGroot, A. M. B., Delmaar, P., & Lupker, S.J. (2000). The processing of interlexical homographs in translation recognition and lexical decision: Support for non-selective access to bilingual memory. *The Quarterly Journal of Experimental Psychology, 53A,* 397-428.

Descheniak, J. D., & Schriefers, H. (2001). Priming effects from phonologically related distractors in picture-word interference. *The Quarterly Journal of Experimental Psychology, 54A,* 371-382.

Dijkstra, T., Grainger, J., & van Heuven, W. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and Language, 41*, 496-518.

Doctor, E. A., & Klein, D. (1992). Phonological processing in bilingual word recognition. In R. J. Harris (Ed.), *Cognitive Processing in Bilinguals* (pp. 237-252). Amsterdam: Elsevier.

Ehri, L. C., & Bouchard Ryan, E. (1980). Performance of bilinguals on a picture-word interference task. *Journal of Psycholinguistic Research, 9, 3*, 285-302.

Feldman, L. B., & Turvey, M. T. (1983). Word recognition in Serbo-Croation is a phonologically analytic. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 288-298.

Goodman, G. S., Haith, M. M., Guttentag, R. E., & Rao, S. (1985). Automatic processing of word meaning; intralingual and interlingual interference. *Child Development, 56*, 103-118.

Grainger, J. (1993). Visual word recognition in bilinguals. In R. Schreuder & B. Weltens (eds.), *The Bilingual Lexicon* (pp. 11-26). Amsterdam: John Benjamins.

Grainger, J., & Dijkstra, A. (1992). On the representation and use of language information in bilinguals. In R. J. Harris (ed.), *Cognitive Processing in Bilinguals* (pp. 207-220). Amsterdam: Elsevier.

Hermans, D., Bongaerts, T., de Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition, 1*, 213-229.

Kaushanskaya, M. & Marian, V. (2004). Activation of non-target language phonology during bilingual visual word recognition: Evidence from eye-tracking. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum.

La Heij, W., & van den Hof, E. (1995). Picture-word interference increases with target-set size. *Journal of Psychological Research, 58*, 119-133.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann,T. H., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review, 98*, 122-142.

Li, P. (1996). Spoken word recognition of code-switched words by Chinese-English bilinguals. *Journal of Memory and Language, 35*, 757-774.

Lukatela, G., Savic, M., Gligorijevic, B., Ognjenovic, P., & Turvey, M.. T. (1978). Bi-alphabetical lexical decision. *Language and Speech,21,* 142-165.

MacNamara, J., & Kushnir, S. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior, 10,* 480-487.

Marian, V. (to appear). Bilingual research methods. In J. Altarriba, & R. R. Heredia (Eds.), *An Introduction to Bilingualism: Principles and Processes.* Mahwah, NJ: Laurence Erlbaum.

Marian, V., Blumenfeld, H., Garstecki, D., Kaushanskaya, M., Fausey, C., & Lu, D. (2003, May). Developing a tool for assessing Language Experience and Bilingual Status (LEABS). Poster presented at the annual meeting of the *Midwestern Psychological Association*, Chicago, IL. Manuscript in preparation.

Marian, V., & Spivey, M. (2003a). Comparing Bilingual and Monolingual Processing of Competing Lexical Items. *Applied Psycholinguistics, 24,* 2, 173-193.

Marian, V., & Spivey, M. (2003b). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6*, 1-19.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*, 71-102.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Nas, G. (1983). Visual word recognition in bilinguals: Evidence for a cooperation between visual and sound based codes during access to a common lexical store. *Journal of Verbal Learning and Verbal Behavior, 22*, 526-534.

Rayner, K., & Springer, C. J. (1986). Graphemic and semantic similarity effects in the picture-word interference task. *British Journal of Psychology, 77*, 207-222.

Schriefers, H. & Meyer, A. S. (1990). Experimental note: Cross-modal, visual-auditory picture-word interference. *Bulletin of the Psychonomic Society, 28,* (5), 418-420.

Sharoff, S. (2002). Meaning as use: Exploitation of aligned corpora for the contrastive study of lexical semantics. Proceedings of Language Resources and Evaluation Conference (LREC02). May, 2002. Las Palmas, Spain.

Smith, M.. C. (1997). How do bilinguals access lexical information. In DeGroot, A. M. B., & Kroll, J. F. (Eds.), *Tutorials in Bilingualism: Psycholinguistic Perspectives* (pp. 145-168). Mahwah, NJ: Lawrence Erlbaum.

Smith, M.C., & Kirsner, K. (1982). Language and orthography as irrelevant features in colour-word and picture-word stroop interference. *Quarterly Journal of Experimental Psychology, 34A,* 153-170.

Soares, C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory and Cognition, 12*, 380-386.

Spivey, M., & Marian, V. (1999). Crosstalk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10,* 3, 281-284.

Starreveld, P. A., & LaHeij, W. (1995). Semantic interference, orthographic facilitation, and their interaction in naming task. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 686-698.

Starreveld, P. A., & LaHeij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 896-918.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268*, 1632-1634.

Tzelgov, J., Henik, A., Sneg, R., & Baruch, O. (1996). Unintentional word reading via the phonological route: The Stroop effect with cross-script homophones. *Journal of Expiremental Pschology: Learning, Memory, and Cognition, 22,* 336-349.

Van Heuven, W. J. B., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language, 39*, 458-483.

Wong, K.F.E., & Chen, H-C. (1999). Orthographic and phonological processing in reading Chinese text: Evidence from eye fixations. *Language and Cognitive Processes*, *14*, 461-480.

www.census.gov

www.artint.ru/projects/frqlist/frqlist-en.asp

The 2004 CogSci proceedings publication

Martin, B. A., Lozano, S. C., and Tversky, B. Detecting Goal Structure Facilitates Learning. Table of contents:897-902 Actual pages: 945-950

has been retracted.

All authors retract this article. Bridgette Martin Hard and Barbara Tversky believe that the research results cannot be relied upon; Sandra C. Lozano takes full responsibility for the need to retract this article.

An ACT-R Modeling Framework for Interleaving Templates of Human Behavior

Michael Matessa (mmatessa@arc.nasa.gov)
NASA Ames Research Center, Mail Stop 262-4
Moffett Field, CA 94035 USA

Abstract

Performance modeling has been made easier by architectures which package psychological theory for reuse at different levels. Both CPM-GOMS, which packages theory at the task level, and ACT-R, which packages theory at the lower level of rules for perceptual-motor interaction, have been shown to be useful. This paper describes ACT-Stitch, a framework for translating CPM-GOMS templates and interleaving theory into ACT-R. The research involved in producing ACT-Stitch will benefit reusable template research by showing how to implement templates and interleaving in a new architecture that processes resource information. ACT-R research will benefit from re-usable productions packaged at a higher task level and from the multi-tasking control structure used that allows ACT-R to interleave productions from different templates. The zero-parameter predictions of ACT-Stitch are empirically validated.

Introduction

Predicting well-practiced human performance in human-computer interaction (HCI) domains by means of computer modeling is a valuable but difficult process. For example, modeling has been used to predict the outcome of a test of new computer workstations, saving a telephone company millions of dollars per year (Gray, John & Atwood, 1993), but much of the modeling was done by hand.

For accurate predictions, a large amount of psychological theory needs to be applied. Several modeling architectures have been developed to make modeling easier by packaging this theory for reuse. CPM-GOMS (John, 1988; 1990) uses templates of behavior to package at a task level (e.g., mouse move-click, typing) predictions of lower-level cognitive, perceptual, and motor resource use. These templates are interleaved to reflect the ability of skilled people to perform parts of one task in parallel with another. For example, an eye-movement study has demonstrated interleaving in a hand-washing task -- while people perform the subtask of first getting their hands wet they interleave a look to the soap dispenser before performing the motor actions in the subtask of soaping their hands (Pelz & Canosa, 2001). The CPM-GOMS theory has been automated (John, et al., 2002) in a computational architecture that schedules blocks of abstract resource use (Freed et al., 2003). ACT-R (Anderson & Lebiere, 1998; Anderson et al., submitted) uses a computational production system architecture for packaging knowledge at the lower level of rules for working with cognitive and perceptual information and motor actions. In contrast with CPM-GOMS, the ACT-R system can interact with an environment to perceive objects and manipulate them. However, ACT-R does not have a built-in theory of multi-tasking which would interleave tasks, although some work has been done in modeling multi-tasking in the ACT-R architecture (Byrne & Anderson, 2001; Lee & Taatgen, 2002; Salvucci, 2002).

This paper presents a new framework, ACT-Stitch, which combines the usefulness of modeling at the task level with the process theory of a lower-level cognitive architecture. It uses a process of macro-compilation similar to that used by Salvucci and Lee (2003) to translate CPM-GOMS templates into ACT-R productions. Their system will be compared to the current system in the discussion section, but one difference is that their system models at the level of KLM-GOMS, which does not interleave cognitive operators (John & Kieras, 1996). The control structure used by ACT-Stitch to achieve the interleaving of cognitive operators from different templates is one of the major contributions of this paper. The research involved in producing ACT-Stitch will benefit reusable template research by showing what aspects of template and interleaving theory are important in a new architecture that processes resource information. ACT-R research will benefit from re-usable productions packaged at a higher task level and from the multi-tasking control structure used that allows ACT-R to interleave productions from different templates.

Templates

Templates are building blocks of human behavior containing a detailed theory of cognitive, perceptual, and motor behaviors. They are beneficial for modelers because they package this theory at the task level and can be reused in different applications (Matessa et al., 2002). Even behavior as simple as a mouse move and click requires coordination of the use of cognitive, perceptual, and motor resources, as Figure 1 shows in PERT chart form with boxes representing resource use and lines indicating dependencies. The template was developed for the simple task of clicking on lit circles by Gray and Boehm-Davis (2000), but has been successfully reused for clicking to operate a simulated automated teller machine (John, et al., 2002).

Templates require a theory of interleaving to reflect the ability of skilled people to perform operations from different tasks in parallel. When CPM-GOMS was first developed, this interleaving was done by hand, with modelers applying their knowledge of the psychology involved. John et al. (2002) codified this knowledge and implemented automated interleaving in a system that scheduled blocks of abstract resource use. Results from this work were used in the construction of ACT-Stitch templates that produce productions which ACT-R can interleave.

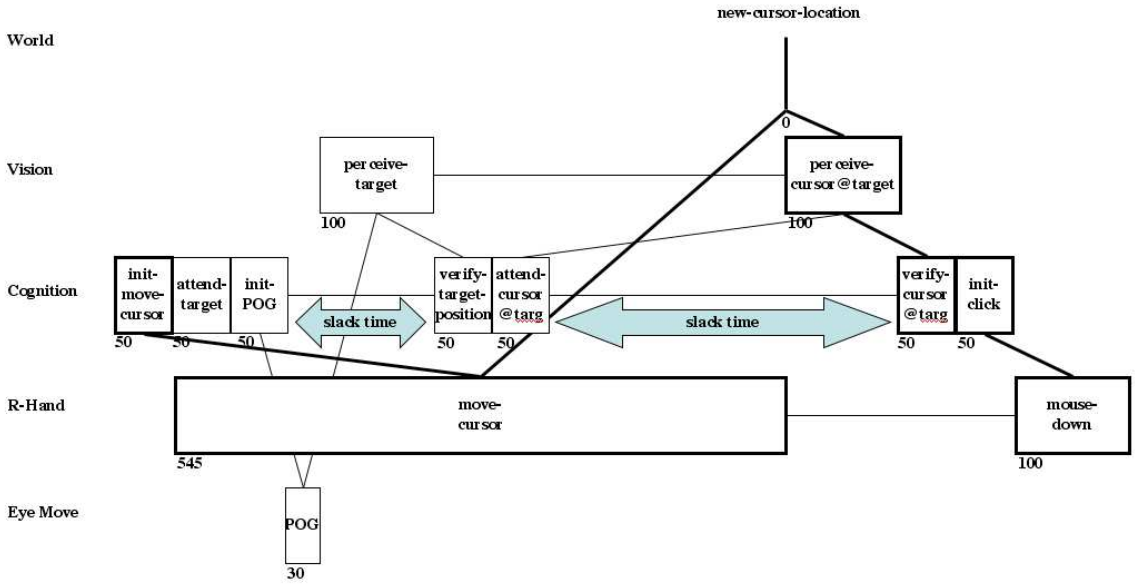


Figure 1: A template of carefully moving the cursor to a target and clicking the mouse (adapted from Gray and Boehm-Davis, 2000).

Macro-Compilation

ACT-Stitch uses a process of macro-compilation to translate CPM-GOMS templates of human behavior into ACT-R productions. More specifically, cognitive operators are translated into productions with ACT-R perceptual-motor commands that represent CPM-GOMS perceptual-motor operators. Productions also contain a control structure that allows ACT-R to implement CPM-GOMS interleaving and have productions from one template execute during the execution of productions from another template. This differs from the ACT-Simple system (Salvucci & Lee, 2003) that compiled a sequence of KLM-GOMS tasks into a series of productions which were controlled by an incrementing state counter.

Macro-compilation should not to be confused with ACT-R production compilation in which two productions are translated into another more efficient production. Salvucci and Lee (2003) argue that macro-compilation facilitates theoretical consistency, inheritance of architectural features, model integration, and model refinement. Theoretical consistency is maintained by having the higher task-level template share a consistent representation with the lower-level ACT-R architecture. The macro-compiled template inherits parameters and limitations that increase psychological plausibility as well as a framework for learning, showing individual differences, and making errors. Model integration is helped by providing a common language where models from different domains can interact.

ACT-Stitch Framework

To understand how ACT-Stitch works, this section will first explain the process of how a modeler uses ACT-Stitch, then describe the ACT-R architecture, then go into more detail about macro-compilation and production execution, and finally give an example of macro-compiled productions.

ACT-Stitch modeling

ACT-Stitch currently has two templates implemented, Slow-Move-Click and Fast-Move-Click, based on templates from Gray and Boehm-Davis (2000). For Gray and Boehm-Davis, Slow-Move-Click represented the selection of a target when there is uncertainty about where the target appears in each trial. Fast-Move-Click represented the selection of a target at a known location. and skipped the verification of the cursor being at the target. These templates were reused by John et al. (2002) in modeling interactions with a simulated automated teller machine. There, Slow-Move-Click represented the selection of difficult targets at far distances, requiring more careful verification of target and cursor location before clicking than the selection of easier targets, which are represented with Fast-Move-Click.

To use ACT-Stitch, the modeler creates two lists, one for target objects and one for a task sequence. The target object list contains target names, positions, and sizes. The task sequence list contains template/target pairs. The system then creates an environment including target objects and macro-compiles templates into productions. The ACT-R system is then run, and information about resource use and

dependencies is automatically stored. This information can be exported to a PERT chart viewing program.

ACT-R

ACT-R (Anderson & Lebiere, 1998; Anderson et al., submitted) is a computational theory of human cognition incorporating both declarative knowledge (e.g., addition facts) and procedural knowledge (e.g., the process of solving a multi-column addition problem) into a production system where procedural rules act on declarative chunks. Chunks are made up of slots containing information, and production rules which match the information in chunk slots are able to execute. The goal chunk represents the current intentions. The ACT-R system includes the capability for modelers to create simulated environments, such as screen interfaces. Production rules have the ability to interact with this environment by perceiving objects and making motor movements through perceptual and motor buffers. With this interaction, ACT-R can make use of Fitts' Law to make predictions of movement time based on distance to target and target size.

ACT-Stitch production creation

CPM-GOMS templates contain predictions of cognitive, perceptual, and motor behavior. When translating a template into ACT-R productions, each cognitive operator in a template corresponds to a production in ACT-R. Cognitive operators and productions are both predicted to take 50 ms to perform by each theory. Both theories predict parallel execution of cognitive, perceptual, and motor processes. In CPM-GOMS, each perceptual and motor operator requires an initiation by a cognitive operator. This corresponds to the ACT-R requirement of productions to initiate vision and motor processes. To move visual attention to a new location and perceive an object, CPM-GOMS predicts that it takes 30 ms to move attention plus some time for perception, while ACT-R predicts that is takes 85 ms to move attention with no additional time for perception. For mouse movement, CPM-GOMS predicts an execution time calculated by Fitts' Law, while ACT-R predicts a 200 ms preparation time plus a time calculated by Fitts' Law plus a 50 ms finish time. For mouse clicks, CPM-GOMS predicts a 100 ms mouse down time plus a 100 ms mouse up time, while ACT-R predicts a 150 ms preparation time plus a 60 ms execution time plus a 90 ms finish time. ACT-R can perform motor preparations in parallel with the motor executions and finishes of previous motor commands, and ACT-Stitch creates productions that take advantage of this capability.

ACT-Stitch creates a set of productions for each template/target pair in the task list, and the productions created from macro-compilation must insure proper sequencing of motor actions, insure the ability to allow the correct productions in future templates to interleave during the execution of productions in the current template, and insure the ability to block the incorrect productions in future templates from interleaving with productions in the current template.

These three requirements are accomplished in productions by using information in the current goal as well as perceptual-motor buffers. Slots in the goal are created for the vision and hand resources for both the intended action and target making use of the resource. This makes four slots in the goal: vision action, vision target, hand action, and hand target. To insure proper sequencing, the action slots in productions of the current template are filled with an intended action appended with the unique number of the current template. Also, the target slots are filled with an intended target. The intended action cannot be used alone since without the template number no sequence information would be stored. The template number cannot be used alone since there may be multiple actions in the same template using the same resource (e.g., mouse move and click). The intended target cannot be used alone since sequence information would be lost if a target appears twice in a sequence (e.g., clicking the same number twice). The intended target cannot be ignored since the same action could be used in a template for two targets (e.g., verify target and verify cursor).

To insure the ability to interleave productions, separate action slots are used for each resource (vision and hand). This allows, for example, a procedure to initiate a vision action from a future template before a procedure initiates a hand action from the current template. To insure the ability to block productions from future templates, the action slots are filled with intended actions appended with the current template number. This prevents, for example, moving to the next target while the hand resource is free between moving to the current target and clicking on the current target. The template number cannot be contained in a separate goal slot because that would not allow productions from the next template to execute before the productions of the current template have finished.

Perceptual-motor buffers are also used in sequencing. Productions that interact with the perceptual-motor buffers check to make sure the buffers are free before using them. Also, the task logic of perception and action makes use of buffers to order productions. For example, the process of verifying a target position before clicking requires filling the visual location buffer with the location of the intended target, then filling the visual object buffer with the object found at that location, and then making a mouse click through the motor buffer.

These goal slots and buffers could be extended to include resources such as a left hand and buffers such as memory retrieval in future template development.

ACT-Stitch production execution

The ACT-R system is initialized with the goal containing the actions and targets of the first template. ACT-R selects productions to execute based on the state of the goal and perceptual-motor buffers. Productions make calls to the perceptual-motor system which has assumptions for how long the resources are used. Slack time corresponds to the time a resource is available during procedure execution. A production that is created from the next template can execute (even if all the productions made from the current

template are not finished executing) when it matches values in the action and target goal slots. Action slots contain intended actions appended with unique template numbers, and target slots contain intended targets. When a resource is no longer needed by a template, a production in the template will fill the action slot with the next intended action appended with the next template number, and the target slot will be filled with the next intended target.

Within-template dependencies are implemented by productions waiting for action and target slots to be filled in the goal and for resources to be available. Template productions are created so that a production will change the contents of action and target slots appropriately. A production (A) from a future template that is waiting for another production (B) in that template to change the contents of action and target slots cannot execute during the execution of productions in the current template until production B is executed.

Relationships across templates are established the same way as within templates, using action and target slots in the goal. Values in these slots allow the blocking of productions that would use resources even if the resource is free.

Example ACT-Stitch productions

To get an idea of what a template looks like after being macro-compiled into ACT-R productions, the following shows pseudo-code for the Fast-Move-Click template. Each instance of a template in the task sequence list would have its own set of productions labeled by the position of the template in the list (x).

```
Tx-Init-Move-Cursor
IF
        right hand action goal is to move in this template
        right hand target goal is this template's object
        motor preparations have completed
THEN
        move cursor
        empty right hand target goal
        set right hand action goal to click in this template

Tx-Attend-Targ
IF
        vision action goal is to attend target in this template
        vision target goal is this template's object
        visual location and object buffers are empty
        vision is available
THEN
        fill visual location buffer with location where
           this template's object should be

Tx-Init-Eye-Move
IF
        vision action goal is to attend target in this template
        vision target goal is this template's object
        visual object buffer is empty
        visual location buffer holds object location
 THEN
        fill visual object buffer with object at location
        empty visual location buffer

Tx-Verify-Targ-Pos
IF
        vision action goal is to attend target in this template
        vision target goal is this template's object
        right hand target goal is empty
        visual object buffer holds object at location y
        location y is the expected location of this template's object
THEN
        empty visual object buffer
        set visual action goal to attend in the next template
        set visual target goal to next template's object
        set right hand target goal to this template's object

Tx-Init-Click
IF
        right hand action goal is to click in this template
        right hand target goal is this template's object
        motor preparations have completed
THEN
        click mouse
        set right hand action goal to move in next template
        set right hand target goal to next template's object
```

Productions that initiate motor movements (Init-Move-Cursor and Init-Click) first check that the motor preparations from previous motor movements have completed Since motor preparations can happen in parallel with motor executions and finishes in ACT-R, this means that preparations can start during previous executions and finishes. Productions could be written to wait for the previous executions and finishes to complete before starting preparations, but they would not be as efficient.

Empirical Validation

ACT-Stitch was applied to the ATM task used by John et al. (2002) to test their automation of CPM-GOMS. The task was to make an $80 withdraw from a checking account on a simulation of an automated teller machine. Users interacted with the ATM by using a mouse to click on simulated keys or slots. The users were instructed to follow the following steps:

Insert card (click on the card slot)
Enter PIN (click on the 4, 9, 0, and 1 keys in turn)
Press OK (click on the OK button)
Select transaction type (click on the withdraw button)
Select account (click on the checking button)
Enter amount (click on the 8 and 0 keys)
Select correct/not correct (click on the correct button)
Take cash (click on the cash slot)
Select another transaction (click on the No button)
Take card (click on the card slot)
Take receipt (click on the cash slot)

This task was repeated 200 times by the users, and results were analyzed using the means of trials 51-100. This level of practice is comparable to that used by both Card, Moran, and Newell (1983) in a text editing task and Baskin and John (1998) in a CAD drawing task when they explored the effects of extensive practice on match to various GOMS models. As in John et al. (2002), Slow-Move-Click templates were used for targets that were difficult to select because of size and distance (e.g. the thin card slot) and Fast-Move-Click templates were used for easier targets (e.g. keypad keys).

Figure 2 compares ACT-Stitch predictions of mouse click times to average subject mouse click times of trials 51-100. The results are highly correlated (r=.96) with a low average absolute difference of 62 ms.

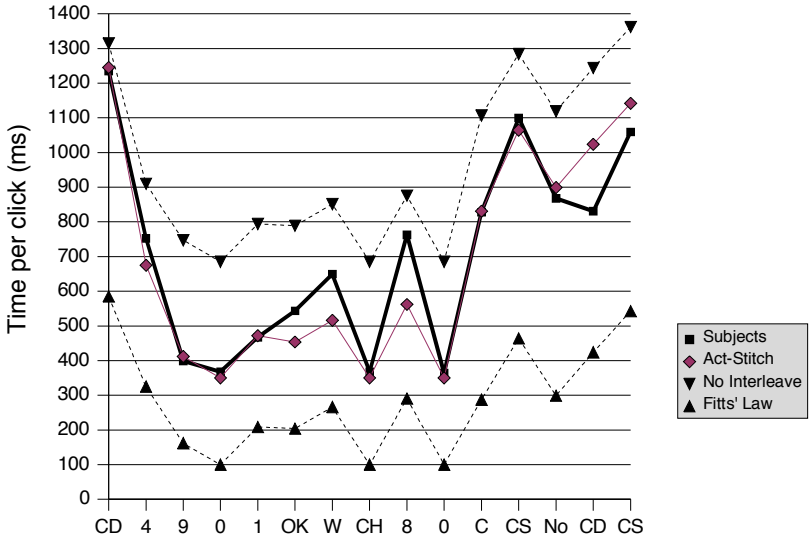


Figure 2: Average subject performance compared to ACT-Stitch predictions, ACT-Stitch predictions with no interleaving, and Fitts' Law predictions.

Figure 2 also shows the value of cognitive modeling over a Fitts' Law only prediction and the value of ACT-Stitch interleaving. A Fitts' Law prediction has a high correlation with subject performance (r=.97) but predicts faster performance, with an average absolute difference of 416 ms. A version of ACT-Stitch was created that did not interleave template productions, and while the correlation with subject performance was still high (r=.95), the predictions are too slow (average absolute distance = 257 ms).

The effect of interleaving on resource use is shown in PERT chart form in Figure 3. This output is from the Sherpa visualization tool developed by John et al. (2002) in their work to automate CPM-GOMS. The top row shows vision resource use, the second shows cognition, the third shows motor preparation, and the bottom shows motor execution and finishing. Resource use is indicated with shaded boxes, and instances of resource use in the same template are shown with the same shade of gray. The figure shows how cognitive, perceptual, and motor resources are interleaved between templates.

General Discussion

ACT-Stitch appears to be a useful framework for modeling the cognitive, perceptual, and motor processes involved in HCI tasks. With a simple description of an environment and task sequence, it is able to produce detailed, zero-parameter predictions that match well to human data.

ACT-Stitch has some similarities and differences with the ACT-Simple framework created by Salvucci and Lee (2003). They both use a process of macro-compilation to translate task-level descriptions of behavior into ACT-R productions, which give a detailed account of the cognitive, perceptual, and motor processes involved in the task. ACT-Stitch adds the ability to easily simulate simple environments, the ability for templates to interleave cognitive operators, and the ability to view resource use of the model with PERT chart tools. With the environment, models can take advantage of Fitts' Law to make detailed predictions of movement times. With a theory of interleaving that is based on fixed resources instead of spontaneous task demands, ACT-R modelers have the ability to start moving away from control theory based on simple chained productions. With PERT chart output, complex interactions of resource use in models can be understood easier.

CPM-GOMS is assumed to model skilled performance, and a CPM-GOMS model translated into ACT-R can be thought of as a state of performance after learning. With the ACT-R compilation process of learning more efficient productions, the whole learning curve from slow reading and remembering instructions to quick interleaving of resources can be studied. There has already been some start on this by Lee and Taatgen (2002), where they describe a model of performance on an air traffic controller task that at first has slow performance to due interpreting instructions, then speeds up due to production compilation creating more efficient productions, and eventually interleaves an optional step to look at wind conditions during multiple keystrokes.

In the ATM task, ACT-Stitch accounts for the data as well as CPM-GOMS automated in another system (see John et al., 2002), but it differs from that system in that it predicts a 200 ms motor preparation that occurs between the movement of attention and motor execution (see Figure 3). ACT-Stitch predicts that during this motor preparation time previous motor operations are taking place. This prediction could be tested with eye-tracking experiments.

This paper offers only a first step of a template and interleaving theory in ACT-R. Many more templates are needed to test the robustness of the representations used for the interleaving theory.

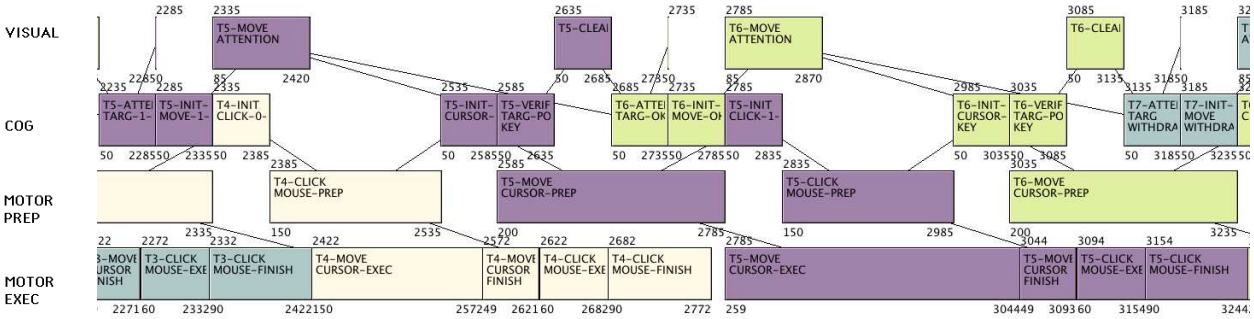


Figure 3: PERT chart of ACT-Stitch interleaving perceptual, cognitive, motor preparation, and motor execution and finishing resources

There are some interleaving abilities that the current framework cannot accomplish, for example, hovering a hand over a key for a key press that occurs in a template that is more than one template away in the future, or blocking an arbitrary combination of resources (such as both hands during typing) from interleaving. But this work is a first step to easier modeling and multi-tasking in ACT-R.

Acknowledgments

This work was supported by Office of Naval Research grant N00014-04-IP-2002 and by funds from the Airspace Operations System project of NASA's Airspace System program.

References

Anderson, J. R., Bothell, D., Byrne M.D. & Lebiere, C. (submitted to Psychological Review). *An Integrated Theory of the Mind*. Available from: http://act-r.psy.cmu.edu/papers/403/IntegratedTheory.pdf

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

Baskin, J. D., and John, B. E. (1998). Comparison of GOMS Analysis Methods. *Proceedings of ACM CHI 98 Conference on Human Factors in Computing Systems* (Summary) 1998 v.2 p.261-262.

Byrne, M. D., & Anderson, J. R. (2001). Serial modules in parallel: The psychological refractory period and perfect time-sharing. *Psychological Review*, 108, 847-869.

Card, S. K., Moran, T.P. & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Freed, M., Matessa, M., Remington, R. and Vera, A. (2003) How Apex automates CPM-GOMS. *Proceedings of the Fifth International Conference on Cognitive Modeling*, pp. 93-98. Bamberg, Germany:Universitats-Verlag.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, 6(4), 322-335.

Gray, W. D., John, B. E. & Atwood, M. E. (1993) Project Ernestine: Validating a GOMS Analysis for Predicting and Explaining Real-World Task Performance. *Human-Computer Interaction*, 8 (3), pp. 237-309.

John, B. E. (1988) *Contributions to Engineering Models of human-computer interaction*. Ph.D. Thesis. Carnegie Mellon University.

John, B. E. (1990) Extensions of GOMS analyses to expert performance requiring perception of dynamic visual and auditory information. *Proceedings of CHI*, 1990 (Seattle, Washington, April 30-May 4, 1990) ACM, New York, 107-115.

John, B. E. & Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: Comparison and Contrast. *ACM Transactions on Computer-Human Interaction*, 3 (4), pp. 320-351.

John, B. E., Vera, A. H., Matessa, M., Freed, M., and Remington, R. (2002) Automating CPM-GOMS. In *Proceedings of CHI 2002: Conference on Human Factors in Computing Systems* (pp. 147-154). ACM, New York.

Lee, F.J. & Taatgen, N.A. (2002). Multi-tasking as Skill Acquisition. *Proceedings of the twenty-fourth annual conference of the cognitive science society* (pp. 572-577). Mahwah, NJ: Erlbaum.

Matessa, M., Vera, A., John, B., Remington, R., & Freed, M. (2002). Reusable Templates in Human Performance Modeling. *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society* (pp. 649-654). Mahwah, NJ: Erlbaum.

Pelz, J. B. and Canosa, R. (2001). Oculomotor Behavior and Perceptual Strategies in Complex Tasks. *Vision Research*, 41, 3587-3596.

Salvucci, D. D. (2002). Modeling driver distraction from cognitive tasks. *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 792-797). Mahwah, NJ: Erlbaum.

Salvucci, D. D., & Lee., F. J. (2003). Simple cognitive modeling in a complex cognitive architecture. *Human Factors in Computing Systems: CHI 2003 Conference Proceedings*. New York: ACM Press.

Do eye movements go with fictive motion?

Teenie Matlock (tmatlock@psych.stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305 USA

Daniel C. Richardson (richardson@psych.stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305 USA

Abstract

Cognitive scientists interested in the link between language and visual experience have shown that linguistic input influences eye movements. Research in this area, however, tends to focus on literal language alone. In the current study, we investigate whether *figurative* language influences eye movements. In our experiment, participants viewed two-dimensional depictions of static spatial scenes while they heard either fictive motion sentences, such as *The palm trees run along the highway,* or non-fictive motion sentences, such as *The palm trees are next to the highway*. Overall, sentence type influenced participants' eye movements. Specifically, gaze duration on the figure (e.g., palm trees) was longer with fictive motion sentences than with non-fictive motion sentences. Our results demonstrate that figurative language influences visual experience. They provide further evidence that fictive motion processing includes mentally simulated motion.

Introduction

Imagine that you and a friend are sitting in a courtyard chatting. During the course of the conversation, you occasionally glance over at a long, thin stationary object on the ground. You assume the object is a tree branch or a walking stick until your friend says, "Oh! Look what slithered onto the courtyard." At that point, your perceptions and conceptions of the object dramatically change. The object goes from a piece of wood to a snake.

Situations like these—in which language influences the interpretation of objects and actions—are ubiquitous. The question addressed here is whether this influence is limited to *literal* language, or whether it also includes *figurative* language. We are especially interested in whether sentences such as *The road goes through the desert* or *The fence follows the coastline* (figurative because they include a motion verb but express no motion) affect eye movements. Our results suggest they do.

What We Know about Fictive Motion

Everyday language is replete with sentences such as (1a) and (1b). These are literal descriptions of static scenes.

(1a) *The road is in the desert*
(1b) *The fence is on the coastline*

Language is also full of sentences such as (2a) and (2b).

(2a) *The road goes through the desert*
(2b) *The fence follows the coastline*

These sentences are figurative because they contain a motion verb (e.g., *goes*, *follows*) but express no actual motion (Matlock, 2001).[1] They contrast with literal sentences with motion verbs, such as *The bus goes through the desert*, or *The herd of sheep follows the coastline*, which feature mobile agents that move from one point in space and time to another (Talmy, 1975; Miller & Johnson-Laird, 1976).

Despite the absence of actual movement with sentences such as (2a) and (2b), they have been claimed to involve *fictive motion*, an implicit mental simulation of "movement" through a construed scene (Talmy, 1983, 1996, 2000). On this view, the conceptualizer subjectively "scans" from one part of the scene to another, most notably, along the figure (i.e., prominent entity, subject noun phrase referent). For (2a), this means "moving" along the road, and for (2b), it means "moving" along the fence. According to the argument, fictive motion is a way to impose motion on what is otherwise a static scene. It enables the language user to compute information about the layout of a scene, for instance, a road in a desert in (2a), or a fence aligned with a coastline in (2b). Importantly, Talmy (1996) and other cognitive linguists do *not* maintain that fictive motion involves vivid imagery whereby the conceptualizer "sees" himself or herself (or any other animate entity) moving point by point along the figure in the scene being described. Instead, they take the motion to be relatively fleeting and tacit. (See also Langacker's *abstract motion*, 1986, 2000, and Matsumoto's *subjective motion*, 1996).[2]

At first, the claim that people simulate motion while processing descriptions of static scenes seems bizarre. Why would motion be processed, for instance, with sentences such as (2a) and (2b) when neither the road nor the fence is

[1] Like Rumelhart (1979) and Gibbs (1994), we do not maintain a hard and fast distinction between "literal" and "figurative". We simply use these terms here to operationalize two types of motion verb constructions: those that express motion and those that do not.
[2] Our study looks at just one type of fictive motion, Talmy's (2000) *co-extension path* fictive motion. There are many others.

capable of movement? Perhaps it is more reasonable to assume that such sentences are yoked to a purely static representation, as proposed by Jackendoff (2002). On this view, the representation underlying sentences such as (2a) and (2b) is static and atemporal. It is not unlike the representation underlying the literal sentences shown in (1a) and (1b), in which all points along the figure are activated simultaneously rather than incrementally.

Recent experimental work suggests that mental simulation, a fundamental part of cognition (e.g., Schwartz & Black, 1999; Freyd, 1983; Barsalou, 1999) generalizes to fictive motion. In several reading studies, Matlock (in press) investigated whether thinking about motion would affect fictive motion processing. In one study, participants read vignettes about fast or slow travel through a large-scale spatial region (e.g., driving in a desert), and then a fictive motion critical sentence, such as *The road goes through the desert*. Participants were quicker to read the critical sentences after reading about fast motion than they were after reading about slow motion. The same effects were also observed with easy versus difficult terrains, and with short versus long distances. Critically, the effect was *not* obtained with non-fictive motion test sentences at the end of the same stories, such as *The road is in the desert*. In sum, the results show that thinking about motion influences the processing of fictive motion sentences, but not the processing of comparable non-fictive motion sentences. They provide evidence that simulating motion is part of fictive motion understanding.

Matlock, Ramscar, and Boroditsky (2003a, 2003b) investigated whether engaging in thought about fictive motion would influence metaphoric construal of time in the way that engaging in thought about real motion has been shown to do (see Boroditsky, 2000; Boroditsky & Ramscar, 2002). In one experiment, participants were primed with fictive motion sentences or non-fictive motion sentences (e.g., *The tattoo runs along his spine* versus *The tattoo is next to his spine*) before reading this ambiguous question: "Next Wednesday's meeting has been moved forward two days. When is the meeting now that it has been rescheduled?"[3] When primed with fictive motion (congruent with an *ego-moving* construal), people were more likely to say, "Friday", suggesting they viewed themselves "moving" forward in time. When primed with non-fictive motion (congruent with a *time-moving* construal), people were more likely to say, "Monday", suggesting they viewed time as "moving" toward them. Another experiment issued this same question with one of two primes: *The road goes all the way to New York* (fictive motion away from conceptualizer), *The road comes all the way to New York* (fictive motion toward conceptualizer). The results indicated that people were more likely to respond "Friday" after the *away* prime and more likely to respond "Monday" with the *toward* prime. They suggest that people take a perspective and simulate motion when thinking about fictive motion, and that that in turn affects the way they perform abstract reasoning, such as reasoning about temporal movement.

Hence, people simulate motion when processing figurative sentences such as *The road runs along the coast*, and this naturally affects conceptual representation. Given this, we would like to know whether fictive motion also influences perceptual processing.

What Eye Movements Can Tell Us

Eye movements have been measured during a range of cognitive and perceptual activities (for review, see Richardson & Spivey, 2004). Scene perception has been studied in terms of the "bottom up" statistical properties of the image that attracts eye fixations, and in terms of the "top down" knowledge, beliefs or expertise that might affect how one person inspects a scene differently from another (for review, see Henderson, 2003). In a separate research tradition, eye tracking has been used to investigate reading, which engages both linguistic and perceptual processing (Rayner, 1998; Tinker, 1946).

Until recently the intersection between language and visual perception—looking at a scene and listening to a voice—had not been studied. The advent of head-mounted and remote eye tracking devices has allowed researchers to place participants in relatively rich, natural visual contexts and record how the eyes respond to spoken instructions and descriptions. Such experiments have yielded a surprisingly close integration between incremental linguistic processing and visual perception, demonstrating that eye movements to possible referents in the world are used to resolve temporary ambiguities in word recognition and syntactic structure (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), and to predict upcoming agents based on thematic role information (Altmann & Kamide, 2004).

Language has also been shown to modulate eye movements even when there is nothing to look at. In studies by Spivey and colleagues (Spivey & Geng, 2001; Spivey, Tyler, Richardson, & Young, 2000), people stared at a blank screen or closed their eyes and listened to a story that was spatially extended along an axis, for example, a story about a train going past, or a sequence of activities occurring on successive floors of a tall apartment block. While listening, participants' saccades tended to be extended along the horizontal or vertical axes that were consistent with those communicated in the story.

Whether it is in the presence of temporary ambiguity, or in the absence of visual input, linguistic input has been shown to influence eye movements. However, with the current surge of interest in language and vision (e.g., Henderson & Ferreira, 2004) one question has been left behind. What about figurative language? Does it influence eye movements? If so, how? This is an important question, for figurative language is not restricted to poetic or literary works. It is at least as pervasive in every day talk as literal

[3] The question is ambiguous because people are just as apt to answer Friday as they are Monday when the question is posed without any prime. (See Boroditsky, 2000; Boroditsky & Ramscar, 2002 for discussion.)

language, if not more pervasive (see Gibbs, 1994; Katz, Cacciari, Gibbs, & Turner, 1998; Lakoff, 1987).

Experiment

In the current study, participants viewed static depictions of scenes while they heard fictive motion and non-fictive motion sentences. Of interest was whether there would be differences between the eye movements that accompanied fictive motion sentences, such as *The palm trees run along the highway,* and those that accompanied non-fictive motion sentences, such as *The palm trees are next to the highway*. On the surface, the sentences convey similar information: Both include a linearly extended subject noun phrase reference (e.g., *palm trees*) and both describe a static spatial scene. However, the former has been argued to involve mentally simulated motion or scanning along the figure, but the latter has not. Would participants spend more time inspecting figures with fictive motion sentences than figures with non-fictive motion sentences? Longer gaze durations on regions of interest with fictive motion sentences would suggest mentally simulated motion or scanning.

Method

Participants

A total of 24 Stanford University psychology students participated for course credit. All had normal or corrected-to-normal vision.

Design

Gaze durations were recorded along the axis referred to by the subject of the sentence (the compatible region) and along the axis not referred to by the subject of the sentence (the incompatible region). Half the sentences included fictive motion language and half did not. Therefore, the experiment was a 2 x 2 design, with compatibility as one factor and sentence type as the other.

Stimuli

Sixteen pictures served as the primary visual stimuli. Each depicted a simple spatial scene and featured both a horizontally extended figure and a vertically extended figure, for example, a river extending from top to bottom, and a fence extending left to right. A further 16 pictures were used as filler items. All pictures were matched on level of color luminance.

Sixteen blocks of recorded English sentences served as primary stimuli. Each block contained two sentence pairs. Each pair included a FM-sentence (fictive motion sentence) and a comparable NFM-sentence (non-fictive motion sentence), for example, *The cord runs along the wall*, and *The cord is on the wall*. One sentence pair referred to the vertical object in a picture and the other referred to the horizontal object in that same picture. Figure 1 displays an example picture and its block of sentences. Sixteen sentences that described the filler pictures were also recorded.

We conducted three norming studies on our sentences and pictures. In the first, 57 Stanford undergraduates judged all FM- and NFM-sentences on a scale of 1 to 7, in which 1 indicated "makes no sense at all" and 7 indicated "makes good sense". The mean for all FM-sentences was 5.85 and the mean for all NFM-sentences was 6.02. A t-test showed no reliable difference between the two, $t(31) = 1.16$, $p > .1$. In the second norming study, 28 undergraduates rated pairs of FM- and NFM-sentences on how similar they were in meaning. They used a scale of 1 to 7, in which 1 indicated "not at all similar" and 7 indicated "very similar". The mean for all sentence pairs across all subjects was 6.04, with the highest average at 7 and the lowest average at 5.25. In a third norming study, 12 undergraduates judged our pictures and sentences on how well they went together. Overall, the sentence-picture combinations were judged as well-matched. The means were 6.63 FM-horizontal, 6.58 FM-horizontal, 6.53 NFM-horizontal, and 6.34 NFM-vertical. A one-way ANOVA yielded no difference, $F(3, 63) = .04$, $p > .1$, showing that they were equally good descriptions.

Together, the norming studies indicate that (a) all FM- and NFM-sentences were equally sensible in meaning, (b) all FM- and NFM-sentences described comparable information, and (c) all FM- and NFM-sentences were equally good descriptions of the pictures.

Horizontal landmark

FM *The books run along the wall*
NFM *The books are on the wall*

Vertical landmark

FM *The cord runs along the wall*
NFM *The cord is on the wall*

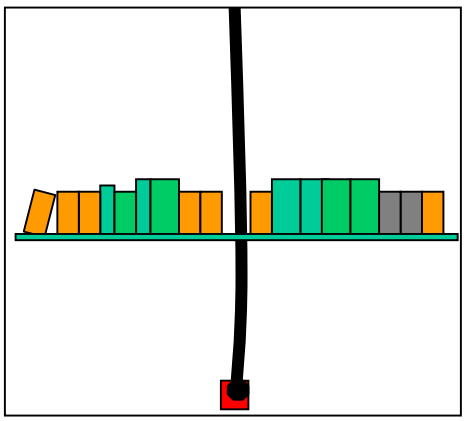


Figure 1. Example of picture with vertical and horizontal fictive motion and non-fictive motion sentences.

Apparatus

An ASL 504 remote eye tracking camera was positioned at the base of a 17" LCD stimulus display. Participants were unrestrained, and sat approximately 30" from the screen. The camera detected pupil and corneal reflection position from the right eye, and the eye tracking PC calculated point-of-gaze in terms of coordinates on the stimulus display. This information was passed every 33ms to a PowerMac G4 which controlled the stimulus presentation and collected gaze duration data. Prior to the experimental session, participants went through a 9 point calibration routine, which typically took between 2 and 5 minutes.

Procedure

Once a successful eye track was established, participants were told to "Look at the pictures and listen to the sentences." On each trial, a picture appeared 1000ms before the sentence began, and then remained in view for 2000ms after the sentence finished. There was a 2000ms inter-stimulus interval, during which participants saw a gray screen, roughly isoluminant with the pictures.

Following 4 practice trials, each participant was presented with a random sequence of 16 experimental and 16 filler trials. Each experimental picture was accompanied by one of four sentences that described the picture (e.g., vertical FM-sentence). Sentence presentation varied such that each participant heard 4 vertical FM-sentences, 4 horizontal FM-sentences, 4 vertical NFM-sentences, and 4 horizontal NFM-sentences.

Coding

The screen was partitioned into 17 non-overlapping regions of interest, corresponding to a central square, six squares spanning the horizontal axis, six squares spanning the vertical axis, and four squares in each corner (see Figure 2). During the period that the experimental picture was onscreen, total gaze durations in each region were recorded.

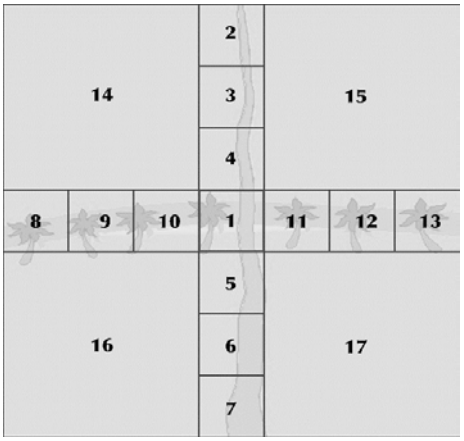


Figure 2. Grid defining relevant regions superimposed on grayed example picture.

Results and Discussion

For a quantitative analysis, we compared total gaze durations to the vertical region of the picture (regions 2 to 7) and the horizontal region of the picture (regions 8 to 13). Each accompanying sentence described either the horizontal or vertical element in the picture. Thus our data could be expressed in terms of gaze durations to the compatible and the incompatible regions. [4]

We conducted a 2 (compatibility) x 2 (sentence) ANOVA. There was a main effect of compatibility, indicating that our participants spent more time inspecting the compatible portion of the grid containing the figure described in the sentences, $F(1,23) = 5.51$, $p < .03$. There was also a main effect of sentence, revealing a reliable difference between inspection time for FM-sentences (M= 995) and inspection time for NFM-sentences (M=827), $F(1,23) = 10.18$, $p < .001$. Most importantly, there was a reliable interaction between these factors, $F(1,23) = 6.00$, $p < .03$, shown in Figure 3. Tukey's HSD revealed that the only cell that differed from all others was gaze duration to the compatible region with fictive motion sentences ($p < .01$).

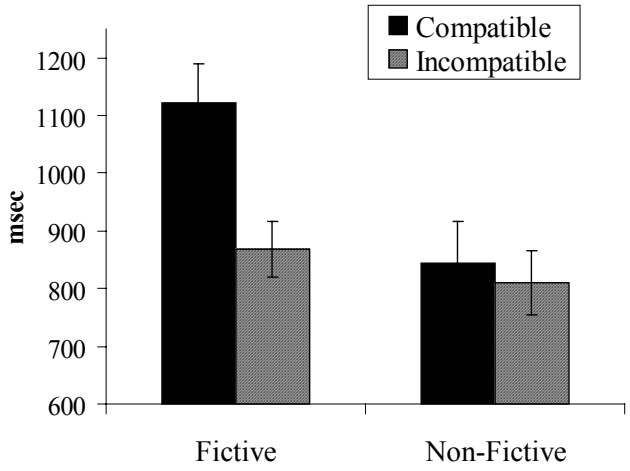


Figure 3. Gaze durations to compatible and incompatible picture regions only differed when the sentence employed fictive motion.

As predicted, people spent more time gazing at the region of a picture associated with the figure in fictive motion input than with the figure in non-fictive motion input, especially when the figure in the picture was compatible with the

[4] Example recordings of eye tracks can be seen at http://psychology.stanford.edu/~richardson/ficmot.

figure in the sentence. Taken together, our results show that fictive motion sentences had a consistent and dramatic effect on eye movements, most notably on the compatible region of interest.

General Discussion

Participants in our preliminary study spent more time inspecting the compatible region of interest in spatial scenes when they heard fictive motion sentences than when they heard non-fictive sentences. The results demonstrate that *figurative* language influences eye movements in consistent and predictable ways. The results are in line with other work on fictive motion (Matlock, in press; Matlock, Ramscar, & Boroditsky, 2003a, 2003b), and they suggest a dynamic mental representation that mirrors perception or enactment of motion (e.g., Barsalou, 1999, Glenberg, 1999).

One explanation for the results obtained here is that when our participants were presented with fictive motion input, they mentally simulated motion along the figure, and then their eye movements mirrored that internal simulation. For example, on hearing the sentence *The road runs through the desert*, participants conceptually "moved" along a road and then their eye movements enacted a congruent simulation. Another not incompatible explanation assigns a more active role to eye movements. It might be that participants' eye movements were central to simulation and building an appropriate representation of the figure. For example, on hearing the sentence *The road runs through the desert* and seeing a depiction of that scene, participants' eye movements allowed them to incrementally construct an appropriate model of the road. If this is the case, then perhaps eye movements allowed participants to simulate and compute some information about the scene externally (for related views, see Spivey, Richardson, & Fitneva, in press; Spivey, et al 2000).

Are there other explanations for longer gaze durations with fictive motion sentences? For instance, could it be that people activated the literal meaning of the motion verb in fictive motion sentences, and that that literal interpretation led to longer inspection times? Based on the results reported here, we cannot rule out this possibility entirely. But we would argue that our compatibility results suggest that this is not likely. Namely, if the verb alone – independent of the figurative meaning of the fictive motion sentence – brought on longer gaze durations, we would not have seen selective differences in the axis of orientation (vertical versus horizontal). After all, the motion verb alone provided no information about direction.

Our data show that figurative language, like literal language, influences eye movements. We argue that this is because fictive language evokes a dynamic mental simulation, and that this simulation determines how the visual system interprets and inspects the world. Further research will reveal how these simulations occur and the extent to which they mirror perception or enactment of physical motion in the world.

Acknowledgements

We thank Elsie Wang and Teresa-Bulygo for helping with our norming studies. We also thank Michael Spivey, Herbert Clark, Paul Maglio, Daniel Casasanto, and Gordon Bower for insightful comments.

References

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: mediating the mapping between language and the visual world. In J. M. Henderson & F. Ferreira (Eds.), *Interfacing Language, Vision, and Action*. San Diego: Academic Press.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences, 22 , 577-660.*

Boroditsky, L. (2000). Metaphoric Structuring: Understanding time through spatial metaphors. *Cognition, 75*, 1-28.

Boroditsky, L. & Ramscar, M. (2002). The Roles of Body and Mind in Abstract Thought. *Psychological Science, 13(2)*, 185-188.

Freyd, J.J. (1983). The mental representation of movement when static stimuli are viewed. *Perception & Psychophysics, 33*, 575-581.

Gibbs, R. W., Jr. (1994). *The Poetics of Mind : Figurative Thought, Language, and Understanding*. Cambridge, UK: Cambridge University Press.

Glenberg, A. M. (1999). Why mental models must be embodied. In G. Rickheit, & C. Habel (Eds.), *Mental models in discourse processing and reasoning*. New York, NY: North-Holland.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences, 7*, 498-504.

Henderson, J. M., & Ferreira, F. (Eds.). (2004). *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Jackendoff, R (2002). *Foundations of language.* New York: Oxford University Press.

Katz, A.N., Cacciari, C., Gibbs, R.W., & Turner, M. (Eds.) (1998). *Figurative language and thought*. New York, NY: Oxford University Press.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, IL: University of Chicago Press.

Langacker, R.W. (1986). Abstract motion. *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society,* 455–471.

Langacker, R. W. (2000). Virtual reality. *Studies in the Linguistic Sciences, 29*, 77-103

Matlock, T. (in press). Fictive motion as cognitive simulation. *Memory & Cognition*.

Matlock, T. (2001). Fictive motion is real motion. Presentation at Seventh International Cognitive Linguistics Conference (ICLC), Santa Barbara.

Matlock, T., Ramscar, M., & Boroditsky, L. (2003a). The experiential basis of meaning. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Matlock, T., Ramscar, M., & Boroditsky, L. (2003b). The experiential basis of motion language. *Proceedings of Language, Culture, and Cognition: An International Conference on Cognitive Linguistics*. Braga, Portugal.

Matsumoto, Y. (1996). Subjective motion and English and Japanese verbs. *Cognitive Linguistics, 7*, 183-226.

Miller, G. & Johnson-Laird, P. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124(3),* 372-422.

Richardson, D. C., & Spivey, M. J. (2004). Eye Movements. In Wenk. G. & Bowlin, G. (Eds.) *Encyclopedia of Biomaterials and Biomedical Engineering*. Marcel Dekker, Inc.

Rumelhart, D.E. (1979). Some problems with the notion of literal meaning. In A. Ortony (ed). *Metaphor and Thought. Cambridge University Press*.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 116–136.

Spivey, M.J., & Geng, J.J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research/Psychologische Forschung, 65*, 235-241.

Spivey, M.J., Richardson, D.C., & Fitneva, S.A. (in press). Thinking outside the brain: Spatial indices to visual and linguistics information. In J. Henderson & F. Ferreira (Eds.), *Interfacing Language, Vision, and Action*. San Diego, CA: Academic Press.

Spivey, M.J., Tyler, M.J., Richardson, D.C., & Young, E.E. (2000). Eye movements during comprehension of spoken scene descriptions. *The Proceedings of the Twenty-second Annual Cognitive Science Society Meeting*, 487-492.

Talmy, L. (1975). Semantics and syntax of motion. In J. P. Kimball (Ed.), *Syntax and Semantics*, Volume 4 (pp.181–238). New York, NY: Academic Press.

Talmy, L. (1983). How language structures space. In H. Pick, & L.P. Acredolo (Eds.), *Spatial orientation: Theory, research, and application* (pp. 225-282). New York: Plenum Press.

Talmy, L. (1996). Fictive motion in language and "ception". In P. Bloom, M.A. Peterson, L. Nadel, & M.F. Garrett (Eds.), *Language and space* (pp. 211-276). Cambridge, MA: MIT Press.

Talmy, L. (2000). *Toward a Cognitive Semantics, Volume I: Conceptual Structuring Systems*. Cambridge: MIT Press.

Tanenhaus, M. K., Spivey Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632-1634.

Tinker, M. A. (1946). The study of eye movements in reading. *Psychological Bulletin*.

Biased stochastic learning in computational model of category learning

Toshihiko Matsuka (matsuka@psychology.rutgers.edu)
RUMBA, Rutgers University – Newark
101 Warren St., Smith Hall 327, Newark, NJ 07102 USA

Abstract

Matsuka and Corter (2003b) presented evidence that people tend to utilize only the minimally necessary information for classification tasks. This approach for categorization was efficient and valid for the stimulus set used in the experiment, but might be considered a statistically or mathematically non-normative approach. In the present paper, I hypothesized human category learning processes are biased toward simpler representation and/or conception rather than complex but normative ones. In particular, a few variants of "biased" learning algorithms are introduced and applied to Matsuka and Corter's stochastic learning algorithm (2003a, 2004). The result of a simulation study showed that the biased learning models account for empirical results successfully.

Introduction

In their recent work, Matsuka and Corter (2003b & 2004) investigated the possibility of using stochastic learning rather than gradient-based methods in neural network models of human category learning. They introduced stochastic learning models to more accurately account for human category learning. The gradient based learning algorithm used in many neural network models may be considered to have a normative justification (i.e., it models how people "should" learn or process information), but may not be descriptively valid at the individual level. Models utilizing a gradient method for learning seem to require a high degree of mental effort and assume that optimal adjustments are made to the vector of parameters on each trial. In contrast, Matsuka & Corter's stochastic learning model (2003a, 2004) does not assume that learning is associated with monotonic increases in accuracy (and attention) or continuous search for better categorization processes by humans. Rather, it models random fluctuations or "errors" in people's memory and learning processes, and how people utilize and "misutilize" such errors.

In their simulation studies (Matsuka & Corter 2004a), the effectiveness of stochastic learning methods applied to an ALCOVE-like model (Kruschke, 1992) was evaluated in several settings. The modified models were shown to be satisfactory in replicating two phenomena observed in empirical studies on categorization; namely, rapid change in attention processes (Macho 1997; Rehder and Hoffman 2003), and individual differences in distribution of attention (Matsuka & Corter 2003b).

Although the stochastic learning model reproduced more realistic individual differences than models with a gradient type learning algorithm, it did not replicate one tendency observed in the empirical study of Matsuka and Corter (2003b). They found that for four dimensional stimulus sets with two diagnostic but perfectly correlated dimensions, the proportion of human participants who paid attention primarily to only one of the two correlated dimensions was higher than that of those who paid attention to both of the two correlated dimensions approximately equally (see Figure 2, top row, third column). In other words, many participants utilized only the minimal necessary information for this task. In contrast, the stochastic learning model inadequately predicted that a higher proportion of participants would pay attention to the two correlated dimensions approximately equally.

The strategy of using minimal information may be a very natural and efficient usage of limited mental resources for humans. This would be particularly true for real world categorization tasks, where the number of feature dimensions could easily exceed a manageable number, in which many are not necessary or crucial (e.g., irrelevant and or highly correlated) for successful categorization. There are several ways that could lead people to use a lesser amount of information, resulting in simple conception of categories. One possible explanation is that there may be an implicit or explicit penalizing mechanism in human cognition that encourages less complete but simpler concepts than more complete but more complex concepts. Another possible explanation is that there may be a mechanism in human cognition that leads to a more thorough search for simple concepts.

In the present research, based on these remarks, I hypothesize and model human category learning as being biased toward simpler and heuristic concepts [1] (or representation) than complex and complete ones.

Biased Stochastic Learning

The proposed algorithm is based on a simulated annealing algorithm (Kirkpatrick, Gelatt, & Vecchi, 1983; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1958) and somewhat resembles Boltzmann Machine (Hinton & Sejnowski, 1986). In the present algorithm, initial association weights are randomly selected from a uniform distribution centered at 0, and initial dimension attention weights are equally distributed across all dimensions. This equal attention allocation in the early stages of learning is motivated by the results of empirical studies (Matsuka, 2002; Rehder & Hoffman, 2003) that showed many participants initially tended to evenly allocate attention to the feature dimensions. In the present algorithm, at the beginning of each training epoch, a hypothetical "move" in

[1] In the present paper, the concepts of categories correspond to the configurations of the association weights and dimensional attention attractiveness.

the parameter space is computed by adjusting each parameter by an independently sampled term. These adjustment terms are drawn from a prespecified distribution. The move (i.e., the set of new parameter values) is then accepted or rejected, based on the computed relative fit or utility (defined below) of the new values. Specifically, if the new parameter values result in a better fit/utility, they are accepted. If they result in a poorer fit/utility, they are accepted with some probability *P*. This probability is a function of a parameter called the "temperature", which decreases across blocks according to the annealing schedule.

Because of the human's biased cognitive processes, possibly as a consequence of our implicit or explicit biased processes and/or preference toward simpler but less complete concept (these processes are discussed in detail in the model section), the learned concepts of categories, thus the configuration of the association weights and attention strengths, are inclined toward simpler ones. Note that in the present algorithm the notion of simplicity (or complexity) is directly related to the number of effective (non-zero, or non-subzero) association weights and attention strengths.

The proposed models would not require computation intensive (back) propagations of classification errors. Rather, in the present biased stochastic learning model framework, a very simple operation (e.g., comparison of two values) along with the operation of stochastic processes are assumed to be the key mechanisms in category learning. These learning algorithms can be applied to virtually any feed-forward NN model of human category learning

General Algorithm for Stochastic Learning

A general framework for the stochastic learning algorithm is discussed in this section. Here, the stochastic learning algorithm is embedded into ALCOVE, which is one of the most studied and applied computational models of category learning incorporating a selective attention mechanism (Kruschke, 1992). Again, it should be noted that this learning algorithm is very general and can be applied to virtually any NN model of category learning.

STEP 0: Initialization:

Problem specific parameters: (T^0,υ)

T^0 : initial temperature.

υ : temperature decreasing rate

Association weights w_{kj},, Attention strengths α_i,, Exemplar ψ_{ji}

STEP 1: Calculate ALCOVE output activations:

$$O_k = \sum_J w_{kj} \exp\left[-c\left(\sum_I \alpha_i \mid \psi_{ji} - x_i \mid\right)\right] \quad \text{(SL-1)}$$

STEP 2: Calculate fit index for the current parameter set:

$$F(w^t, \alpha^t) = \sum_{n=1}^{N}\sum_{k=1}^{K}(d_k - O_k^t)^2 \quad \text{(SL-2)}$$

where K = # categories, N = # input in one block, d_k is a desired output for category node k. Here, the superscript t indicates time.

STEP 3: Accept or reject of parameter set, $\boldsymbol{\alpha}$ & $\boldsymbol{w}$:

Accept all weight and attention parameters at the probability of:

$$P(w^t, \alpha^t \mid w^s, \alpha^s, T^t) = \left\{1 + \exp\left(\frac{F(w^t, \alpha^t) - F(w^s, \alpha^s)}{T^t}\right)\right\}^{-1} \quad \text{(SL-3)}$$

if $F(w^t, \alpha^t) > F(w^s, \alpha^s)$, or 1 otherwise, where $F(w^s, \alpha^s)$ is the fit index for the previously accepted parameter set, and T^t is temperature at time t.

STEP 4: Reduce temperature:

$$T^t = T^o \delta(\upsilon, t) \quad \text{(SL-4)}$$

where δ is the temperature decreasing function that take temperature decreasing rate, υ, and time t as inputs.

STEP 5: Generate new w_{kj} and α_i.

$$w_{kj}^t = w_{kj}^s + r^w,\ r^w \sim \Phi^w(\cdot) \quad \text{(SL-5)}$$

$$\alpha_i^t = \alpha_i^s + r^\alpha,\ r^\alpha \sim \Phi^\alpha(\cdot) \quad \text{(SL-6)}$$

where r^w and r^α are random numbers generated from prespecified distributions Φ^w and Φ^α.

REPEAT STEPS 1~5 until stopping criterion is met.

Biased Stochastic Learning Models

There are several approaches to model biased learning processes using stochastic learning model. Here, two simple approaches are introduced. The first biased learning model is based on the parameter regularization in which complex parameter configurations are penalized. The second model based on asymmetric random distributions, searches simpler parameter configurations more thoroughly.

Model 1: Bias via penalizing fitness function

In the present algorithm the utility index rather than the fit index is used for the decision on acceptance and rejection of the current parameter set. The utility of a particular parameter configuration is defined as a weighted sum of the accuracy in classification and the mental effort required by the parameter configuration. Thus, the utility index consists of two independent indices, namely "classification accuracy", *L* and "mental effort", *Q*, both dependent on learnable parameters $\boldsymbol{w}$ and $\boldsymbol{\alpha}$ at time *t*.

$$U(w^t, \alpha^t) = L(w^t, \alpha^t) + Q(w^t, \alpha^t) \quad \text{(M1-1)}$$

The *L* function can be the same function for the fitness index (i.e., Eq. SL-2). Here, the *Q* function may be considered as a penalty function, penalizing "complex" parameter configurations that are believed to require more mental effort. The general form of *Q* function is given as follows:

$$Q(w^t, \alpha^t) = \gamma_w \phi^w(w_m^t) + \gamma_\alpha \phi^\alpha(a_m^t) \quad \text{(M1-2)}$$

where ϕ^w and ϕ^α are functions calculating mental effort required for specific parameter configurations at time *t* (i.e., $\boldsymbol{w}^t$ and $\boldsymbol{\alpha}^t$), and γ_w and γ_α are coefficients weighting these mental efforts. Note that γ_w and γ_α also control relative importance of *L* and *Q* functions (i.e., accuracy vs. simplicity). That is the hypothetical coefficient, γ_Q, weighting importance of *Q* function relative to *L* function is

included in γ_w and γ_α. I.e., $\gamma_w = \gamma_Q \gamma_w^*$ and $\gamma_\alpha = \gamma_Q \gamma_\alpha^*$. Thus Equations M1-1 and M1-2 may be rewritten as:

$$U(w^t, \alpha^t) = L(w^t, \alpha^t) + \gamma_Q Q^*(w^t, \alpha^t) \quad \text{(M1-1R)}$$

$$\gamma_Q Q^*(w^t, \alpha^t) = \gamma_Q \gamma_w^* \phi^w(w_m^t) + \gamma_Q \gamma_\alpha^* \phi^\alpha(a_m^t) \quad \text{(M1-2R)}$$

There are several functions applicable for ϕ:

$$\phi^w = \sum_j \sum_k w_{kj}^2 \quad \text{(M1-3a)}; \qquad \phi^\alpha = \sum_i \alpha_i^2 \quad \text{(M1-3b)}$$

$$\phi^w = \sum_j \sum_k I(|w_{kj}| > \zeta^w) \quad \text{(M1-4a)}$$

$$\phi^\alpha = \sum_i I(\alpha_i > \zeta^\alpha) \quad \text{(M1-4b)}$$

where ζ^w and ζ^α are threshold values, and $I(expression)$ is the indicator function that returns 1 if the *expression* is satisfied. Equations M1-3a and M1-3b, often referred to as ridge penalty function or weight decay, encourage parameter settings that have small parameter values, whereas Equations M1-4a and M1-4b encourage parameter settings that have large number of parameters with less than the threshold values ζs. More general ϕ function is given as follows:

$$\phi^w = \sum_j \sum_k \frac{\left(w_{kj}/q\right)^2}{1+\left(w_{kj}/q\right)^2}, \qquad \phi^\alpha = \sum_i \frac{\left(\alpha_i/q\right)^2}{1+\left(\alpha_i/q\right)^2} \quad \text{(M1-5)}$$

where q, which can be either time dependent or independent, controls types of penalization or encouragement. That is, Equations M1-5 approach Equations M1-3s as $q \to \infty$, and approach Equations M1-4s as $q \to 0$ (Cherkassky & Mulier, 1997).

In many simulation studies, relative, but not absolute predicted attention allocation strengths are analyzed and compared (e.g. Matsuka, 2002). In such cases, the relative attention strengths $a_i = \alpha_i / \Sigma(\alpha_m)$ should be used as inputs for the penalty function. In addition, the penalization functions do not have to be in the same form for association weights and attention strengths. For example, in order to pay attention to a smaller number of feature dimensions it seems more sensible to use M1-4b or M1-5 with small q values for the attention parameters, because the relative but not absolute attention strength values are usually considered. In contrast, either choice seems appropriate for the association weight parameters where raw values are usually used.

Model 2: Bias via asymmetric distribution.

In the present model, random numbers are drawn from an asymmetric distribution with its mode equal to zero. Thus, as in the previous model, the probability of drawing a random number r from the vicinity of current values (i.e., vicinity of zero) is still the highest

$$P(0-\varepsilon < r < 0+\varepsilon) > P(M-\varepsilon < r < M+\varepsilon) \quad \text{(M2-1)}$$

for all $M \neq 0$.

However, unlike the previous model, for a particular parameter value, the probability of drawing a random number which will lead its updated value toward zero is higher than that of a random number that leads to the opposite direction. In other words, when the association weight value, w_{kj} is negative, then the probability of drawing a positive number is greater than a negative number; when the weight is positive, then the opposite is true, or

$$P(r^w > 0 \mid w_{kj} > 0) < P(r^w < 0 \mid w_{kj} > 0)$$
$$P(r^w > 0 \mid w_{kj} < 0) > P(r^w < 0 \mid w_{kj} < 0) \quad \text{(M2-2)}$$

For the attention strength parameter α_i the probability of drawing a negative random move is larger than for a positive move, assuming that α_i is constrained to be positive, thus,

$$P(r^\alpha > 0) < P(r^\alpha < 0). \quad \text{(M2-3)}$$

Parameter updates are accomplished by the following functions:

$$w_{kj}^{t+1} = w_{kj}^t + r_{kj}^w \quad \text{(M2-4)}$$

$$\alpha_i^{t+1} = \alpha_i^t + r_i^\alpha \quad \text{(M2-5)}$$

where $r_{kj}^w \sim \text{sgn}(w_{kj}^t) \cdot \Phi^w(\cdot)$ and $r_i^\alpha \sim \Phi^\alpha(\cdot)$

The random movement r_m is drawn from the negatively skewed distributions for α_i and w_{kj} if w_{kj} is positive, and from the positively skewed distributions for negative w_{kj}. Thus, the expected value of the distance of the random movement leading the learnable parameters to zero is greater than that of the opposite direction. This makes the model to decrease values of "irrelevant" parameters quickly.

There are several asymmetric distributions, and the χ^2 (Eq. M2-6, Figure 1, left panel) and Rayleigh (M2-7 & Figure 1, right panel) distributions are examples of asymmetrical distributions.

$$f(x \mid v) = \frac{1}{\Gamma(v/2)} \left(\frac{1}{2}\right)^{v/2} x^{\frac{v}{2}-1} \exp\left(\frac{-x}{2}\right) \quad \text{(M2-6)}$$

where $\Gamma(\cdot)$ is a gamma function, v is the degree of freedom.

$$f(x \mid b) = \frac{x}{b^2} \exp\left(\frac{-x^2}{2b^2}\right) \quad \text{(M2-7)}$$

where b is the Rayleigh distribution parameter.

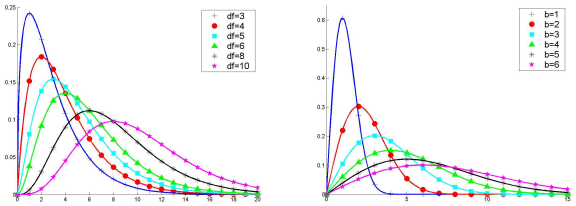


Figure 1. Example asymmetric distributions. Left panel: χ^2 distributions with several different distribution parameters. Right panel: Rayleigh distributions with several different distribution parameters.

Since the modes of these asymmetric non-negative distributions are not zero, and the distribution parameters affect both central tendencies and spreads of the distributions, the random numbers should be transformed as:

$$r = -s^t(f(x) - MODE(f(x))) \quad \text{(M2-8)}$$

where s^t is a time-dependent scalar controlling the width of the search areas. This ensures that the mode of the transformed random variable is zero and thus satisfies M2-1. Note that the distribution parameter v or b may be selected a priori and held constant throughout the training, or they can be time dependent so that the model starts with a highly skewed distribution and terminates with a near normal distribution, or vice versa.

While the present biased learning model (bias via thorough searches around zero) may be interpreted as active bias, actively trying to reduce the effective number of parameters or simplifying concepts, the bias via regularization (Model 1) may be interpreted as passive bias, involuntary resulting in simpler concept because of the limitation of mental capacity.

Simulations

Here, I examined how the two new biased stochastic learning models account for individual differences in attention learning. To do this, I simulated the results of an empirical study on classification learning, Study 2 of Matsuka (2002). In this study, there were two perfectly redundant feature dimensions, Dimension 1 & Dimension 2 (see Table 1), and those two dimensions are also perfectly correlated with category membership. Thus, information from only one of the two correlated dimensions was necessary and sufficient for perfect categorization performance. Besides classification accuracy, data on the amount of attention allocated to each feature dimension were collected in the empirical study. The measures of attention used were based on feature viewing time, as measured in a MouseLab-type interface (Bettman, Johnson, Luce, & Payne, 1993).

The empirical results that I am trying to simulate indicated that 13 out of 14 subjects were able to categorize the stimuli almost perfectly (Figure 2, top left panel). The aggregated results suggest that on average subjects paid attention to both of the correlated dimensions approximately equally (Figure 3, top middle panel). However, more interestingly when the attention data were analyzed per individual, it was found that many subjects tended to pay attention primarily to only one of the two correlated dimensions, particularly in the late learning blocks as shown in Figure 2, top row third column (Matsuka & Corter, 2003). This suggests that subjects used only the minimal necessary information for this task.

Simulation method: There were three ALCOVE-type models in the present simulation study, namely ALCOVE with stochastic learning (ASL; Matsuka & Corter, 2003a, 2004); ALCOVE with a regularized stochastic learning (ARSL); and ALCOVE with the Rayleigh distribution-based stochastic learning (ARAY). The standard ALCOVE will not be evaluated in the present simulation study, because its standard gradient learning method was shown to be unsuccessful in replicating individual difference when attention allocation is initialized equally (Matsuka & Corter, 2003a, 2004).

All three models were run in a simulated training procedure to learn the correct classification responses for the stimuli of the experiment. ARAY was run for 300 blocks of training, where each block consisted of a complete set of the training instances, while ASL and ARSL were run for 500 training blocks. For each model, the final results are based on 50 replications.

The model configurations (e.g., type of distribution, temperature decreasing rate & function, search ranges) for ASL and ARSL were the same except for the additional parameter-penalization functions incorporated in RSL to model biased processes in category learning. The random numbers for these two models were drawn from the Cauchy distribution, and its random number generation algorithm was based on Ingber (1989). For ARSL, the ridge penalty (Equation M1-3a) was imposed on the association weights, and a subset selection method (M1-4b with $\zeta = 0.1$) was used for the *relative* attention strengths.

For ARAY, a (pseudo) random number generator function from MATLAB Statistical Toolbox (MathWorks, 2001) was used to generate random numbers, and its transforming scalar s (see Eq. M2-8) was exponentially decreased during the learning. For all models, an exponential function was used as the temperature decreasing function. Models' user-definable parameters (e.g., initial temperature, temperature decreasing rate, ζ, and etc…) were selected arbitrarily.

Table 1: Stimulus structure used in Study 2 of Matsuka

Category	Dim1	Dim2	Dim3	Dim4
A	1*	1*	3	4
A	1*	1*	4	1
A	1*	1*	1	2
B	2*	2*	2	1
B	2*	2*	3	2
B	2*	2*	4	3
C	3*	3*	1	3
C	3*	3*	2	4
C	3*	3*	3	1
D	4*	4*	4	2
D	4*	4*	2	3
D	4*	4*	1	4

*Diagnostic feature

Results: All three models correctly replicated aggregated or averaged relative attention allocations to the four feature dimensions (Figure 2, second column). However, there are some minor differences in their predictions; ARAY paid less attention to non-diagnostic dimensions than ASL, which in turn paid less attention to those dimensions as compared with ARSL. Qualitatively, ARSL appears to be the most successful in replicating not paying attention to both Dimension 1 and 2 equally, while ASL appears to be least successful in this regard. ARAY was similarly unsuccessful, overestimating the proportion of people who would attend to both of the correlated dimensions equally. A noticeable difference between ARAY and other two

models is that ARAY virtually ignored non-diagnostic feature dimensions and paid attention exclusively to either or both Dimensions 1 and 2.

Among all three models, the proportion of sub-zero association weights for ARAY was the largest (Figure 2, fourth column), indicating it yielded simpler category conceptions than the other two models. Here, the notion of simplicity (or complexity) is directly related to numbers of effective (i.e., non-zero, or non-subzero) association weights and attention parameters. When compared with the distribution of the association weights of ASL, the proportion of sub-zero weights for ARSL was larger, indicating penalizing processes incorporated in ARSL resulted in simpler configuration. Note that the model configurations and settings for ASL and ARSL were the same expect for the regularization process incorporated in ARSL. Thus, the straightforward comparison of ASL and ARSL seems reasonable. However, because ARAY and ARSL had different parameter settings, interpreting the comparisons of distributions of the weights for ARAY and ARSL or ASL should be done with care.

In sum, the stochastic learning model with the regularizing processes penalizing mentally-expensive complex category conceptions (i.e. ARSL) appears to be the most successful model capturing human category learning trends that appeared biased, heuristic, and/or less optimal.

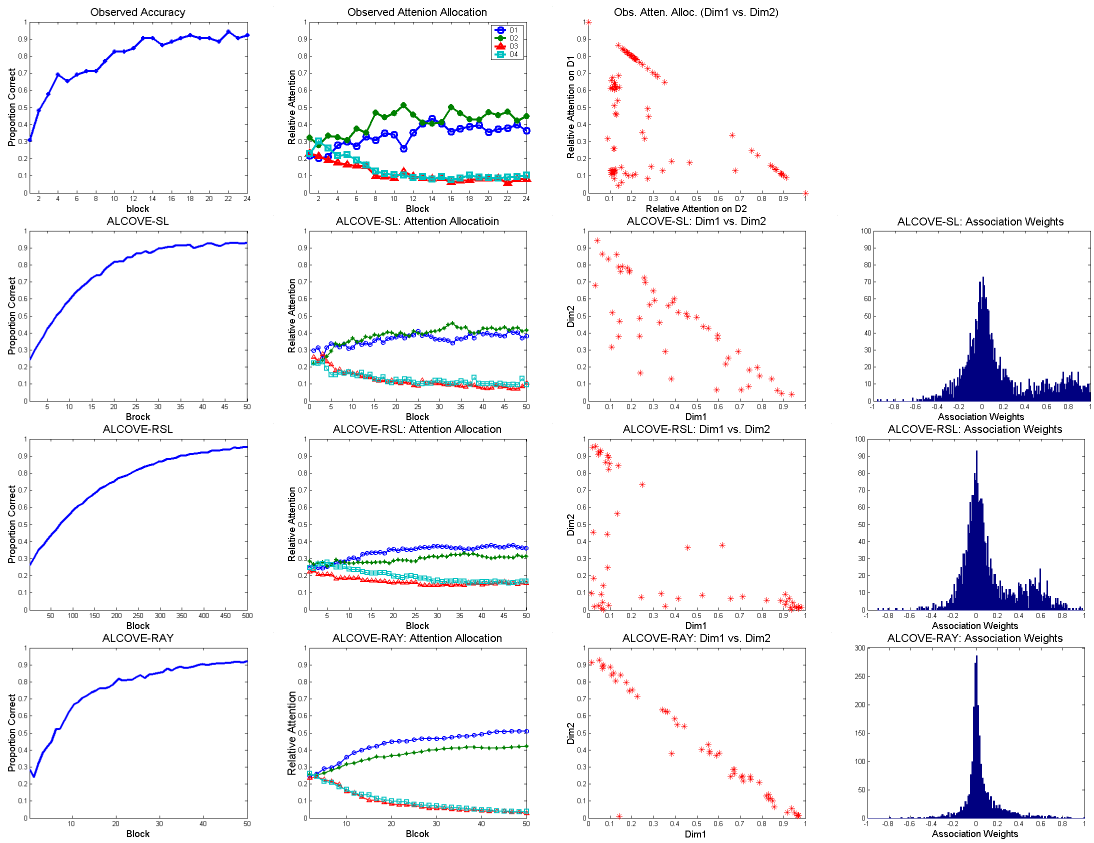


Figure 2. Results of the simulation study. Top row: Observed empirical results of Matsuka & Corter (2003b). The graphs on the first column show observed and predicted classification accuracy, second column shows relative attention allocation for the four feature dimensions; third column compares relative attention allocated to Dimensions 1 and 2 for the last four blocks, where each dot represents an observation. Fourth column shows histograms for the final association weights. Second row shows results of ALCOVE-SL; Third row, ALCOVE-RSL; Fourth Row, ALCOVE-RAY.

Discussion: Although there are 12 unique exemplars in the stimulus set, there are only four exemplars (one from each category) needed for a perfect categorization. Then, one might wonder if people would utilize all the exemplars or not. The distribution of ARAY's association weights may suggest that there are several "dead" or inactive exemplars whose association weights are all zero or near-zero, not being utilized for categorization. This characteristic along with not paying attention to irrelevant feature dimensions may suggest that ARAY replicates learning of an efficient learner, who utilizes a lesser amount of information. In contrast, ARSL predicts that people would utilize more than necessary information. In terms of attention allocation, the empirical results indicate that some people do try utilizing irrelevant information, suggesting that ARSL is more descriptive than other models. This suggests that people may not actively being biased, searching for simpler concepts (i.e., Model 2). Rather it suggests that biases may be caused by the limited mental capacity, involuntarily resulting in simpler concepts.

Discussion

RULEX vs. Stochastic Learning: The stochastic learning's take-all-or–none parameter updating strategy may be considered as a type of hypothesis testing learning model, which makes it similar to the RULEX model (Nosofsky, Palmeri, & McKinley, 1994). However, its random search method, interpreted as unstructured hypothesis generation and search, is very distinct from RULEX whose hypothesis generation algorithm is very strategic and well-structured. Thus, for Matsuka's (2002) stimuli set, RULEX would predict that everyone would allocate his/her attention exclusively to the one of the two diagnostic dimensions. Whereas the stochastic learning would predict some paying attention to either one of the two dimensions, another paying attention to both, and others distributing attention in some other combinations, since, as an exemplar-based model, it can minimize classification error with several different attention allocation patterns (i.e., it can learn to classify stimuli without "optimal" or "rational" attention distribution). In other words, when there are several minima, which is probably true for real world category learning task, stochastic learning can result in several different learning trajectories and parameter (i.e., association weight & attention allocation) configurations, corresponding to possible individual differences. In contrast, RULEX would always predict that people pay attention to the least number of dimensions, which may be a too normative prediction.

Gradient-type vs. Stochastic Learning: For two perfectly redundant feature dimensions, a gradient-type learning algorithm in general would allocate the same amounts of attention to the two dimensions, or its attention learning curves for the two dimensions would be parallel. In contrast, (biased) stochastic learning could result in asymmetric attention allocation to the two dimensions, and its attention learning curves are not necessarily parallel. In these regards, stochastic learning's predictions appear more realistic than those of gradient-type learning. However, this point alone does not necessarily indicate stochastic learning is what people would do. Perhaps, a gradient-type learning with some stochastic elements or errors might, as well, result in more "realistic" predictions.

Conclusion

Biased stochastic learning is a descriptive model of heuristic learning that prefers a simpler conception of categories in which less mental effort seems to be needed. Although the present two stochastic learning algorithms are intended to model such bias, the algorithms appear to be modeling two different types of learners, namely "ordinary people" and "proficients". The simulation study indicates that modeling biased learning via parameter-configuration regularization was the most successful in replicating the empirical results (i.e., ordinary people). In contrast, biased learning via asymmetric distributions appears to be more optimal or rational model, paying attention to only diagnostic feature dimensions and having smaller numbers of effective association weights (proficient-like concepts).

Although the present study supports biased stochastic learning's descriptive validity, more comprehensive simulation studies would be useful in evaluating the present learning models.

References

Bettman, J.R., Johnson, E.J., Luce, M.F., Payne, J.W. (1993). Correlation, conflict, and Choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 931-951.

Cherkassky, V. & Mulier, F. (1997). *Learning from data: Concepts, Theory, and Methods.* New York: Wiley

Hinton, G E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D.E. Rumelhart & J.L. McClelland (Eds.) *Parallel distributed processing: Explorations in microstructure of cognition.* Cambridge, MA: MIT Press.

Ingber, L. (1998). Very fast simulated annealing. *Journal of Mathematical Modelling, 12:* 967-973.

Kruschke, J. E. (1992). ALCOVE: An exemplar-based connectionist model of category learning, *Psychological Review, 99.* 22-44.

Kirkpatrick, S., Gelatt Jr., C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220,* 671-680.

Macho, S. (1997). Effect of relevance shifts in category acquisition: A test of neural networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 30-53.

MathWorks. (2001). *MATLAB* [Computer Software]. Natick, MA: Author.

Matsuka, T. (2002). Attention processes in computational models of category learning. Unpublished doctoral dissertation. Columbia University, NY.

Matsuka, T. (2003). Generalized exploratory model of human category learning. Accepted for publication.

Matsuka, T & Corter, J. E. (2003a). Stochastic learning in neural network models of category learning. Proceedings of the 25th Annual Meeting of the Cognitive Science Society.

Matsuka, T. & Corter, J. E. (2003b). Empirical studies on attention processes in category learning. Poster presented at 44th Annual Meeting of the Psychonomic Society. Vancouver, BC, Canada.

Matsuka, T. & Corter, J.E (2004). Stochastic learning algorithm for modeling human category learning. Accepted for publication.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.

Nosofsky, R.M., Palmeri, T.J., & McKinley, S.C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79 .

Rehder, B. & Hoffman, A. B. (2003). Eyetracking and selective attention in category learning. Proceedings of the 25th Annual Meeting of the Cognitive Science Society, Boston, 2003.

Comparisons of prototype- and exemplar-based neural network models of categorization using the GECLE framework

Toshihiko Matsuka (matsuka@psychology.rutgsers.edu)
RUMBA, Rutgers University – Newark
101 Warren St., Smith Hall 327, Newark, NJ 07102 USA

Abstract

In the present study, GECLE (Matsuka, 2003) was used as a general modeling framework to systematically compare the plausibility of two prominent assumptions about internal representations of neural network (NN) models of human category learning. In particular, exemplar-model friendly Medin and Schaffer's 5/4 stimulus set (1978) was used for comparing prototype- and exemplar-based NN models. The results indicate that some prototype-based models performed as good as or better than an exemplar-based model in replicating the empirical classification profile. In addition, a phenomenon called A2 advantage (i.e., people tend to categorize the less "prototypical" stimulus A2 more accurately than more "prototypical" stimulus A1) reported in empirical studies (e.g., Medin & Schaffer 1978) was also successfully reproduced by these prototype-based NN models.

Introduction

There have been an increasing number of studies debating how stimuli are internally represented in human cognition during the last few decades (e.g., Minda & Smith 2002; Nosofsky & Zaki 2002). Most of these debates have been based on quantitative models of categorization, and only a few have considered representational aspects of adaptive, network, or learning models of categorization. Several studies (Matsuka, 2002; Matsuka, Corter, & Markman, 2003) have compared exemplar-based (EB) and prototype-based (PB) adaptive network models of categorization, but there has been no systematic comparison of specific assumptions in EB and PB modeling. Although these comparative studies provided information on the models' capabilities for reproducing human-like categorization learning, they did not necessarily provide information that can lead to specific understanding of the nature of human category learning. That is because model-to-model comparisons are not informative for testing the plausibility of each specific assumption, rather such model comparisons are essentially omnibus tests collectively comparing all variations in assumptions at once. In other words, it has been difficult to use the results of these previous comparative studies to understand which specific assumptions are supported by the empirical data. Therefore, it seems desirable to make systematic comparisons between competing model assumptions using a general modeling framework that allows us to manipulate and test one or a limited number of model assumptions at a time.

In the present study, a generalized exploratory modeling approach for human category learning is introduced. Then, using this general framework two assumptions about how categories are internally represented, namely prototypes and exemplars, are compared in a systematic fashion.

GECLE

GECLE (for Generalized Exploratory models of Category LEarning) is a general and flexible exploratory approach for modeling human category learning, that is capable of modeling human category learning with many variants using different model assumptions (Matsuka, 2003). This general modeling framework allows model assumptions to be manipulated separately and independently. For example, one can manipulate assumptions about how stimuli are internally represented (e.g. exemplars vs. prototypes), or about how people selectively pay attention to input feature dimensions (e.g., paying attention to dimensions independently or not).

The GECLE model uses the Mahalanobis distances (in the quadratic form) between the internally represented reference points (RP: corresponding to either exemplars or prototypes) and the input stimuli as the measure of similarity between them. Thus, unlike other neural network models of category learning, GECLE does not necessarily assume that attention is allocated independently dimension-by-dimension. Rather, it assumes that humans in some cases might pay attention to correlations among feature dimensions. This allows GECLE to model processes interpretable as dimensionality reduction or mental rotation in the perception and learning of stimuli. Such processes may increase the interpretability of stimuli in categorization tasks. Another motivation for using the Mahalanobis distance is that the capability for paying attention to correlations among feature dimensions may be necessary for classification tasks defined on integral stimuli.

In the GECLE framework, the attention parameters (which are the diagonal and off-diagonal elements of the covariance matrices) can be considered as *shape* and *orientation* parameters for receptive fields or attention coverage areas of the reference points. It should be noted, however, that one can constrain GECLE to incorporate the "dimensional attention processes" assumption (i.e., attention is allocated independently on a dimension-by-dimension basis) by forcing the off-diagonal entries in the covariance matrices to be equal to zero.

Another unique feature of GECLE's attention mechanism is that it allows each reference point to have uniquely shaped and oriented attention coverage area, which is referred to as "local attention coverage structure" (Matsuka 2003). Again, one can impose a restriction on the model's attention mechanism by fixing all covariance matrices to be the same, which may be called "global attention coverage structure". Many NN models of category learning, ALCOVE (Kruschke, 1992) for example, incorporate the global attention coverage structure.

The local attention coverage structure model is complex, but may plausibly model attention processes in human category learning. For example, it allows models to be sensitive to one particular feature dimension when the input stimulus is compared with a particular reference point that is highly associated with category X, while the same feature dimension receives little or no attention when compared with another reference point associated with category Y. Thus the local attention coverage structure causes models to learn and be sensitive to within-cluster or within-category feature configurations, while the global attention coverage structure essentially stretches or shrinks input feature dimensions in a consistent manner for all RP receptive fields and all categories.

Another way of interpreting GECLE's capabilities for paying attention to correlations among feature dimensions and having local attention coverage structures is that the model learns to define what the feature dimensions are for each RP and to allocate attention to those dimensions independently. In contrast, for almost all previous adaptive models of category learning, the definition of the feature dimensions is static and supplied by individuals who use the models.

Some studies showed that humans learn much better in "filtration" tasks, in which information from only one dimension is required for (perfect) categorization, than in "condensation" tasks, in which information from two dimensions is required (e.g., Gottwald & Garner, 1975). This finding has been used as evidence that people pay attention to each dimension independently, rather than dependently (i.e., paying attention to correlations). Thus, a model paying attention to correlations or having diagonal attention coverage, as GECLE does, may not replicate filtration advantage. However, Matsuka (2003, 2004) successfully replicated the filtration advantage using a prototype based correlation-attentive GECLE with local attention coverage structure. He suggested that for a prototype based GECLE, the condensation stimuli require a stricter correspondence or synchronization between prototype search (i.e., shifting centroids of prototypes) and psychological scaling of the two feature dimensions (i.e., attention processes) as compared with the filtration stimuli. This is because the "correct" prototypes and "correct" scaling are defined by two dimensions in the condensation stimuli as compared to one dimension in the filtration stimuli.

In its natural form, the GECLE may be considered as a model using prototype internal representation, because it tries to learn to locate its reference points at the centers of each category cluster. However, with proper user-defined parameter settings, it can behave like a model with an exemplar-based internal representation.

Quantitative Descriptions (Algorithm)

The feedforward and learning algorithms of the GECLE are typical for implementation of the Generalized Radial Basis Function (Haykin, 1999; Poggio & Girosi, 1989, 1990). GECLE uses the following function to calculate the distances or similarity between internally represented reference points and input stimuli:

$$D_j^n\left(x^n, r_j\right) = \left(x^n - r_j\right)^T \Sigma_j^{-1}\left(x^n - r_j\right) \quad (1)$$

where x^n is an I-ruple vector representing an input stimulus consisted of I feature dimensions presented at time n, r_j, also an I-ruple vector, that corresponds to the centroids of reference point j, expressing its characteristics, and Σ_j^{-1} is the inverse of the covariance matrix, which defines the shape and orientation of the attention coverage area of reference point j. For a model with global attention coverage structure, there is only one global Σ^{-1} for all reference points.

The psychological similarity measures $D_j(x,r)$ cause some activations in internal "hidden" units or reference points (i.e., exemplars or prototypes). The activation of "hidden" basis unit j, or h_j, is obtained by any differentiable nonlinear activation transfer function (ATF), or

$$h_j = G\left(D_j(x,r)\right) \quad (2)$$

given that its first derivative $G'(\cdot)$ exists. An exponential function, $\exp(-cD_j(x,r))$, is an example of an ATF. The ATF must be a differentiable function, because GECLE uses a gradient method for learning, where the partial derivatives are used for updating the learnable parameters. However, it is possible to eliminate this restriction by incorporating a form of derivative-free learning algorithm such as stochastic learning (Matsuka & Corter 2004).

The activations of hidden units are then fed forward to output nodes. The activation of the kth output node, O_k, is calculated by summing the weighted activations of all hidden units connected to the output node, or

$$O_k = \sum_{j=1}^{J} w_{kj} h_j \quad (3)$$

where w_{kj} is the association weight between output node k and reference point j. The probability that a particular stimulus is classified as category C_k, denoted as P(C), is assumed equal to the activity of category k relative to the summed activations of all categories, where the activations are first transformed by the exponential function (Kruschke, 1992)

$$P(C) = \frac{\exp(\phi O_c)}{\sum_k \exp(\phi O_k)}. \quad (4)$$

ϕ is a real-value mapping constant that controls the "decisiveness" of classification responses.

GECLE uses the gradient method to update parameters. The error function is defined as the sum of squared differences between targeted and predicted output values (i.e., L_2 norm), or

$$E(w, r, \Sigma^{-1}) = \frac{1}{2}\sum_{k=1}^{K} e_k^2 = \frac{1}{2}\sum_{k=1}^{K}\left(d_k - O_k\right)^2 \quad (5)$$

Then the following functions are used to update parameters.

$$\Delta w_{jk} = \frac{\partial E}{\partial w_{jk}} = -\eta^w e_k h_j \quad (6)$$

where η^w is the learning rate for the association weights.

$$\Delta r_j = \frac{\partial E}{\partial r_j} = -\eta^r \sum_{k=1}^{K} e_k w_{jk} G'\left(D_j(x,r)\right)\Sigma_j^{-1}\left(x - r_j\right) \quad (7)$$

where $G'(\cdot)$ is a derivative of $G(\cdot)$. Equation 7 can be considered as a function that locates or defines prototypes of

stimuli. For the exemplar-based modeling η^r must be set to zero to maintain the static nature of reference points.

$$\Delta\Sigma_j^{-1} = \frac{\partial E}{\partial \Sigma_j^{-1}} = \eta^{\Sigma} \sum_{k=1}^{K} e_k w_{jk} G'\left(D_j(x,r)\right)\left(x - r_j\right)\left(x - r_j\right)^T \quad (8)$$

For models with global attention coverage structure, Equation 8 should be summed over both k and j.

Hierarchy of Constraints on Attention Parameters

There is a hierarchy of constraints that one can impose on the attention parameters Σ^1 to manipulate GECLE's attention mechanisms. There are two levels of uniqueness of Σ^1 (global and local attention coverage structure), in each of which there are three levels of constraints on entries in Σ. The following is a list of six possible levels of restriction. Note that regardless of the types of restriction, the entries (s_{im}) in Σ_j are assumed and constrained to satisfy the following conditions: $s_{ii} \geq 0$ & $|s_{im}| \leq$ MIN(s_{ii}, s_{mm}).

Global Attention Coverage Structures

A. Global Pure Radial (GPR): Constraints on Σ_j: $s_{ii} = s$, for all i: $s_{im} = 0$, for all $i \neq m$; $\Sigma_j = \Sigma$, for all reference points j.

B. Global Uncorrelated Non-radial (GUN): Constraints on Σ_j: $s_{im} = 0$, for all $i \neq m$; $\Sigma_j = \Sigma$, for all reference points j.

C. Global Correlated Non-radial (GCN): Constraints on Σ_j: $\Sigma_j = \Sigma$, for all reference points j.

Local Attention Coverage Structures

D. Local Pure Radial (LPR): Constraints on Σ_j: $s_{ii} = s$, for all i; $s_{im} = 0$, for all $i \neq m$.

E. Local Uncorrelated Non-radial (LUN): Constraints on Σ_j: $s_{im} = 0$, for all $i \neq m$.

F. Local Correlated Non-radial (LCN): Constraints on Σ_j: none.

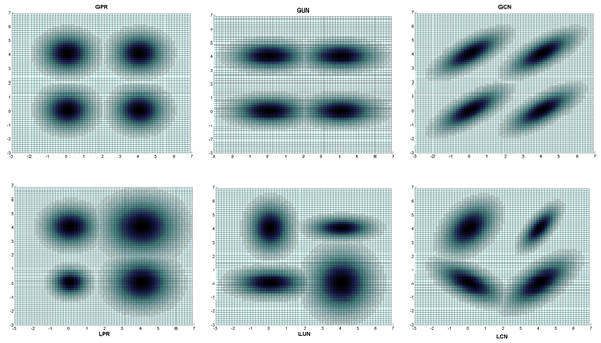


Figure 1. Six types of attention structures in the GECLE framework. Clockwise from top left. GRP, GUN, GCN, LCN, LUN, and LRP.

Simulations

In this section, three simulation studies were conducted to compare adaptive network models of category learning utilizing prototypes or exemplar internal representations using the GECLE framework. Here, a classical category learning study (Medin & Schaffer 1978) was replicated with several variants of GECLE. Simulation 1 reports the predictions by several GECLE models based on "optimal" parameter values. In Simulation 2, the general tendencies in some key aspects associated with the stimulus set were investigated with the same GECLE models used in Simulation 1. The plausibility of prototype models was further investigated using two variants of prototype-based GECLE in Simulation 3.

Simulation 1

In Simulation 1, I simulated category learning using the well-known Medin and Schaffer's 5/4 stimulus set (1978). Table 1 shows the schematic representation of the stimulus set. Eight different GECLE-based models were involved in the present simulation study. Among them there were seven prototype-based models (PB) with 2,3,4,5,6,7, or 8 prototypes and one exemplar-based model (EB) with all 9 unique exemplars. The global attention structure with dimensional attentional processes (i.e., GUN) was used for all eight models. They were run in a simulated training procedure to learn the correct classification responses for the training set. The models were run for 100 blocks of training, where each block consisted of a complete set of the training instances. The final parameter values used for each model were chosen by a simulated annealing method to minimize the objective function (i.e., sum of squared error: SSE) in reproducing the classification profile reported in the original Medin & Schaffer's work (1978). There are a total of 50 simulated subjects in each condition.

The following one-parameter exponential activation transfer function was used for the models:

$$h_j = \exp\left(-c \cdot D_j(x,r)\right)$$

One of the main interests of the present simulation study was how well the eight models could reproduce observed classification profile reported in Medin & Schaffer (1978). The other related interest was how well each model performs on stimuli A1 and A2 (see Table 1). These two stimuli have been considered to be very important and diagnostic, because PB and EB tend to give different predictions for these particular stimuli (e.g., Nosofsky & Zaki, 2002). Specifically, EB models are used to explain empirical results that show that humans are better able to categorize less "prototypical" A2 than more "prototypical" A1 (e.g., Medin & Schaffer 1978). Moreover, simulation studies (e.g., Nosofsky & Zaki 2002) indicate that EB gives a better fit for differential performance on these particular stimuli.

Table 1. Stimulus set used in Simulation 1

	Training Set					Transfer Set			
	Cat	D1	D2	D3	D4	D1	D2	D3	D4
A1	A	1	1	1	0	1	0	0	1
A2	A	1	0	1	0	1	1	1	1
A3	A	1	0	1	1	0	1	0	1
A4	A	1	1	0	1	0	0	1	1
A5	A	0	1	1	1	1	0	0	0
B1	B	1	1	0	0	0	0	1	0
B2	B	0	1	1	0	0	1	0	0
B3	B	0	0	0	1				
B4	B	0	0	0	0				

Results: Table 2 shows two fit indices for the eight models, namely SSE as an absolute fit index, and SSE multiplied by

the number of learnable parameters (NLP) as a (crude) relative fit index that may account for the model complexity. A pure prototype model (here a pure prototype is defined as a model that has as many RPs as the number of categories) performed worst before and after controlling for the model complexities. In addition, it failed to show the A2 advantage. Rather as in many previous studies, it predicted that A1 was easier than A2. However, other PB models performed well; PB8 resulted in the best absolute fit, and PB5 resulted in the best relative fit.

When the PB models are compared with the EB model, some PBs fit the observed profile better than EB, particularly after controlling for the model complexities. More interestingly, as the EB model, almost all PBs were able to predict the A2 advantage (Table 2, last column).

Although, this Medin and Schaffer 5/4 stimulus set has been used as evidence supporting exemplar-based models and undermining prototype-based models, the results of the present simulation study appear to show no competitive advantage of the exemplar-based model. Instead, some PB models were able to reproduce the observed classification profile and the A2 advantage equally successfully with smaller numbers of learnable parameters.

Table 2. Results of simulation 1

Model	NLP	NRP	SSE	SSE x NLP	A2-A1
PB2	16	2	0.1438	2.301	-5.633
PB3	22	3	0.0694	1.527	3.643
PB4	28	4	0.0361	1.011	5.444
PB5	34	5	0.0250	0.850	9.046
PB6	40	6	0.0215	0.860	2.663
PB7	46	7	0.0193	0.888	4.314
PB8	52	8	0.0182	0.946	3.273
EB9	58*	9	0.0201	1.166	8.011

NLP: Number of Learnable Parameters
NRP: Number of Reference Points (e.g. prototype or exemplar)
* Location parameters for exemplar were static & not subject for learning, but assumed that optimized locations were learned when the exemplars were created.

Discussion of Simulation 1: All GECLE models that were capable of learning to locate the reference points were interpreted as prototype-based models in the present simulation study. However, it might not have been a sensible interpretation for some of those models, particularly for models with larger numbers of prototypes (e.g., PB5 ~ PB8). That is, it does not seem logical to create eight prototypes from only nine unique stimuli. Rather, there may be better interpretations for these models. Two possible alternative interpretations are discussed below.

First, it might be more sensible to interpret PB GECLE with larger numbers of prototype as models utilizing "fuzzy" or modular prototypes (or simply modules) as the reference points (RP) in a combinatorial fashion: it tries to create and memorize modules (defined by or being prototypes of subsets of stimuli belonging to a particular category) that summarize characteristics of particular feature dimensions more correctly than the other feature dimensions for a particular category, and uses combinations of the module activations triggered by similarities between the modules and input stimuli for categorizing. This combinatorial coding seems to be a very efficient use of limited mental resources for categorizing virtually unlimited number of unique instances.

Alternatively, those models that were interpreted as prototype-based GECLE with many prototypes might have been utilizing RPs that were more sensible to be interpreted as probabilistic, partial, or erroneous exemplars, instead. That is, although the models might have tried to store correct exemplars in their memory, the process was not fully completed because of the limited mental resources, resulting in imprecise exemplars memorization, in which a particular feature of a particular exemplar was more correctly memorized than other features. Then, these imprecise exemplars were utilized for categorizing the stimuli.

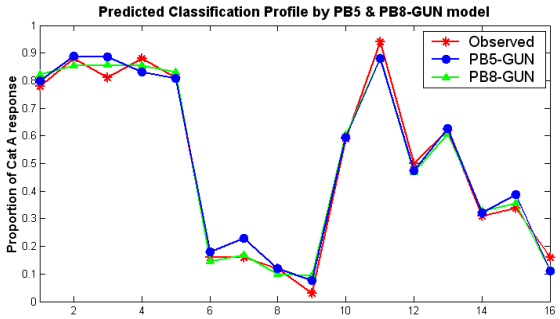


Figure 2. Predicted classification profiles by two best prototype based GECLE models (i.e., PB8-GUN: lowest absolute fit; PB5-GUN: lowest relative fit).

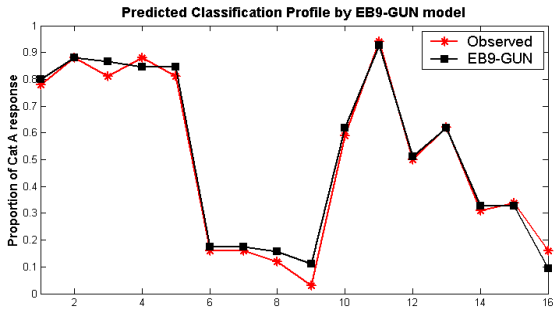


Figure 3. Predicted classification profiles by exemplar based GECLE model (i.e., EB9-GUN).

Simulation 2

Simulation 2 is a replication of Simulation 1 with 10,000 randomly chosen parameter configurations to investigate general tendencies in the A2 advantage by the same eight models used in Simulation 1. Here, the 10,000 simulated subjects with randomly assigned parameter values were trained to classify the 5/4 stimulus set. The ranges of parameters were [0.1 10] for c and ϕ, [0.001 1] for the three learning rates.

Results & discussion: Table 4 summarizes the results of Stimulation 2. In short, the A2 advantage was observed in almost all PB and EB models, indicating that the results of Simulation 1 are reasonably generalizable in this regard.

More interestingly, the EB model showed lesser magnitude of the A2 advantage than several PBs. This was mainly because EB9 learned to produce network output activations correctly with many parameter configurations (i.e., minimizing the error defined as Equation 5 perfectly) since the model was supplied the correct locations of all unique stimulus exemplars from the beginning of the training. This in turn, resulted in very small differences in classification responses for Stimuli A1 and A2, because the activations triggered by Stimulus A1 and A2 for the output nodes were almost identical (i.e., L_2 was minimized). This implies that any EB-based GECLE or any EB-based model such as ALCOVE would find this learning task (here, learning task does not correspond to categorization, but L_2 minimization, i.e., Eq. 5) easy because it can satisfactorily complete the task with virtually any parameter settings inasmuch as the locations of exemplars were well defined. Although this may be true if the condition of correctly memorizing exemplars is met, there is no guarantee for satisfying the condition in real human cognition. But, more likely, the condition would not be tenable for some people (i.e., some memorize exemplars more correctly and/or faster than other individuals). This difference in memorization ability may be one of the factors creating individual differences in category learning. This aspect of exemplar type modeling alone does not invalidate the assumption of exemplar-type internal representation, but it does suggest that exemplar-based (computational) models of categorization could be benefited from integrating an algorithm or quantitative explanation of how people learn and memorize exemplars.

On the contrary, exemplar theorists may argue that the upper limits of the randomly selected learning rate parameters (or the number of training epochs) were set unrealistically high. Although this argument is likely valid and thus the interpretation of the results may require some caution, it is still true that exemplar model may need to have learning algorithm for exemplar initialization, maintenance, and memorization.

Table 4. Results of Simulation study 2: Differences in classification accuracies for A2 and A1. (numbers of observed cases shown in parentheses).

Model	Overall	Classification Accuracy (CA) in training	
		100≥ CA >90%	90≥ CA >80%
PB2	1.011	-8.725(117)	-8.178(162)
PB3	2.184	0.056(250)	0.539(295)
PB4	2.521	0.331(556)	1.261(369)
PB5	3.071	0.885(905)	3.007(342)
PB6	2.711	0.661(1212)	3.816(365)
PB7	2.962	0.342(1690)	4.029(367)
PB8	2.446	0.330(2037)	2.885(393)
EB9	0.050	0.014(7660)	0.087(837)

Note: Observed classification accuracy for the training set is 0.85

Simulation 3

Simulations 1 and 2 showed that the pure prototype model, PB-2, accounted poorly for phenomena associated with the Medin and Schaffer's stimuli. However, these results might have resulted from incorrect assumptions about the prototype modeling. For example, I assumed that the locations of prototypes were continuously updated throughout the training, but in reality, people may quickly identify prototypes which may be less likely to be updated unless absolutely necessary. Another possible explanation is that people may have a uniquely shaped activation area for each prototype and/or pay attention to correlation among feature dimensions. For example, Matsuka (2003 & 2004) showed that there may be an interaction between types of internal mental representation and types of attention mechanism: the prototype-based model performed better when it incorporated unique attention structure with the capability of paying attention to dimensional correlations; whereas the exemplar-based model performed better with global attention structure with independent dimensional attention processes (i.e., no attention to correlations).

In the present simulation study, pure prototype modeling was reinvestigated using two variants of the original PB2 GECLE. The first one, SPB-2 is a static version of PB-2. That is SPB-2 is identical to PB2 appeared in Simulations 1 and 2, but the locations of prototypes were supplied from the beginning of the training and the learning rate for RPs was set to zero. Thus, this model resembles EB-based GECLE (except that RPs were prototypes) in that the locations of RPs were static. The second one, CPB2, is PB2-GECLE with the most complex attention mechanism, namely LCN (see Figure 1, lower right panel), having a unique receptive field for each prototype and the capability of paying attention to correlation.

For SPB2, the prototype for each category was created by averaging the feature values of each dimension of every object in a particular category, thus [0.8 0.6 0.8 0.6] for Category A and [0.25 0.5 0.25 0.25] for Category B. The rest of the procedures of the present simulation study follow those of Simulations 1 and 2.

Table 5a. Simulation 3: Results based on optimal parameters

Model	NLP	NRP	SSE	SSE x NLP	A2-A1
SPB2	16	2	0.1972	3.155	-9.208
CPB2	32	2	0.0377	1.206	11.130

Table 5b. Simulation 3: A2 advantage based on randomly drawn parameters.

Model	Overall	Classification Accuracy (CA) in training	
		100≥ CA >90%	90≥ CA >80%
SPB2	-2.346	-3.740 (2263)	-4.535(1920)
CPB2	2.931	-0.814(2215)	5.964(1505)

Results & discussion: A great decrease in SSE was obtained for CPB2 as compared with the original PB2, and after controlling for the model complexity by the simple linear adjustment (i.e., SSE x NLP) it performed nearly as good as EB9 (1.206 vs.1.166). In addition, unlike PB2, CPB2 was able to replicate the A2 advantage, and it was

shown to be generalizable to some extent in the second part of the present simulation study using the randomly drawn parameters (Table 5b). In contrast SPB2 performed worse than PB2 for replicating the observed classification profile. Moreover, SPB2 consistently failed to replicate the A2 advantage in the randomized simulation study.

Discussion on Simulations

Medin and Schaffer's 5/4 stimulus (1978) has been used as a benchmarking stimulus set for computational models of categorization and category learning, usually favoring exemplar models (e.g. Matsuka et al. 2003; Minda & Smith 2002; Nosofsky & Zaki, 2003). However, the results of the present simulation studies showed that several GECLE models with prototype internal representation performed as good as or better than the exemplar-based GELCLE. One type of those successful prototype-based GECLE was the model that created and utilized multiple *modular* prototypes for categorization. The modular prototype is a prototype defined by subsets of stimuli belonging to a particular category that summarize characteristics of particular feature dimensions more correctly than the other feature dimensions for the particular category (however, the modular prototypes may be interpreted as imprecise exemplars). The other type of the successful prototype-based GECLE was the one with uniquely shaped and oriented attention coverage areas and with the capability of paying attention to correlations among feature dimensions.

There are at least few concerns associated with the present simulation studies. First one, as discussed in Simulation 1, is that as the number of GECLE's reference points (RP) increases, it become philosophically difficult within the cognitive science paradigm to interpret what these RP are representing (e.g., modular prototypes vs. imprecise exemplars). The other concern is the way the numbers of learnable parameters were counted for the exemplar-based GECLE (see notes on Table 2). That is, in the present simulation studies, the location parameters of the exemplars were counted as learnable parameters. On one hand, the locations of exemplars may be learnable, because they are initialized at the "optimal" location without error. On the other hand, they may not be learnable, because they reside in static locations.

Conclusions

Generalized Exploratory model of human Category LEarning (GECLE) is a flexible and general framework for modeling human category learning that is capable of manipulating a limited number of assumptions independently and systematically. In the present study, the plausibility of two different assumptions about internal representation was investigated with GECLE using exemplar-model-friendly Medin & Schaffer 5/4 stimulus set (1978). The results of simulations showed no competitive advantage of previously favored exemplar-based modeling. Rather, they appeared to suggest some prototype models performed better than an exemplar model. In addition, the exploratory nature of GECLE yielded new plausible prototype-based adaptive models of category learning with different structures and model assumptions.

Although, several models were examined in some depth in the present research, the results were based only on a simulation of one empirical study. More simulation studies with several other stimulus sets should help identify models or assumptions with descriptive validities more accurately. In addition, measurements of several different cognitive processes associated with category learning, such as, attention allocation should be collected in empirical studies, in order to restrict model parameters and to better differentiate among models.

References

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation (2nd ed.).* Upper Saddle River, NJ: Prentice Hall.

Kruschke, J. E. (1992). ALCOVE: An exemplar-based connectionist model of category learning, *Psychological Review, 99.* 22-44.

Matsuka, T (2002). Attention processes in computational models of categorization. Unpublished Doctoral Dissertation. Columbia University, NY.

Matsuka, T. (2003). General exploratory model of human category learning. Accepted for publication.

Matsuka, T. (2004). Interactions between representation and attention processes in category learning. Poster presented at the 11th Annual Meeting of Cognitive Neuroscience Society. San Francisco.

Matsuka, T. & Corter, J.E (2004). Stochastic learning algorithm for modeling human category learning. Accepted for publication.

Matsuka, T., Corter, J. E. & Markman, A. B. (2003). Allocation of attention in neural network models of categorization. Under revision.

Medin, D.L. & Schaffer, M.M. (1978). Context theory of classification learning, *Psychological Review, 85*, 207-238.

Minda, J.P. & Smith, J.D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 275-292.

Nosofsky, R.M. & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 924-940.

Nosofsky, R.M., Gluck, M.A., Palmeri, T.J., McKinley, S.C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins

Poggio, T. & Girosi, F. (1989) A Theory of Networks for Approximation and Learning). *AI Memo 1140/CBIP Paper 31*, Massachusetts Institute of Technology, Cambridge, MA.

Poggio, T. & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science, 247,* 978-982.

Studying Human Face Recognition with the Gaze-Contingent Window Technique

Naing Naing Maw (nnmaw@cs.umb.edu)
University of Massachusetts at Boston, Department of Computer Science
100 Morrissey Boulevard, Boston, MA 02125-3393, USA

Marc Pomplun (marc@cs.umb.edu)
University of Massachusetts at Boston, Department of Computer Science
100 Morrissey Boulevard, Boston, MA 02125-3393, USA

Abstract

In eye-movement experiments using gaze-contingent windows, the stimulus display is continuously updated in response to the participant's current gaze position. Usually, a window is centered at the participant's gaze position and follows it wherever the participant looks. Within the window, all stimulus information is visible, while outside of the window at least part of the information is masked. In the present paper, we apply this technique to a face recognition task. By varying the size of the window, we gain insight into face recognition processes in humans and characterize the visual information on which face recognition relies. The results also motivate the use of gaze-contingent windows to study visual perception.

Introduction

Face recognition is a very important function of the human visual system and is fundamental to our complex social behavior. Therefore, it is not surprising that face recognition in humans has been extensively studied. Many studies concluded that face recognition relies more strongly on holistic information than does object recognition in general (see Maurer, Le Grand & Mondloch, 2002, for a review). In other words, ideally, the recognition process uses the entire visual information available from a face.

What makes faces so special in this regard? An important reason seems to be our everyday-life expertise in identifying people by their faces. It was found that people can be trained to recognize individual non-face objects, and thereby develop analysis patterns that are similar to those used in face recognition (e.g., Diamond & Carey, 1986; Gauthier, Williams, Tarr & Tanaka, 1998).

Other researchers presented participants with face images filtered by different spatial frequencies and found that only a rather narrow band of spatial frequencies (about 6 to 12 cycles per face width) contributes significantly to the recognition of a face (e.g., Näsänen, 1999). Again, this finding does not apply to non-face objects, even if individual objects of the same class are to be distinguished (Biederman & Kolacsai, 1997).

Moreover, face recognition has received considerable attention in machine vision research (e.g., Phillips, Moon, Rizvi & Rauss, 2000; Senior, Hsu, Mottaleb & Jain, 2002; Zhou., Krueger & Chellappa, 2003). Despite these immense efforts, however, even the currently best vision algorithms achieve face recognition rates that are far below the ones of a human observer.

We believe that a better understanding of the mechanisms underlying human face recognition will be beneficial to both the fields of medicine and machine vision. In the present study, we applied the sophisticated method of gaze-contingent windows to a psychophysical eye-movement study of a face recognition task in order to broaden our understanding of the underlying perceptual and attentional processes. The gaze-contingent window technique provides powerful experimental control and has been used extensively in reading, scene perception, and more recently in visual search studies (e.g. Bertera & Rayner, 2000; McConkie & Rayner, 1975; Pomplun, Reingold & Shen, 2001; Saida & Ikeda, 1979; see Rayner, 1998, for a review).

In most of its applications, this technique obscures all objects from view except those within a certain window that is continually centered on the participant's current gaze position. The window position changes across fixations to follow the gaze position. For example, in a study by McConkie and Rayner (1975), participants read text that was masked outside a visual window that included the fixated character and a number of characters to the left and to the right. Only the text within the window was legible. The visual span in reading was assessed by varying the window size across trials and determining the smallest window size that allowed participants to read with normal speed.

In the present study, participants were presented with images of famous and non-famous faces and had to indicate whether they recognized the displayed person or not. While viewing the images, a gaze-contingent window was administered with its size varying across trials. This allowed us to address the following questions: First, to what extent does face recognition rely on the simultaneous availability of the entire face features? Second, from which positions in the image and in what manner do participants acquire information about a face when their peripheral vision is restricted? It is well known that saccades during unrestricted face viewing tend to be aimed at the region formed by the eyes, nose, and mouth (e.g., Yarbus, 1967). However, where and how do participants gather information if they have to do it sequentially and be as efficient as possible? Third, how can the moment of recognition be characterized? Is it possible to determine this moment based on psychophysical data?

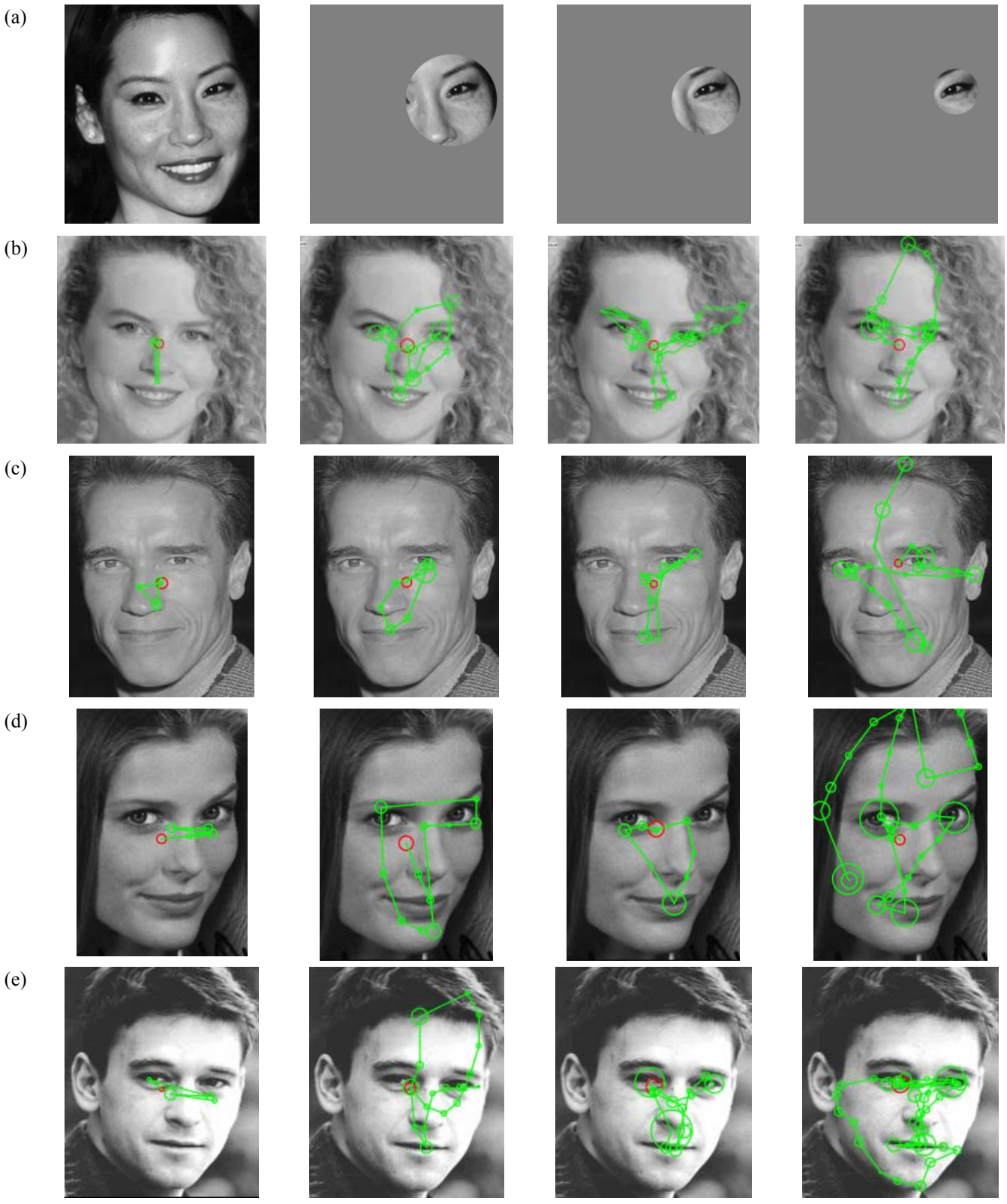


Figure 1: Sample stimuli used in the present study – each column represents one of the viewing conditions. From left to right column: unrestricted, large window, medium window, and small window. (a) Illustration of the different window sizes. (b) - (e) Sample gaze trajectories for each of the four stimulus categories (from top to bottom row): famous females, famous males, non-famous females, and non-famous males. Fixations are shown as circles with their size indicating fixation duration; the initial fixation is displayed in red color.

Method

Participants. Twenty students of the University of Massachusetts at Boston (ten females and ten males) participated in the present study. All of them had normal or corrected-to-normal vision. They were naïve with respect to the purpose of the study and were paid $10 for their participation.

Materials. We prepared 80 face images to serve as stimuli – 20 in each of the following four categories: famous females, famous males, non-famous females, and non-famous males. We chose the most popular American actresses and actors for the "famous" categories, while foreign actresses and actors, who had never appeared in international movies, were chosen for the "non-famous" categories. These grayscale images subtended an area of about 18° horizontally and 24° vertically on the screen of a 21-inch monitor. In the gaze-contingent window trials, the display area outside a circular, gaze-centered window was replaced with plain gray color. Four different viewing conditions were included in the experiment: unrestricted, large window (diameter of 8.2°), medium window (diameter of 5.5°), and small window (4.1°). These window sizes are illustrated in Figure 1a.

Apparatus. Eye movements were measured with an SR Research Ltd. EyeLink-II system. After a calibration procedure that was typically completed in less than a minute, gaze-position error was below or equal to 0.5 degrees of visual angle. The temporal resolution of the system was 2 ms. The gaze-contingent window followed the participant's gaze position with an average delay of 12 ms.

Procedure. Prior to each trial, participants were asked to fixate a marker in the center of the display. Following a button press, a face display was presented. As soon as participants had decided whether the depicted person was famous or non-famous, they terminated the trial by pressing one out of two buttons indicating their decision. Each participant was presented with each of the 80 stimuli exactly once, resulting in 80 trials per participant. The trials were administered in eight blocks of ten successive trials. Each of the four viewing conditions was applied in two of these blocks. The order of blocks and stimuli as well as the combination of stimuli with viewing conditions was systematically varied across participants.

Results and Discussion

Figures 1b to 1e show sample gaze trajectories for different stimuli across the four viewing conditions. Notice that the four trajectories for the same stimulus were generated by different participants, because each participant saw each stimulus only once. Two things can clearly be observed: First, in the unrestricted viewing condition, only a few central fixations were performed; the parafoveal and peripheral information of most of the face seems to be sufficient for successful face recognition. Second, when the gaze-contingent window was implemented, participants produced more fixations and directed them also at features that would normally not require foveal inspection, such as the hair or the ears, but obviously hold important information for the face recognition process. This effect of the gaze-contingent window on the eye-movement patterns generally increased with decreasing window size.

The quantitative analysis of the empirical data included the "standard" variables response time, proportion of correct responses, fixation duration, and saccade amplitude, but also the variables area coverage per trial and relative pupil size (see below). Interestingly, four-way analyses of variance (ANOVAs) for each of these variables (factors: viewing condition, stimulus recognizability, stimulus gender, and participant gender) revealed no significant effect by the factors stimulus gender or participant gender or their interaction. In other words, the gender of the participants or the people shown in the stimuli had no significant influence on any of the obtained variables. Therefore, in the following analyses, data were collapsed over these factors and only two-way ANOVAs (factors: viewing condition and stimulus recognizability) were conducted.

Response time was found to significantly depend on the viewing condition, $F(3; 57) = 80.93$, $p < 0.001$ as well as on recognizability (famous faces vs. non-famous faces), $F(1; 19) = 8.80$, $p < 0.01$. The interaction between the two factors also reached significance, $F(3; 57) = 2.77$, $p = 0.05$. As can be seen in Figure 2a, response time increased with smaller window size for both famous faces (no window: 1.90 s; large window: 4.57 s; medium window: 6.57 s; small window: 9.34 s) and non-famous faces (1.82 s, 5.69 s, 7.60 s, and 11.33 s). With more severe viewing restriction, response time became increasingly longer for non-famous faces as compared to famous ones. The pattern of these findings was expected, because restricting the participants' parafoveal and peripheral vision obviously makes their task more difficult. Detecting a familiar face should on average be faster than deciding that a face is unfamiliar, because before a negative decision can be made, all reasonable possibilities for a match have to be considered. With a smaller field of view, this effect increases, because more information needs to be obtained for a negative decision. The only unexpected finding is the pure magnitude of the response time difference imposed by the window manipulation.

The *proportion of correct responses* was also significantly influenced by the viewing condition, $F(3; 57) = 21.92$, $p < 0.001$, while there was neither an effect by recognizability, $F < 1$, or an interaction between the factors, $F < 1$. As shown in Figure 2b, there was no tradeoff between participants' response time and accuracy, but actually the opposite effect was found: The proportion of correct responses strongly decreased with more severe viewing restriction (91.3%, 78.0%, 67.8%, and 68.7%). Given that a participant who just gives random responses would reach an average of 50% correct responses, this result indicates a dramatic decrease in performance accuracy. The fact that there was no significant difference between famous and non-famous faces demonstrates that participants were not biased towards giving positive responses or towards giving negative responses. The latter would have been found if participants had just clicked the "non-famous" button

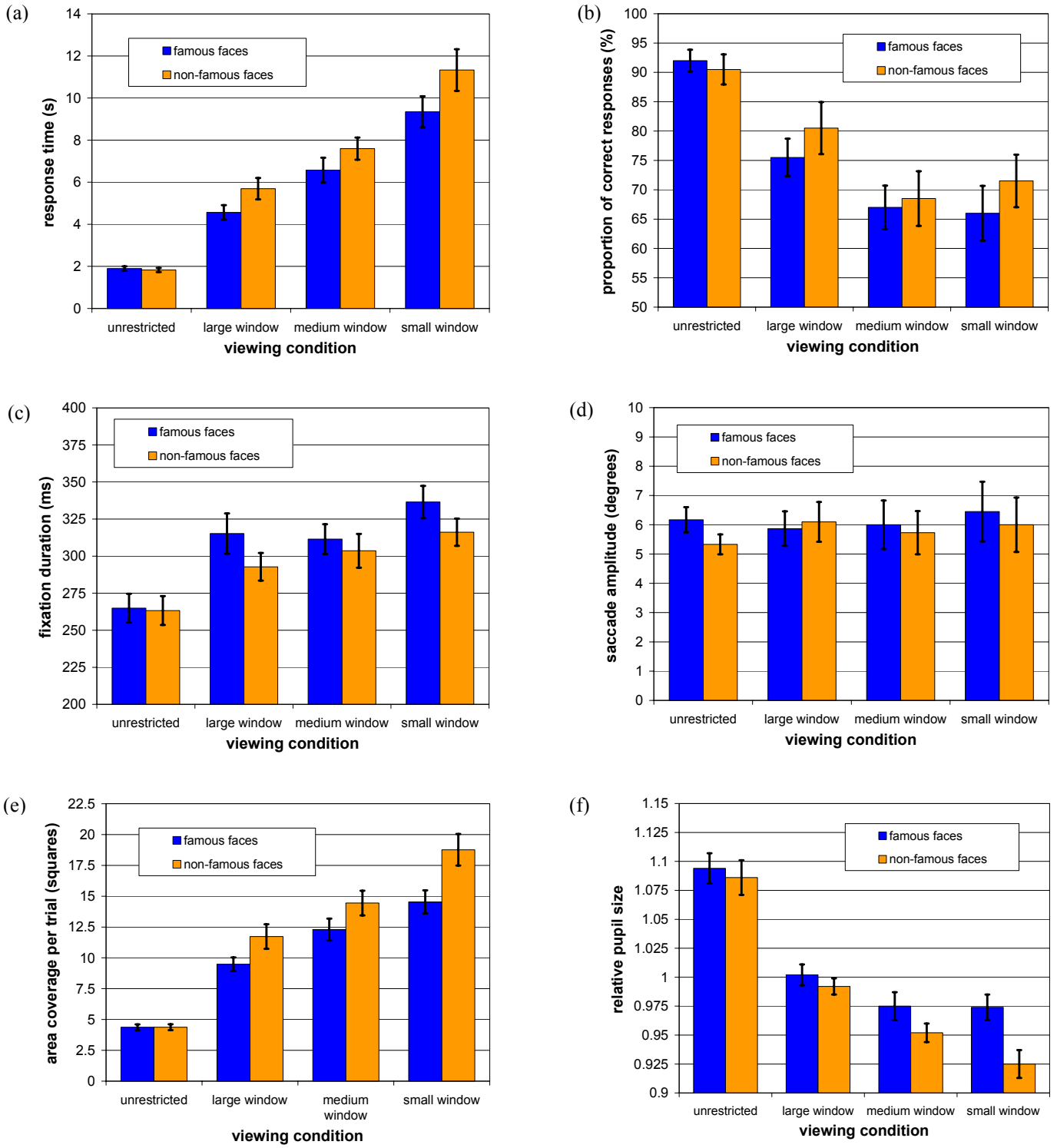


Figure 2: Psychophysical measurements obtained in the present study: (a) response time, (b) proportion of correct answers, (c) fixation duration, (d) saccade amplitude, (e) area coverage per trial, and (f) relative pupil size. Notice that the depicted interval of the variable values does not always start at 0.

whenever they did not recognize a face immediately, instead of making an effort to verify their first impression. Therefore, the analysis of the proportion of correct responses provides evidence for the participants of the present study to perform their task according to the instructions.

A variable that is analyzed in almost all eye-movement experiments is *fixation duration*. The duration of a fixation indicates how long the local information in a display was processed, which includes the duration necessary to program the subsequent saccade. In the present experiment, we found a significant effect by the viewing condition on fixation duration, $F(3; 57) = 23.15$, $p < 0.001$. The factor recognizability also exerted a significant effect, $F(1; 19) = 7.84$, $p < 0.05$, while there was no interaction between the two factors, $F(3; 57) = 1.33$, $p > 0.2$. In Figure 2c, fixation duration can be seen to increase with smaller windows for both famous faces (266 ms, 315 ms, 312 ms, and 336 ms) and non-famous faces (263 ms, 293 ms, 304 ms, 316 ms). Since smaller gaze-contingent windows reduce the amount of information that is available near the fixation point, there are two likely factors that determine this pattern of results: First, the smaller the window, the more effort is required to merge the available visual information with the current representation of the face in visual working memory. Second, for efficient task performance, a smaller window increases the necessity to aim saccades at locations where the most useful information is assumed to be located; consequently, the programming of these saccades requires more time. The finding of longer fixations for famous than for non-famous faces (see Figure 2c) is more puzzling; one possible explanation is that the recognition process itself causes one or more prolonged fixations – such fixations do not occur in non-famous faces. We conducted another analysis, reported later in this section, to test this hypothesis.

Another variable is routinely analyzed in eye-movement studies, namely *saccade amplitude*, measuring the length of saccades in degrees of visual angle. Short saccades can indicate fine-grained processing of local information, whereas long saccades often signify low information content or superficial scanning of local visual input. In the present context, we might expect saccades to become shorter with decreasing window size in order to uncover contiguous patches of an image. However, no influence by the viewing condition on saccade amplitude was found, $F < 1$, and recognizability showed no effect either, $F(1; 19) = 2.83$, $p > 0.1$. There was no interaction between the factors, $F(3; 57) = 1.74$, $p > 0.1$. As shown in Figure 2d, saccade amplitude (famous faces: 6.17°, 5.87°, 6.00°, and 6.45°; non-famous faces: 5.33°, 6.10°, 5.73°, and 6.00°) is unaffected by the recognizability of faces or the viewing condition. This result suggests that saccadic endpoints are not chosen to completely inspect local areas of the image by patching together adjacent pieces of visual information. Instead, saccades are aimed at positions that are assumed to contain the most significant information for the famous versus non-famous decision. This interpretation is in line with the view that fixations become longer with decreasing window size because of increased effort in the programming of saccades.

If it is true that the additional saccades induced by smaller gaze-contingent windows are aimed at "foraging" for useful information wherever in the image it is suspected, rather than inspect focused areas more thoroughly, then this behavior should be quantitatively reflected in the eye-movement data. In order to investigate this, we analyzed the eye-movement variable *area coverage per trial*. To compute this variable, we divided the stimulus area into 9 (horizontally) by 12 (vertically) squares. For each trial, we calculated the area coverage as the number of different squares that contained at least one fixation. This number should increase with the amount of "information foraging" (as opposed to focused examination) performed by participants. We found the area coverage per trial to be significantly influenced by the viewing condition, $F(3; 57) = 96.42$, $p < 0.001$, and by recognizability, $F(1; 19) = 13.50$, $p < 0.01$. The interaction between these factors was also significant, $F(3; 57) = 5.78$, $p < 0.01$. As shown in Figure 2e, the pattern of results for area coverage per trial is very similar to the one for response times (Figure 2a): Area coverage per trial increases strongly with decreasing window size, both for famous faces (4.37, 9.49, 12.30, and 14.54 squares) and for non-famous ones (4.38, 11.73, 14.45, and 18.77 squares). Area coverage is also smaller for famous faces than for non-famous ones, with this difference being more pronounced for smaller windows. This finding supports our assumption that peripheral restriction of information induces an exploration strategy that guides saccades towards the most promising new locations in the stimulus.

Finally, a variable that usually receives less attention, although it is a "by-product" of video-based eye tracking, is *pupil size*. Pupil size is known to depend on factors such as the luminance in the visual field or the cognitive activation of a person (e.g. Kahneman, 1973; Pomplun & Sunkara, 2003). Since we were only interested in relative changes in pupil size, we divided all measurements by the participants' initial pupil size after the eye tracker setup. Consequently, values above 1 indicate a dilated pupil, while values below 1 signify a contracted pupil. We found a significant influence of the viewing condition on pupil size, $F(3; 57) = 42.58$, $p < 0.001$, and also a significant influence by recognizability, $F(1; 19) = 8.37$, $p < 0.01$. There was no interaction, $F(3; 57) = 1.81$, $p > 0.1$. Figure 2f illustrates that pupil size clearly decreases with smaller window size for both famous faces (1.094, 1.002, 0.975, and 0.974) and non-famous faces (1.086, 0.992, 0.952, 0.925). This could be explained by the smaller amount of information that is available for processing at any given time during the trial. However, why was the pupil larger for famous than for non-famous faces?
A possible explanation is that the *moment of recognizing* a famous face has an impact on the eye-movement data. In an earlier, unpublished study on a face search experiment, we observed that the moment of recognizing the presence of a face tended to coincide with a prolonged fixation and temporarily dilated pupils. If such a reaction also occurs at the moment of recognizing a known face, then in the present study this would not only explain the greater in pupil size, but also the longer fixations measured for famous as compared to non-famous faces. To test this hypothesis, we

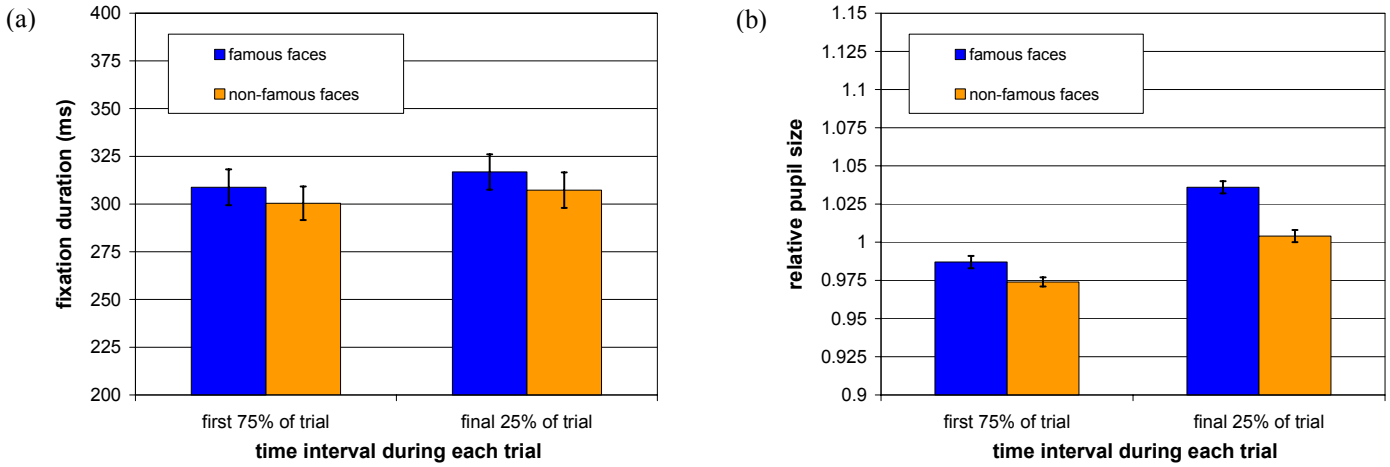


Figure 3: (a) Fixation duration and (b) pupil size analyzed separately during the first 75% and the last 25% of each trial.

separated the data for the first 75% of the duration of each trial from the final 25%. The moment of recognition is most likely to occur during the last 25% of a trial, so if we found the recognizability effects on fixation duration and pupil size to only occur during the last phase, as indicated by an interaction of the factors recognizability and time interval, it would support the moment of recognition interpretation.

For each of the two variables, we therefore conducted a two-way (recognizability and time interval) ANOVA. While there was no significant interaction for fixation duration (Figure 3a), $F < 1$, it was found for pupil size, $F(1; 19) = 6.92$, $p < 0.05$. Figure 3b shows that the difference in pupil size between famous and non-famous faces clearly increases during the final 25% of the trials. This finding suggests that the moment of recognizing a face may be associated with pupil dilation.

All in all, the present study has shown that the simultaneous availability of the entire face information is crucial for efficient face recognition, which supports the view that face recognition is a holistic process that heavily relies on parafoveal and peripheral input. Restricting this input has provided us with insight into which essential information was eliminated and needed to be foveally processed instead. Finally, we have found that the moment of recognizing a face may be indicated by a dilated pupil. This line of research is only at its beginning, and we hope to inspire other researchers to consider the technique of gaze-contingent windows for their face recognition and other perceptual studies.

References

Bertera, J.H. & Rayner, K. (2000). Eye movements and the span of the effective visual stimulus in visual search. *Perception & Psychophysics*, 62, 576-585.

Biederman, I. & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London*, 352, 1203-1219.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, 115, 107-117.

Gauthier, I., Williams, P., Tarr, M.J. & Tanaka, J. (1998). Training "greeble" experts: A framework for studying expert object recognition processes. *Vision Research*, 38, 2401-2428.

Kahneman, D. (1973). *Attention and effort*. New Jersey: Prentice Hall.

Maurer, D., LeGrand, R. & Mondloch, C.J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, 6, 255-260.

McConkie, G.W. & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17, 578-586.

Näsänen, R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Research*, 39, 3824-3833.

Phillips, P.J., Moon, H., Rizvi, S.A. & Rauss, P.J. (2000). The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 1090-1104.

Pomplun, M., Reingold, E.M. & Shen, J. (2001). Peripheral and parafoveal cueing and masking effects on saccadic selectivity in a gaze-contingent window paradigm. *Vision Research*, 41, 2757 - 2769.

Pomplun, M. & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In D. Harris, V. Duffy, M. Smith & C. Stephanidis (Eds.), *Human-Centred Computing: Cognitive, Social, and Ergonomic Aspects.* Vol. 3 of the Proceedings of the 10th International Conference on Human-Computer Interaction, HCI 2003, Crete, Greece.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.

Saida, S. & Ikeda, M. (1979). Useful visual field size for pattern perception. *Perception & Psychophysics*, 25, 119–125.

Senior, A., Hsu, R., Mottaleb, M.A. & Jain, A.K. (2002). Face detection in color images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 696-706.

Yarbus, A.L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Zhou, S., Krueger, V. & Chellappa, R. (2003). Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91, 214-245.

The Recognition Heuristic: Fast and frugal, but not as simple as it seems.

Rachel McCloy (r.a.mccloy@reading.ac.uk)
School of Psychology, University of Reading,
Reading, U.K., RG6 6AL.

C. Philip Beaman (c.p.beaman@reading.ac.uk)
School of Psychology, University of Reading,
Reading, U.K., RG6 6AL.

Abstract

Two experiments examine the use of the recognition heuristic which states that, in the absence of other information, individuals make judgments on the basis on recognition alone. This has been shown to be adaptive (Borges, Goldstein, Ortmann & Gigerenzer, 1999) and in Experiment 1 we demonstrate that the heuristic is reliably employed when participants are placed under time pressure. Experiment 2 considers a possible confound of the adaptive recognition heuristic with a less-adaptive recognition-preference strategy and shows that both may be employed but that the recognition-preference strategy is not sufficient to account for the recognition heuristic. We discuss the implications of our results for the recognition heuristic and the rest of the adaptive toolbox (Gigerenzer & Todd, 1999).

Introduction

A recent approach to human judgment and rationality put forward by Gigerenzer and colleagues (Gigerenzer, 2000; Gigerenzer & Todd, 1999) emphasizes the real-time constraints of many decision-making and reasoning tasks. In doing so they have suggested that many so-called "biases" in human judgment are actually adaptive within real-world situations. The approach uses "fast and frugal" heuristics that have been shown to be highly effective in a number of situations (Goldstein & Gigerenzer, 1999; Borges et al., 1999) even when compared to more sophisticated methods that take into account multiple sources of information.

The recognition heuristic is one such strategy and is, furthermore, the first step in a number of fast and frugal strategies within what has been termed the adaptive toolbox (Gigerenzer & Goldstein, 1996). Simply stated, the recognition heuristic provides the following rule of thumb: "*If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value"* (Goldstein & Gigerenzer, 1999, p. 41). So, for example, if an experimental participant is asked to judge which of two cities has the larger population, the participant will be following the recognition heuristic if they choose the city which they recognize. This leads to the less-is-more effect whereby participants using the recognition heuristic outperform other participants who recognize both cities and should, therefore, have more information available upon which to base their decision. The reason for this is that recognition of city correlates with the size of the city and hence may be a more effective cue than those used by more knowledgeable participants.

Having demonstrated the usefulness of the heuristic, Goldstein & Gigerenzer (2002) examined whether the heuristic was actually employed in practice. Data from 22 participants showed that all of them produced choice behavior consistent with use of the recognition heuristic. These data were, however, disputed by Oppenheimer (2003) who noted that the American participants tested by Goldstein & Gigerenzer may have accessed information other than mere recognition in making their choices. Goldstein & Gigerenzer required their participants to select the larger of the two in pairs of German cities. Oppenheimer suggested that the stimuli employed conflated recognition with knowledge that the recognized city was one of the largest cities in Germany.

In his study, Oppenheimer presented participants with towns or cities that were local to them and that were known to be small. In doing so, Oppenheimer demonstrated that the recognition heuristic is not an inevitable strategy when faced with forced-choice tasks where only one of the choices is recognized. Oppenheimer's participants proved smarter than the recognition heuristic by choosing the recognized city significantly less often than would be expected by chance. However, the differences between the studies by Goldstein & Gigerenzer (2002) and by Oppenheimer (2003) go beyond the choice of stimuli. One aim of the current paper is to consider how differences in procedure may have contributed to the reported contradictions in choice behavior between the two studies. In doing so we will provide a more balanced view of the place of the recognition heuristic in decision-making generally and in the adaptive toolbox in particular.

The key methodological difference between the two studies was the time pressure that participants experienced. In the Oppenheimer study (Experiment 1), participants made 10 choices over a five-minute period, an average of 30 seconds for each choice. In Oppenheimer's Experiment 2, participants were given a week to return the booklet containing their answers. In contrast, in the Goldstein & Gigerenzer study (Experiment 1), participants made between 300 and 435 choices during a single experimental session. Although Goldstein and Gigerenzer did not specify how long their participants had to complete the task, it is likely that their participants had substantially less time per

choice than the 30 seconds for each choice taken by participants in the first Oppenheimer study (as, if this were the case, participants who had 435 choices to make would have taken over 3 ½ hours). We suggest that participants will be much more likely to use so-called "fast and frugal" strategies when tasks put them under time pressure. This situation basically reinstates the constraints under which boundedly rational approaches such as the use of fast and frugal heuristics are presumed to operate (Simon, 1956). It alos provides an alternative explanation, besides the difference in stimuli, of why participants were much more likely to use the recognition heuristic in the Goldstein and Gigerenzer study than in Oppenheimer's experiment.

In the two experiments that we report we therefore replicated the general procedure of Oppenheimer (2003) but gave participants a strict time limit for the experimental session. In order to address the issue of the confound in the stimuli identified by Oppenheimer we chose English towns or cities whose soccer teams played in the UK First Division, not in the Premier League, as the recognizable stimuli. The town or city names would thus have been familiar to the participants without being considered a large or major city as large cities in the UK (e.g., London, Liverpool, Manchester) tend to have soccer teams in the Premier League. The city names used as the (hopefully) unrecognizable stimuli were the fictional cities invented by Oppenheimer, all of which had made-up but (to UK participants) foreign-sounding names.

The general situation experienced by the participants therefore is one where the recognition heuristic is applicable (only one of the names is recognized) and there is no other information upon which to base a choice. The recognized name is not known to be a particularly large city (unlike the Goldstein & Gigerenzer study). Equally, the participant has little time with which to consider what they know of the recognized town or city or attempt to infer anything regarding the unrecognized city (unlike the Oppenheimer study). Under these circumstances the recognition heuristic is the only tool available in the adaptive toolbox. In Experiment 1 we consider whether, with these potential confounds identified and controlled for, participants will make use of the recognition heuristic. Failure to observe the use of the single simplest heuristic in the toolbox under such circumstances would be a severe setback to the fast and frugal heuristics research agenda.

Experiment 1

Participants and Procedure

The participants were 50 adult volunteers. The 30 men and 20 women who took part had an average age of 28 years (range 17-62; standard deviation 11.2). Each participant was presented with a four-page experimental booklet. The instructions told them that they would be presented with pairs of names of towns, and that their task was to circle the town with the largest population in each pair. Participants were given one minute to complete the task, timed with the stopwatch. In order to encourage them to work quickly, participants were given updates on the time at 15-second intervals. On completion of the first part of the task participants were then given a list of all of the towns used in the experiment (both real and fictional) and were asked to circle those out of the list that they recognized.

Materials and Design

The materials used in this experiment were based on those used by Oppenheimer (2003). We created a group of stimuli where participants could show the recognition heuristic by pairing the names of 10 real English towns with the names of 10 fictional towns (taken from Oppenheimer; see Appendix). The English towns were selected from the list of towns with First Division soccer teams, and each was paired with three different fictional towns, giving 30 recognition heuristic items in all. In addition, we created two groups of filler items. The first group consisted of pairs of real towns and cities taken from a list of 8, which contained four international towns/cities (e.g., Limerick) and four English towns (e.g., Bradford). Each participant received 10 pairs of this type. The second type of filler item consisted of pairs of the fictional towns. Participants each received 9 pairs of this type[1]. Therefore, each participant received 49 choice pairs in total, of which 30 were of the critical recognition heuristic type. The order of presentation of these pairs was randomized across participants.

Results

Coding Some participants did not complete all 49 choices in the allotted time. In addition, some participants either failed to recognize a real place name, or erroneously recognized a fictional place name. Therefore, on the basis of the number of items that they had completed, and their responses to the recognition task, we calculated for each participant: (a) the number of times that they could have used the recognition heuristic, (b) the number of times that they did use the recognition heuristic. The second figure divided by the first gave us a figure for the proportion of responses that conformed to the recognition heuristic.

Analysis One sample t-test showed that, by participant, the proportion of responses attributed to the recognition heuristic was significantly greater than would be expected by chance, $t = 3.55$, $df = 49$, $p = .001$ (2-tailed). These data are shown in Figure 1 (overleaf).

[1] There should have been 10 pairs of this type, but, due to a printing error with the materials, participants only received 9 pairs.

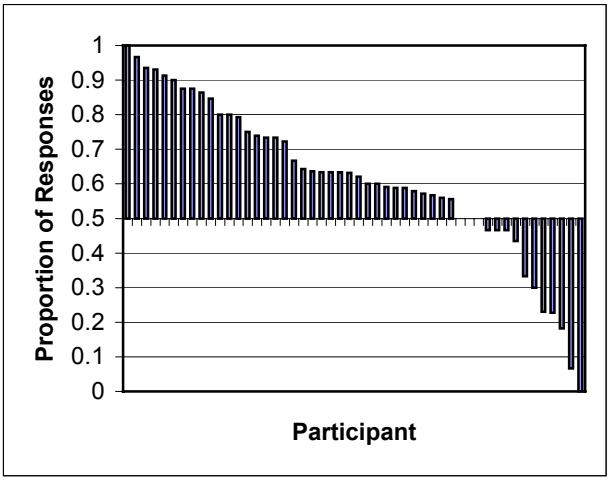


Figure 1: Proportion of Recognized Items Chosen by Participant (Experiment 1).

Further analysis of the choices made by individual participants shows that 23 out of the 50 participants showed evidence of use of the recognition heuristic at levels greater than chance, binomial $z > 1.28$, $p < .05$. Interestingly, a further 6 participants showed the reverse pattern, using the recognition heuristic significantly less often than chance, binomial $z > 1.66$, $p < .05$.

Discussion

These data confirm Goldstein & Gigerenzer's contention that participants use recognition in choice behavior when no other information is available. Use of this recognition heuristic may, however, be limited to situations when the participant is under time or other pressure. We should note that our participants did not use the recognition heuristic as frequently as those of Goldstein and Gigerenzer. This may have been due to the materials we used, and hence may provide some support for Oppenheimer's position. Additionally, not all the participants used the recognition heuristic consistently in their responses in our experiment and a significant subgroup of participants appeared to be using quite the opposite strategy. This was confirmed by the spontaneous self-reports of some participants. There are drawbacks to analyzing individual participants' data in this manner, for example, if participants were responding at random we might expect some individual participants to appear to use the recognition heuristic purely by chance. However, we are following a precedent in the literature (Goldstein & Gigerenzer, 2002; Rieskamp & Hoffrage, 1999) in attempting to identify individual strategies rather than averaging over potentially very different strategies. To answer some of these questions, we therefore ran a further experiment to examine whether altering the form of the question for the same choice stimuli would influence use of the recognition heuristic.

One possible explanation of our data is that participants, rather than using the recognition heuristic in the manner suggested by Goldstein & Gigerenzer (1999), were using recognition in a different way. For example, one participant reported deliberately choosing city names he did not recognize on the assumption that these foreign-sounding cities must be larger than the local towns that he knew. This use of recognition could explain the pattern of choice displayed by those participants who showed significantly less choice of recognized towns than would be expected by chance. This explanation is consistent with the data reported by Oppenheimer (Experiment 1).

If participants are capable of using recognition in a strategic and less rigid way than suggested in the formulation of the recognition heuristic, we would expect to see recognition effects not only in judgments of which of two cities is the larger but also in judgments of which of two cities is the smaller. According to the formal account of the recognition heuristic asking which of two cities is the smaller is equivalent to asking which is the larger. So participants would, paradoxically, be expected to use the recognition heuristic to choose the unrecognized city (since it is inferred that the unrecognized city is smaller of the two). However, if participants merely choose the recognized city because of some learned preference (e.g., Zajonc, 1968) or strategy we might expect them to continue to choose the recognized city. This hypothesis is tested in Experiment 2.

Experiment 2

Method

The participants were 42 adult volunteers. The 24 men and 18 women who took part had an average age of 21 years (range 18-48; standard deviation 4.5). The procedure and materials for this experiment were the same as that of Experiment 1. The only difference being that instead of making judgments of which of two towns was the larger, participants made judgments of which of two towns was the smaller.

Results

Coding We once again assessed participants' usage of the recognition heuristic taking into account the number of items completed and the participants responses to the recognition task in calculating a proportion of choices of the recognized item.

Analysis In this experiment, one sample t-test by participant failed to show that choice of recognized item varied significantly from chance, $t = 1.0$, $df = 41$, $p > .05$. These data are shown in Figure 2. Examination of Figure 2 also suggests, however, that some individual participants did use the recognition heuristic reliably in their responses.

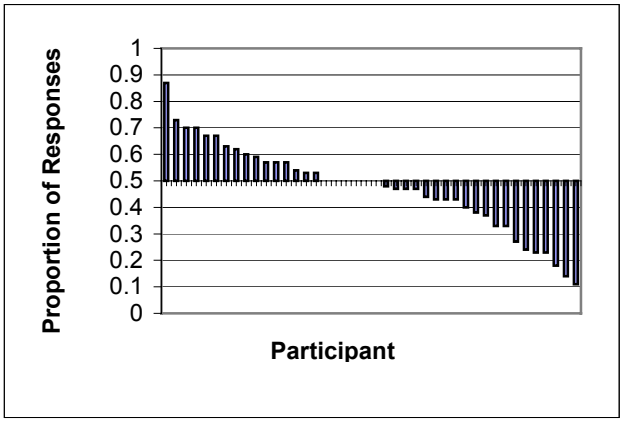


Figure 2: Proportion of Recognized Items Chosen by Participant (Experiment 2).

Further analysis of the choices made by individual participants shows that 10 out of the 42 participants showed evidence of use of the recognition heuristic at levels greater than chance, binomial $z > 1.28$, $p < .05$. A further 7 participants showed the reverse pattern, using the recognition heuristic significantly less often than chance, binomial $z > 1.28$, $p < .05$.

Discussion

The results of this experiment are intriguing because exactly the same stimuli and presentation conditions were used as in Experiment 1 yet we find a different pattern of results. The choice that participants needed to make was also identical to the previous experiment - judging the relative sizes of two towns or cities. The only thing that changed was the framing of the question, from asking which of the two was larger, to asking which was smaller. Some participants ($n = 10$) used a recognition heuristic to judge the smaller of the two towns, however, others ($n = 7$) used a diametrically opposed strategy. Consequently, the sample as a whole did not significantly differ from chance in their choice behavior. In this experiment, therefore, although all the preconditions for using a recognition heuristic were met, only a minority of participants did so.

Analysis of individuals' data showed that some participants *did* reliably use the recognition heuristic in their choices. It also showed that, as in the previous experiment, some participants used the even simpler strategy of always picking a town they recognized, regardless of the framing of the question. However, only a small group of participants appear to use this strategy, which, in the previous experiment, would have been indistinguishable from the recognition heuristic.

General Discussion

The recognition heuristic is an adaptive strategy in decision-making because of the correlation between recognition and magnitude. When a choice is made between the larger, or the more numerous, of two items it is frequently the case that the recognized item is, in fact, the larger or more numerous of the two. This is formalized within the recognition heuristic by stating that "when an individual only recognizes one of two items, the individual will judge the recognized item to be greater in whatever dimensions are positively correlated with recognition" (Oppenheimer, 2003, p. B2; see also Goldstein & Gigerenzer, 1999; 2002). Oppenheimer questioned the unthinking use of the recognition heuristic in his study. However, both Oppenheimer's study and the earlier reports by Goldstein & Gigerenzer confound choosing the recognized object because of the inferred correlation between recognition and magnitude and choosing the recognized object on some other basis, for example preference due to mere exposure (Zajonc, 1968).

In our studies, the two possible strategies of use of the recognition heuristic as a means of inferring relative magnitude and simple choice of the recognized item regardless of the question were examined in Experiment 2. We found that some participants do indeed choose the recognized item regardless of the framing of the question, a strategy indistinguishable from the recognition heuristic in standard formulations of the problem. However, the number of participants who use this strategy is small and although it might exaggerate the effect ascribed to recognition heuristic elsewhere, it cannot account for it.

The recognition heuristic was demonstrated in our Experiment 1 using similar materials to Oppenheimer (2003) and a similar procedure to that of Goldstein & Gigerenzer (2002). The majority of our participants did use the recognition heuristic as a "fast and frugal" means of decision-making when placed under time pressure, a constraint that was absent in the Oppenheimer (2003) study. However, the heuristic is not automatically applied as the number of participants showing it was reduced in our Experiment 2. This was despite the fact that the choice to be made was identical and the heuristic would therefore have equivalent adaptive value in both situations. The recognition heuristic is the single simplest heuristic in the adaptive toolbox and makes up the first principle in more complex decision-making algorithms such a take-the-best (Gigerenzer & Goldstein, 1996). Establishing the situations when the recognition heuristic is employed is a necessary prerequisite for evaluating the applicability of the fast and frugal tools within the adaptive toolbox. There have been very few experiments on this. The current study goes some way towards addressing this issue. We suggest that our results also throw up some interesting avenues for future research. For example, a future study could vary the degree of time, or other, pressure on participants and examine the effects of this on the frequency with which a recognition strategy is used.

Acknowledgments

Thanks to Alexandra Marshall for collecting and scoring the data for Experiments 1 and 2.

References

Borges, B., Goldstein, D. G., Ortmann, A., & Gigerenzer, G. (1999). Can ignorance beat the stock market? In: G. Gigerenzer, & P. M. Todd, (Ed.s). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world.* Oxford: Oxford University Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650-669.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In: G. Gigerenzer, & P. M. Todd, (Ed.s). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109,* 75-90.

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): rethinking the recognition heuristic. *Cognition, 90,* B1-B9.

Rieskamp, J. & Hoffrage, U. (1999). Why do people use simple heuristics and how can we tell? In: G. Gigerenzer, & P. M. Todd, (Ed.s). *Simple heuristics that make us smart.* Oxford: Oxford University Press.

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review, 63,* 129-138.

Zajonc, R.B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1-27.

Appendix

Towns and Cities used in Experiments 1 & 2

Fictional
Papayito
Al Ahbahib
Las Besas
Weingshe
Rio del Sol
Heingjing
Rhavadran
Gohaiza
Schretzberg
Svatlanov

Real
Norwich
Ipswich
Preston
Wigan
Sunderland
Crewe
Coventry
Gillingham
Sheffield
Burnley

Filler
Limerick
Toledo
Berkley
Haifa
Stoke
Rotherham
Bradford
Derby

Don't teach me 2 + 2 equals 4: Knowledge of arithmetic operations hinders equation learning

Nicole M. McNeil (nmmcneil@wisc.edu)
University of Wisconsin-Madison
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

This study investigated whether children's knowledge of arithmetic operations hinders their ability to solve novel equations after instruction. Second- and third-grade children completed a timed arithmetic pretest as a means for assessing their proficiency with arithmetic operations. Next, they received lessons on the principle of mathematical equivalence either in a context designed to activate their knowledge of arithmetic operations (e.g., 15 + 13 = 28), or in a context designed to not activate their knowledge of arithmetic operations (e.g., 28 = 28). Then, children completed an equation-solving posttest (e.g., 3 + 9 + 5 = 6 + __). After the posttest, children switched lesson contexts and completed the posttest again. Children solved more equations incorrectly after receiving lessons in the operational context. Additionally, the operational context led children who were most proficient with arithmetic operations to solve more equations using the typical addition strategy of adding up all the numbers. Results highlight that the activation of existing knowledge can interfere with the acquisition of new information.

Some domains of knowledge are particularly difficult for people to learn, even after significant amounts of training or instruction. There are many examples of this in our formal education system, including reading, mathematics, science, and foreign language. Over the past several years, a number of scientists (e.g., Flege, Yeni Komshian, & Liu, 1999; Kuhl, 2000; McNeil & Alibali, 2002; Schauble, 1990; Zevin & Seidenberg, 2002) have begun to consider how existing knowledge may contribute to these difficulties. The general theoretical view is that later learning is strongly constrained by early learning (cf. Tolman, 1948). If this is true, it obviously has implications in domains, like second language learning, where people learn one thing for many years (native language) before switching gears and learning something new, but closely related (second language).

The domain of mathematics is another domain in which people learn one topic for many years before switching gears and learning a new, but closely related, topic. Specifically, in most American mathematics classrooms, children learn arithmetic operations for many years (i.e., grades K-6) before eventually reaching algebra and being introduced formally to equations and the principle of mathematical equivalence. Mathematical equivalence is the principle that the two sides of an equation represent the same quantity.

Elementary school children (ages 7-11) have significant difficulties with equations and the principle of mathematical equivalence (Carpenter & Levi, 2000; Kieran, 1981; Baroody & Ginsburg, 1983). Their difficulties are most apparent when they are presented with equations that have operations on both sides of the equal sign (e.g., 3 + 4 + 5 = 3 + __). In the absence of instruction, approximately 80% of second- through fifth-grade children solve these types of equations incorrectly (Alibali & Goldin-Meadow, 1993; Alibali, 1999; McNeil & Alibali, 2000; NCISLA, 2000; Perry, Church, & Goldin-Meadow, 1988; Rittle-Johnson & Alibali, 1999).

Although there are many possible accounts of children's difficulties, including immature working memory function (Adams & Hitch, 1997; Gathercole & Pickering, 2000) or insufficient knowledge of necessary prerequisite skills (Haverty, 1999), the *change-resistance account* suggests that children's equation-learning difficulties are due, at least in part, to children's existing knowledge (McNeil, 2004). More specifically, the account posits that children construct knowledge on the basis of their early experiences with arithmetic operations and that this knowledge contributes to children's difficulties with more complex equations.

There are at least three knowledge structures that children learn from their early experiences with arithmetic operations that may ultimately hinder the ability to learn complex equations (see McNeil & Alibali, 2002). First, children may learn an operational strategy for solving math problems—perform all the given operations on all the given numbers. For example, in a typical addition problem like 3 + 4 + 5 + 3 = __, a problem solver simply needs to add up all the numbers and put the total in the blank. Second, children may learn an operational perceptual pattern related to the structure of math problems—the traditional "operations = answer" problem structure. For example, in the typical addition problem above, all of the numbers and operations are on the left hand side of the equation, and the answer blank is on the right side of the equation (directly following the equal sign). Third, children may learn an operational concept of the equal sign—the equal sign means "the total." Although these three operational patterns facilitate fast and accurate

performance on typical addition problems, they do not map onto more complex equations. For example, when presented with the equation "3 + 4 + 5 = 3 + _", a problem solver cannot just add up all the numbers. He or she cannot assume that the equation will conform to the traditional "operations = answer" problem structure. And, he or she needs to understand that the equal sign denotes an equivalence relationship between the two sides of the equation in order to generate a correct solution.

According to the change-resistance account, children learn these operational patterns from their experience with arithmetic operations. They store these operational patterns in memory. Then, when they are presented with a novel equation, their representations of the operational patterns are activated. Once activated, the representations guide attention and can hinder the ability to encode and interpret novel equations that do not directly map onto the patterns (cf. Bruner, 1957; Luchins, 1942; Knoblich, Ohlsson, & Raney, 2001).

In accordance with the change-resistance account, studies have shown that children do, indeed, rely on their knowledge of arithmetic operations when presented with complex equations. For example, when asked to solve the equation "3 + 4 + 5 = 3 + __", most students use their knowledge of the "perform all given operations on all given numbers" strategy and just add up all the numbers and put 15 in the blank (McNeil & Alibali, 2000, 2002, in press b). When asked to reconstruct the equation "3 + 4 + 5 = 3 + __" after viewing it briefly, many students use their knowledge of the traditional "operations = answer" problem structure and write "3 + 4 + 5 + 3 = __" (McNeil & Alibali, 2002, in press b). When asked to define the equal sign, many students use their knowledge of operational symbols (e.g., +) and say that it means, "the total" (McNeil & Alibali, in press a). Thus, children rely on their knowledge of the operational patterns when presented with complex equations.

McNeil and Alibali (2002) provided additional evidence for the change-resistance account by showing that children's reliance on the operational patterns can hinder the ability to learn about equations. In the study, they documented a significant negative linear relationship between children's reliance on the operational patterns on a pretest and the generation of correct equation-solving strategies after a brief lesson on equations. Children who were most reliant on the operational patterns at pretest were the least likely to generate correct equation-solving strategies following a lesson, and children who did not rely on the operational patterns at pretest were the most likely to generate correct equation-solving strategies following a lesson.

Although the results of McNeil and Alibali (2002) support the change-resistance account, they provide only correlational evidence about the relationship between children's knowledge of arithmetic operations and equation-learning difficulties. The change-resistance account argues that the activation of children's knowledge of arithmetic operations *causes* equation-learning difficulties. Thus, in the present study, the activation of children's knowledge of arithmetic operations is manipulated. Children are given lessons about the principle of math equivalence either in a context designed to activate their knowledge of arithmetic operations (operational context), or in a context designed to *not* activate their knowledge of arithmetic operations (non-operational context). If the activation of knowledge of arithmetic operations contributes to difficulties with equations, then the operational lesson context should be inferior to the non-operational lesson context. That is, after receiving lessons in the operational context, children should solve more equations incorrectly, and they should solve more equations with the strategy that is the most often used in the absence of instruction (i.e., they should rely on their knowledge of the operational strategy and just add up all the numbers in the equations).

Additionally, if knowledge of arithmetic operations contributes to difficulties with equation learning, then children who are most proficient with arithmetic operations should be least likely to benefit from the lessons. This, of course, is assuming that children who are most proficient with arithmetic operations have the strongest representations of the operational patterns. Children who are proficient with arithmetic operations should solve more equations incorrectly, and they should solve more equations by just adding up all the numbers in the equations.

Continuing this rationale, the combination of the operational lesson context and proficiency with arithmetic operations should be a "double whammy." That is, proficient children who have just received lessons in the operational context should solve more equations incorrectly, and they should solve more equations by just adding up all the numbers in the equations.

Method

Participants

Ninety-three second- and third-grade children from a public elementary school in Youngsville, North Carolina participated. Eleven children were excluded from the analysis because they were absent on one or more days of the study. Two additional children were excluded because their performance on the equations was three standard deviations away from the mean. The final sample contained eighty children (38 boys and 42 girls).

Measures

Timed Arithmetic Pretest The timed arithmetic pretest was used to assess children's proficiency with arithmetic operations. Participants were given 30 seconds to solve as many arithmetic problems (out of 20) as possible. The problems involved only addition and subtraction (no multiplication or division).

Equation-solving Posttest Participants were given unlimited time to solve twelve equations with operations on both sides of the equal sign (e.g., 5 + 4 + 7 = 5 + __, 6 + 4 + 8 = __ + 3).

Procedure

The study was conducted over a two-week period in children's regular mathematics classrooms. Children's mathematics teachers collected the measures and administered the lessons in the classroom setting. Children first completed the timed arithmetic pretest. Then, teachers taught a set of lessons about the principle of mathematical equivalence. The following is an excerpt from the spoken lesson script: "The correct answer is 28! That's because whatever is on one side of the equal sign has to be **the same amount as** [teachers were told to stress words in bold] whatever is on the other side of the equal sign."

Because the lessons were scripted, all children received the same *spoken* lessons. Children were randomly assigned to lesson contexts through the use of individual booklets. The booklets enabled children to follow along with the spoken lessons. Children received lessons in one of two contexts. In the operational context, booklets contained problems designed to activate children's knowledge of arithmetic operations (e.g., 15 + 13 = 28). In the non-operational context, booklets contained problems designed to *not* activate children's knowledge of arithmetic operations (e.g., 28 = 28). Children received two days of lessons (approximately 15 minutes per day) before they completed the first equation-solving posttest. After the first posttest, children received lessons in the other context (e.g., children who had already received lessons in the operational context now received lessons in the non-operational context). Finally, children once again completed the equation-solving posttest.

Coding

Proficiency with Arithmetic Operations Children were categorized as proficient on the timed arithmetic test if they both solved three (median) or more arithmetic problems correctly, *and* solved one (median) or fewer arithmetic problems incorrectly. This coding system led to approximately equal numbers in the proficient (N = 37) and not proficient (N = 43) groups.

Equation-solving Performance Children's strategies were coded using a system developed by Perry, Church and Goldin-Meadow (1988). Strategies were assigned based on the solutions that children wrote in the answer blank. Examples are presented in Table 1. Solutions were coded as reflecting a particular strategy as long as they were within ±1 of the solution that would be achieved with that particular strategy. Again, we were especially interested in children's use of the add-all strategy (see Table 1) because it is the most commonly used strategy in the absence of instruction.

Table 1: Example solutions and corresponding strategy codes for the given equation.

5 + 4 + 7 = 5 + __	
Solution	Strategy
11	Correct
21	Add all
16	Add to equal sign
4	Carry
1	Idiosyncratic

Results

Number of Incorrect Solutions

Overall, performance on the equation-solving posttest was abysmal. Children solved 11.27 (SD = 0.98) equations incorrectly (out of 12). We performed a 2 (proficiency with arithmetic operations: proficient or not proficient) x 2 (lesson context: operational context or non-operational context) ANOVA with repeated measures on lesson context and number incorrect on the equation-solving posttest (out of 12) as the dependent measure. As expected, the analysis revealed a significant main effect of lesson context, $F(1, 78) = 5.24$, $p = .025$. Children solved more equations incorrectly after receiving lessons in the operational context (M = 11.44, SD = 0.90) than after receiving lessons in the non-operational context (M = 11.11, SD = 1.03). Neither the main effect of proficiency nor the interaction of proficiency and lesson context was significant. Although, as mentioned, children's performance was very poor overall, so there was not a great deal of variability on the dependent measure to predict.

Number of Add-all Solutions

Consistent with prior work, the add-all strategy was the most popular strategy. On average, children solved 5.29 (SD = 3.67) equations (out of 12) by just adding up all the numbers in the equations. We performed a 2

(proficiency with arithmetic operations: proficient or not proficient) x 2 (lesson context: operational context or non-operational context) ANOVA with repeated measures on lesson context and number of equations solved with the add-all strategy (out of 12) as the dependent measure. As expected, the analysis revealed a significant interaction of proficiency and lesson context, $F(1, 78) = 4.90$, $p = .03$. As shown in Figure 1, the children who solved the most equations using the add-all strategy were the ones who were proficient with arithmetic operations and had just received lessons in the operational context ($M = 6.11$, $SD = 3.60$).

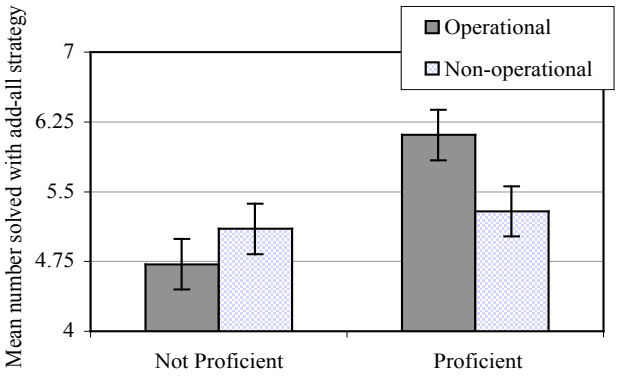


Figure 1: Mean number of equations solved with the add-all strategy (out of 12) as a function of arithmetic proficiency and lesson context. Error bars represent the pooled standard error.

Discussion

Consistent with the change-resistance account of children's equation-learning difficulties, results of the present study suggest that children's knowledge of arithmetic operations hinders their ability to learn about more complex equations. Children solved the fewest equations correctly after receiving lessons in contexts designed to activate their knowledge of arithmetic operations. Moreover, children who were most proficient with arithmetic operations and had just received lessons in the operational context solved the most problems by just adding up all the numbers, which is the most common strategy used by children who have not received any instruction at all.

Results suggest that it is vital to consider the state of children's existing knowledge when theorizing about children's learning difficulties. Thus, prevailing theories that focus on children's immature working memory system or their lack of prerequisite knowledge are missing a layer of complexity. More generally, results contribute to a growing body of work that suggests that knowledge can be detrimental to learning in some cases (Adelson, 1984; Flege et al., 1999; Kuhl, 2000; Schauble, 1990). Because knowledge typically facilitates learning, cases like these in which knowledge hinders learning can provide a unique window onto how the mind works (cf. Luchins, 1942).

Although the present study suggests that knowledge of arithmetic operations hinders equation learning. The results are not definitive. Performance on the equations with operations on both sides of the equal sign was abysmal, even after four, fifteen-minute classroom-based lessons. Thus, most children in the study had difficulties learning from the lessons on the principle of math equivalence. However, this is not surprising when viewed from the perspective of the change-resistance account. Children in the study are learning math on a day-to-day basis from a traditional, skills-based mathematics curricula. Thus, they are deeply entrenched in the operational patterns that are predicted to hinder learning. It is not that surprising that the brief lessons in the present study were not able to override this deeply entrenched way of thinking.

In terms of educational implications, results conflict with both intuition, and traditional mathematics practices. Intuition suggests that the children who are best at one topic in math should be best at another topic in math. And indeed, most schools assign, or track, children to algebra based on their performance in elementary school with arithmetic operations. This policy certainly makes sense if children need to be highly proficient with arithmetic operations before they are able to learn algebraic equations. However, it makes less sense if arithmetic proficiency hinders equation learning. Thus, our schools may be holding back children who would thrive in an early algebra course.

Equally important, American schools often implement spiral curricula in which old information is reintroduced year after year. The idea is that old information provides a framework within which new material can be introduced. In relation to mathematics instruction, this means that basic arithmetic operations are reintroduced year after year. Indeed, data from the Third International Mathematics and Science Study (Beaton et al., 1996) show that, unlike students from higher-achieving countries, students in American mathematics classrooms spend substantial amounts of class time practicing and reviewing basic arithmetic skills throughout the elementary and middle school years, when they should be concentrating on more advanced topics.

This type of spiral, review-based instruction has received some support from the scientific community. For example, Nathan et al. (2004) argue that the most effective instructions are the ones that "bridge" from children's existing knowledge to the new material. However, results of the present study suggest that teachers need to be careful about what they are trying to build bridges between. In some case, to-be-learned information does not map well onto existing knowledge, and in these cases, bridging might not be the most effective instructional strategy.

Instead of reintroducing basic arithmetic facts year after year, mathematics educators may wish to develop creative ways to integrate more algebraic ways of thinking into the math curricula as early as possible. One recommended strategy for integrating algebraic thinking into the earlier grades is to focus on equality and the equal sign (e.g., Carpenter, Franke, & Levi, 2003). For example, instead of simply reviewing and practicing basic arithmetic "facts" such as "3 + 4 = 7" year after year, young students can learn "3 + 4 = 7," "7 = 3 + 4," "3 + 4 = 5 + 2," and "7 = 7." Instructional strategies such as this may prevent the entrenchment of operational patterns and facilitate the notoriously difficult transition from arithmetic to algebra.

Acknowledgments

This research was supported by a Research Award from the University of Wisconsin Department of Psychology to N. M. McNeil. I thank members of the Cognitive Development Research Group at the University of Wisconsin for helpful discussions about the study and Jerry Haeffel for comments on a previous version of this paper. I also thank the students, teachers, and administrators at Youngsville Elementary School in North Carolina. Special thanks go to third-grade teacher Heather Shipley for her organization and enthusiasm.

References

Adams, J. W., & Hitch, G. J. (1997). Working memory and children's mental addition. *Journal of Experimental Child Psychology, 67*, 21-38.

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 483-495.

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35*, 127-145.

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468-523.

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the "equals" sign. *Elementary School Journal, 84*, 199-212.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 2*, 123-152.

Carpenter, T. P., & Levi, L. (2000). *Developing conceptions of algebraic reasoning in the primary grades*. Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school*. Portsmouth, NH: Heinemann.

Flege, J. E., Yeni Komshian, G. H., & Liu, F. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41*, 78-104.

Gathercole, S. E., & Pickering, S. J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years. *British Journal of Educational Psychology, 70*, 177-194.

Haverty, L. A. (1999). *The importance of basic number knowledge to advanced mathematical problem solving.* Unpublished Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA.

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12*, 317-326.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory and Cognition, 29*, 1000-1009.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Science, 97*, 11850-11857.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs, 54 (6, Whole No. 248)*.

McNeil, N. M., & Alibali, M. W. (2000). Learning mathematics from procedural instruction: Externally imposed goals influence what is learned. *Journal of Educational Psychology, 92*, 734-744.

McNeil, N. M., & Alibali, M. W. (2002). A strong schema can interfere with learning: The case of children's typical addition schema. In C. D. Schunn & W. Gray (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

McNeil, N. M., & Alibali, M. W. (in press a). Knowledge change as a function of mathematics experience: All contexts are not created equal. *Journal of Cognition and Development*.

McNeil, N. M., & Alibali, M. W. (in press b). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science*.

Nathan, M. J., Masarik, D. K., Stephens, A. C., Alibali, M. W., & Koedinger, K. R. (2004). Enhancing middle-school students' representational fluency: A classroom study. *Manuscript under review*.

National Center for Improving Student Learning and Achievement in Mathematics Education (2000). *Building a foundation for learning algebra in the elementary grades*. Madison, WI: Author.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359-400.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*, 175-189.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31-57.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*, 189-208.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in reading and other tasks. *Journal of Memory and Language, 47*, 1-29.

Processing Ambiguous Words: Are Blends Necessary for Lexical Decision?

David A. Medler (dmedler@mcw.edu)
Language Imaging Laboratory
Department of Neurology, Medical College of Wisconsin
Milwaukee, WI

C. Darren Piercey (piercey@unb.ca)
Department of Psychology, University of New Brunswick
Fredericton, NB

Abstract

A previous computational model (Joordens & Besner, 1994) has suggested that during lexical access, ambiguous words tend toward a blend state; that is, network activations settle into an incorrect state that is a mixture of the multiple representations of the ambiguous item. It has been suggested that this blend state actually aids lexical decision (LD) for ambiguous items as the blend state creates a larger "feeling of familiarity" which lexical decision may exploit. This theory, however, is based on the results of a computational model (a simple Hopfield network) in which multiple representations cannot be learned. Here we use a Symmetric Diffusion Network (SDN) to effectively learn and retrieve multiple mappings for a single input (i.e., ambiguous items). The model consists of three main processing regions–orthographics, phonology, and semantics–and is trained on a corpus of unambiguous items and ambiguous items that range in their degree of balance (probability distribution) between the multiple meanings. Following training, the SDN is able to reproduce the correct probability distributions for the ambiguous items; that is, it does not produce blend states. Furthermore, the model qualitatively captures the processing advantage for ambiguous items. Consequently, the notion of a blend state being used for LD is re-evaluated, and further assumptions about semantic processing are explored.

Introduction

From a computational perspective, we can break basic language processing into three main components: the semantic representation (what a word means), a phonological representation (the sound of a word), and an orthographic representation (the written form of a word) (e.g., Seidenberg & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996; Harm & Seidenberg, 1999). The relationship between the phonological and semantic representations is initially established in early childhood, and then the mapping between the orthographic representation and the phonological representation (spelling-to-sound conversion) along with the mapping between the orthographic representation and the semantic representation (spelling-to-meaning conversion) is learned later in life (e.g.,Harm & Seidenberg, *in press*).

Ideally, there would be a one-to-one mapping between any of the representations, such that one spelling would correspond to one pronunciation, which would correspond to one meaning. Unfortunately, one-to-one mappings are far from the norm in English. That is, words that sound the same (homophones) can have different semantic representations (/flaI/: fly [insect]; fly [zipper]), different orthographic representations (/laIt/; light [fewer calories]; lite [fewer calories]), or different semantic and orthographic representations (/beɚ/: bear[furry animal]; bare[naked]). Similarly, words that are spelled the same (have the same orthographic representation) can have different phonological representations (either: /ˈaI.Dɚ/ [one or the other], /ˈIːDɚ/ [one or the other]), or phonological and semantic representations (wind: /waInd/ [twist]; /wInd/ [moving air]).

In fact, many words in English have polysemous or ambiguous semantics. For example, WordNet® (Fellbaum, 1998) lists a total of 146,350 noun, verbs, adjectives, and adverbs. Table 1 shows the percentage of unique and ambiguous words, as well as sense data. Whereas ambiguity is often defined as a word having multiple meanings across semantic categories or word classes, a word's sense is defined as it's meaning within a semantic category and can vary dramatically from the prior definition of ambiguity. For example, although Borowsky & Masson (1996) consider "deep" to be an unambiguous word, WordNet® lists "deep" with 3 noun senses, 15 adjective senses, and 3 adverb senses. It is clear that ambiguity is prevalent in English, and there is evidence that it has an effect on how we process words.

Table 1. Percentage of words having unique, ambiguous, and multisense meanings.

Word Class	Unique	Ambiguous	Senses
Noun	86.7	13.3	29.7
Verb	53.4	46.6	75.4
Adjective	74.5	25.5	48.7
Adverb	82.9	17.1	33.1

For example, the behavioral data from word ambiguity studies produces a paradox. In a lexical decision (LD) paradigm, ambiguity aids in word identification; that is, ambiguous words are identified as words more quickly and more accurately than unambiguous words (Gernsbacher,

1984; Borowsky & Masson, 1996). In contrast, in connected text studies (Rayner & Duffy, 1986; Rayner & Duffy, 1987; Duffy, Morris, & Rayner, 1988), ambiguous words are processed more slowly than unambiguous words. In other words, when semantic decisions (SD) (decisions on word meaning) are required, words with multiple meanings pose more difficulty than words with single meanings. This ambiguity paradox was illustrated in a single experiment in which participants first made a lexical decision, and then had to make a relatedness judgement on a subsequently presented word (Piercey & Joordens, 2000). In this study, participants showed an ambiguity advantage for lexical decision, and an ambiguity disadvantage on the subsequent relatedness decision. The importance of the ambiguity paradox lies in the fact that it leads directly to the question of how words are represented in the brain, and how we get access to these words. Any model of language will have to account for the ambiguity paradox if it is to be successful.

Previous models of the ambiguity advantage in LD, however, have shown mixed results. For example, Joordens & Besner (1994) trained a two layer Hopfield network consisting of 125 binary nodes (75 perceptual nodes and 50 conceptual nodes; activations of either +1 or -1). The perceptual nodes represented perceptual features and were never updated during retrieval (that is, they were clamped to a specific pattern). The conceptual nodes represented semantics, and the network was effectively fully connected. Learning was via a Hebbian learning algorithm.

They had two criteria for deciding if a PDP model could successfully account for the ambiguity effect; (a) the network had to retrieve one of the semantic patterns associated with the ambiguous words, and (b) the network had to retrieve ambiguous words faster than unambiguous words. Joordens and Besner (1994) were able to produce an ambiguity advantage within the *conceptual* nodes of their network when it into a stable pattern. This only occurred, however, when the network was relatively small and when the ambiguous meanings had equal probability. Most of the time (over 50% of the trials), their networks failed to settle into a correct pattern and formed a "blend" of the two learned meanings of the words over the conceptual units. Their initial conclusion from these simulations was that distributed models trained with Hebbian learning rule may not be suitable for capturing ambiguity effect.

In a different computational model, Kawamoto, Farrar & Kello (1994) trained a recurrent neural network with the Least Mean Square learning algorithm. Their model contained both "spelling" nodes and "meaning" nodes using a distributed representational coding scheme. During recall, the "spelling" nodes were given environmental activation, and the network was allowed to settle into a stable state. They found that they could produce an ambiguity advantage within the units representing "spelling", but showed the opposite effect in units representing "meaning" (an ambiguity disadvantage in semantics?). It has been suggested, however, that Kawamoto et al.'s (1994) network also settled into blend states in the meaning units (Kello, 2003, *Personal Communication*).

Although both of these models produced an ambiguity advantage (albeit in different processing regions), the networks failed to differentiate between the ambiguous items and produced blended representations. However, in later commentaries (Masson & Borowsky, 1995; Rueckl, 1995; Besner & Joordens, 1995), it was concluded that it may be possible for lexical decisions to be made prior to the network settling into these blend states. In other words, correct lexical decisions could be based on the "blend" states for ambiguous words resulting in a greater feeling of familiarity which could then be used to produce LD.

Using the model of Joordens and Besner (1994) as a basis for their theory, Piercey and Joordens (2000) developed the "efficient then inefficient" hypothesis for the processing of ambiguous words. They concluded that a lexical decision is made based on early processing and that a blend state (i.e., when all meanings of a word are simultaneously activated) produces an advantage for lexical decision but a disadvantage for the relatedness decision. That is, lexical decisions are made based on a feeling of familiarity that occurs during the early stages of processing, before a complete representation of the current item forms (i.e., efficient processing). Therefore, these decisions could be made regardless of an eventual blend state. However, when the participants need to determine which meaning of the word is appropriate to a particular context, processing slows down. The participant continues to process the ambiguous word and each of the word's meanings compete with each other. That is, the participant needs to leave the blend state and choose a meaning for the item so that further semantic processing can occur. This disambiguation of the blend state is an inefficient process that unambiguous words do not share. It should be noted that this theory is based specifically on the fact that the model of Joordens and Besner (1994) produced blended states for ambiguous words.

In this paper, we readdress the ambiguity advantage for lexical decision using a computational model that is able to learn multiple mappings for a single input. These models do not produce blend states; therefore, if the ambiguity advantage can be reproduced, then the notion of blend states existing should be questioned.

Symmetric Diffusion Networks

Symmetric Diffusion Networks (SDNs) are a class of computational models based upon the principles of continuous, stochastic, adaptive, and interactive processing (Movellan & McClelland, 1993). From a computational perspective, SDNs can be viewed as a continuous version of the Boltzmann machine; that is, time is intrinsic to the dynamics of the network. Furthermore, SDNs embody Bayesian principles in that they develop internal representations based upon the statistics of the environment. One of the main advantages of SDNs is that they are able to

learn multiple mappings for a single concept, something previous models often have difficulties with. In other words, SDNs are able to learn ambiguous mappings.

Recent work (Medler & McClelland, 2001) has shown that when biologically inspired constraints (i.e, activations within the range [0,1], positive between layer projections, lateral inhibition) are applied to SDN's, their effective performance is increased substantially in terms of the number of patterns they can be trained on, the rate at which patterns are learned, and their ability to separate out independent sources in an unsupervised manner.

Network Dynamics and Learning

Network dynamics are based upon continuous activations that develop over time, and are governed by the following equation:

$$\Delta a_i(t) = \Delta t[net_i(t) - n\hat{e}t_i(t)] + \sigma \cdot \sqrt{\Delta t} Z_i(t) \quad \textbf{Eq. } 1$$

where,
is the summed activation of all the activities coming into the unit–including its bias–passed through a squashing function

$$net_i = h(\sum_{j=1}^{n} a_j w_{ij})$$

such as the logistic, $h(u)$ = 1-exp(-u), and

$$n\hat{e}t_i = 1/g_i \cdot f(a_i)$$
$$= 1/g_i \cdot \log[(a_i - \min)/(\max - a_i)]$$

represents the net input required to maintain an activation value of a_i. Here we use the inverse logistic, where min and max are the minimum and maximum activation bounds respectively. g_i is a gain function, and $Z_i(t)$ is the standard Gaussian variable with zero mean and unit variance. The last term in the equation adds stochasticity to the network, which allows it to learn multiple meanings for a single input.

SDNs are trained with the *Contrastive Hebbian Learning* (CHL) algorithm, which performs both supervised and unsupervised learning depending on the environmental inputs to the network. Basically, learning occurs by presenting a pattern to the network and letting it settle for a set number of cycles. During this positive phase, co-occurrence statistics are computed for all the units. A negative phase then follows where the pattern is removed, the network is allowed to re-settle, and co-occurrence statistics are collected once again. Weights are then adjusted using the difference between the negative phase statistics and the positive phase statistics.

$$\Delta w_{ij} = \varepsilon\left[\sum\left(a_i^+ a_j^+\right) - \sum\left(a_i^- a_j^-\right)\right] \quad \textbf{Eq. } 2$$

In essence, the CHL algorithm makes weight adjustments based upon subtracting out the statistics of the base activity of the network (negative phase) from the statistics of the environment plus base activity (positive phase). Weight adjustments in this model were computed after each pattern presentation, as opposed to batch learning which adjusts weights only after all patterns have been presented (Movellan & McClelland, 1993).

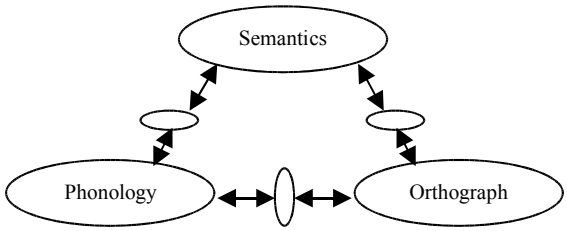


Figure 1. Network architecture showing the three main processing layers and connecting hidden layers.

Network Architecture, Stimuli, & Training

In keeping with previous models of language (e.g., Harm & Seidenberg, 1999), the network consisted of three main processing layers: an "orthographic", a "phonological", and a "semantic processing" layer. To capture the gross relationship between semantics and the orthographic and phonology representation of words, there were twice as many units (10) in the semantic layer as in the orthographic and the phonology layers (5 units each). Each layer was connected to the other via a set of hidden layers (5 units). Between layer connections were excitatory, while within layer connections were inhibitory (Medler & McClelland, 2001).

Stimuli were arbitrary, distributed binary patterns [0,1] that encoded the orthography, phonology, and semantics of a given "word". It is recognized that the abstract, distributed codes used in this simulation are not true representations of semantics, phonology, and orthography; however, future simulations using the same architecture will use more systematic encodings for these representations. Half of the training patterns (20) were unambiguous words, and half (20) were ambiguous words. In this model, only semantics had ambiguous patterns (as opposed to ambiguous orthography or phonology). Hence, ambiguous words had two possible meanings, and were selected with either a 70/30 distribution or a 50/50 distribution. Two representational training patterns are shown in Table 2; the presentation probability is the likelihood of that specific pattern being selected during the positive phase. Nonwords were simply random patterns across the orthographic and phonology units that had not been previously trained[1].

During training, the orthography and phonology units were clamped on, and the semantic and hidden units were modified during the positive and negative phases. Following training, the network was able to correctly produce the probability structure of the training stimuli. That is, the network was able to successfully recall the semantic patterns with the same probabilities that it was trained on. The

[1] As previous results have suggested that non-word foils need to be word-like for the ambiguity advantage to be stable (Borowsky & Masson, 1996), and we are assessing LD over the semantic units, we clamped both the orthographic and phonology units for the non-words.

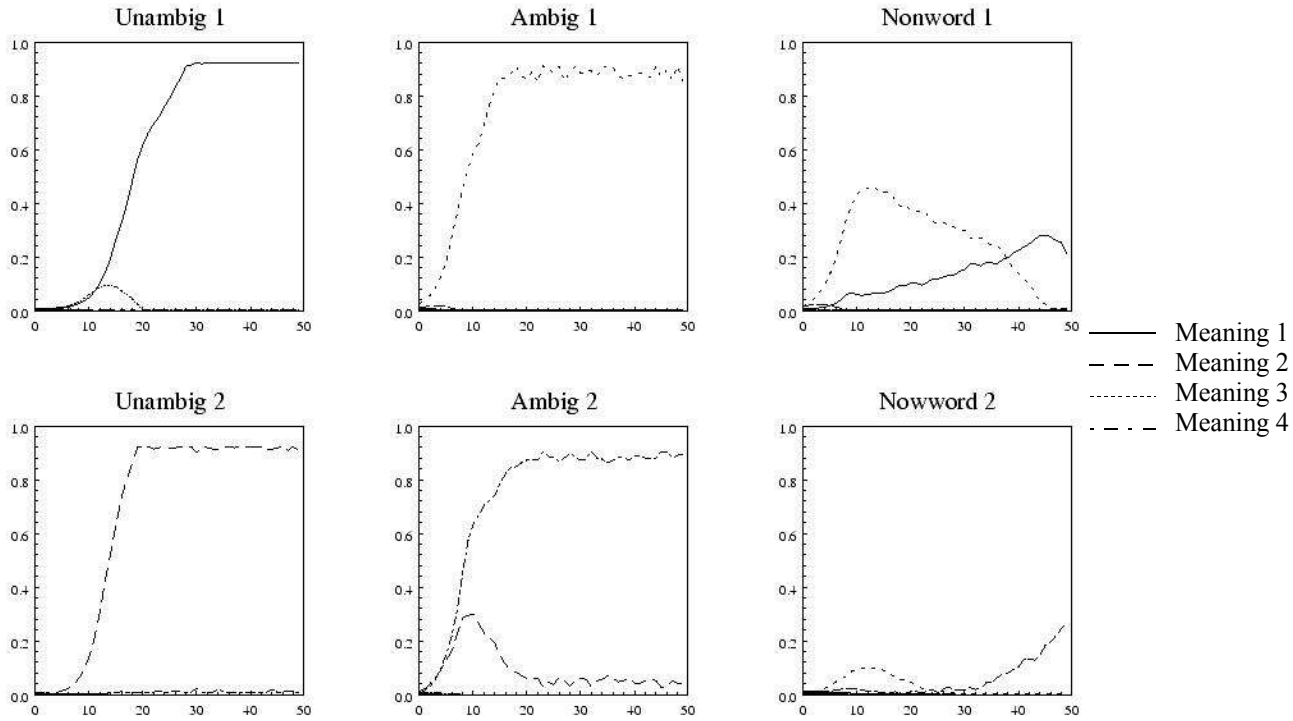


Figure 2. Sample differentiation scores for a subset of unambiguous, ambiguous, and non-words. Note that the first non-word is mistaken for a word at a criterion of 0.25.

network did not produce blend states for the ambiguous items.

Table 2. Sample Patterns Showing Positive and Negative Training Phases for Unambiguous and Ambiguous Words

Present.	Unambiguous		
Prob.	Orthography	Phonology	Semantics
+1.0	1 0 0 1 1	0 1 1 0 1	1 0 1 0 1 1 1 0 1 0
-1.0	1 0 0 1 1	0 1 1 0 1	* * * * * * * * * *
	Ambiguous		
	Orthography	Phonology	Semantics
+0.7	0 1 1 1 0	1 1 0 0 0	0 1 1 0 0 1 1 0 0 1
+0.3	0 1 1 1 0	1 1 0 0 0	1 0 1 1 0 0 0 0 1 0
-1.0	0 1 1 1 0	1 1 0 0 0	* * * * * * * * * *

During testing, the orthography and phonology units were clamped on, and the semantic units were allowed to settle. Previous models waited for the networks to settle into a stable state, and took this measure as a reaction time. In our model, we assume a speeded decision based on a differentiation measure (McClelland & Chappell, 1998) computed over the known words, *k*:

$$diff_k = \prod_{i=1}^{n} |1 - \mathrm{P}(T_i) - \mathrm{P}(G_i)|$$

where G_i is the generated pattern, and T_i is a target. If a generated pattern does not match a target pattern, then the differentiated score should approach zero. A matched pattern, on the other hand, should produce a score that approaches one. If multiple patterns are partially activated (i.e., a blend), then several words should show a differentiation score that approaches a middle value.

When $diff_k$ exceeds a threshold (in this case, an arbitrary value of 0.25), a decision of "word" is made. If a word (or non-word) fails to reach the threshold within a certain time limit (an arbitrary point such as 20 time steps plus or minus some random time to introduce stochasticity in the response times), then a nonword decision is made. This nonword time limit can be adjusted to reflect task instructions (e.g., "respond as quickly as possible" vs. "respond as quickly and accurately as possible"). Consequently, we can produce both accuracy and reaction time measurements from our model.

Results

Figure 2 shows some representative differentiation scores for a sub-sample of the testing stimuli. As can be seen, no blends were formed (a single score tended towards one, whereas all other scores tended towards zero). Furthermore, the figure shows how using a threshold criterion of 0.25 for a speeded decision leads to the first non-word being misclassified as a word. Finally, it should also be noted that although the second ambiguous word looks like it is initially activating two word meanings (heading towards a blend

perhaps?), the second word meaning (i.e., the dashed line) is actually associated with the second unambiguous word.

In terms of reaction times, the network showed an ambiguity advantage. The network made a lexical decision for unambiguous items in an average of 13.0 time steps, whereas ambiguous items took 11.5 time steps. In contrast to previous empirical work (Piercey & Joordens, 2000), however, there was not a clear advantage of ambiguous items in terms of accuracy. Lexical decision for unambiguous items was approximately 97% correct, while ambiguous items were only 95% correct within the speeded decision.

One last note to make is that the final differentiation score for ambiguous items was often lower and more variable than for unambiguous words. The differentiation score averaged over the last ten time steps for the unambiguous items was 0.92 (var = 3.7×10^{-4}) whereas ambiguous items had an average differentiation score of 0.87 (var = 5.1×10^{-4}). This suggests that, for ambiguous items, the final settled state for the networks was more unstable than unambiguous items, and that if speeded decisions were not made, then the ambiguity advantage in reaction times may disappear.

Discussion

We have shown how a network trained with the *CHL* produces the ambiguity advantage over the semantic nodes based on speeded decision. Furthermore, the model was able to produce the approximate correct probability distributions of the training corpus, thereby avoiding "blend" states. Consequently, the theory of blend states having to exist to aid in LD for ambiguous items may have to be re-evaluated. Furthermore, the efficient-then-inefficient hypothesis of Piercy and Joordens (2000) may have to be recast.

The results from this network stimulation suggest an alternative theory as to why ambiguous items show an advantage for lexical decision. Given that there are multiple distinct attractor states in semantics for ambiguous items, and given a random start state, then the probability of starting near an attractor is greater for ambiguous items than unambiguous items. Consequently, if a decision is based on traveling toward an attractor basin, then ambiguous items should—on average—reach a basin sooner than unambiguous items. This is similar to the attractor basin theory proposed by Plaut and Booth (2000). Consequently, lexical decisions are efficient for ambiguous items because they have a higher probability of starting near an attractor basin.

Note that this theory would require lexical decisions to be made at the semantic level. That is, if LD could be completed at the orthographic level or at the phonological level (say by having non-words that either violated the orthographic rules or the phonological rules of English), then the ambiguity advantage would disappear (cf., Borowsky & Masson, 1996).

Interestingly, this theory would also explain the disadvantage seen for ambiguous items during semantic decisions. If we assume that the network has settled into a stable state following the lexical decision (processing is automatic and continues even after the decision process), then both the unambiguous and ambiguous items will have activated a meaning in semantics. For unambiguous items, the semantic comparison would be relatively easy as there would only be one meaning to compare. For ambiguous items, however, the comparison becomes more unsettling. On some trials, the semantic decision would be relatively quick[2] as the network would be in the correct semantic attractor. On other trials, however, the network would be in an incorrect attractor, and would have to switch attractor states. Therefore, when trials are averaged, ambiguous items should show a disadvantage for semantic decisions. Consequently, semantic decisions are inefficient for ambiguous items because of the need to visit multiple attractor basins. Hence, this theory predicts that if we prime an ambiguous item towards one meaning or another, then the disadvantage should be lessened. Indeed, preliminary behavioral results show this to be the case (Piercey, Medler, & Hebert, 2003).

One area of potential criticism for the current model is that although it showed an ambiguity advantage for reaction times, it did not show an ambiguity advantage for accuracy. This discrepancy may be due to the choice of the differentiation score to evaluate network performance. This scoring mechanism assumes that the currently presented pattern is simultaneously compared to all learned words (thus, assuming that the learned patterns are stored somewhere exterior to the current model). Consequently, as unambiguous and ambiguous words are learned to criteria in the model, a decision based on the learned representations should show equal performance (where failure to recognize a word is based on a combination of the threshold criterion and the nonword decision time limit). One possible solution to this would be to use a different type of LD process, such as the harmony/referent model (Piercey, 2002; Joordens, Piercey, & Azarbehi, 2003) of lexical decision.

Future models will focus on training all processing levels (orthographic, phonological, and semantic) to address the theory of non-word background driving the ambiguity advantage in LD. As well, we will explicitly address the semantic relatedness decision issue to evaluate the theory predicted by the current simulations.

Reference List

Besner, D., & Joordens, S. (1995). Wrestling with ambiguity--further reflections: Reply to Masson and Borowsky (1995) and Rueckl (1995). *Journal of Experimental Psychology: Learning Memory and Cognition, 21,* 515-519.

[2] It is unclear whether the decision would be as fast as the unambiguous trials, as simulation results show the final state for the ambiguous words to be less robust and more variable. Consequently, one might expect this variability to slightly slow decision processes.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition, 22,* 63-85.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory & Language, 27,* 429-446.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113,* 256-281.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106,* 491-528.

Harm, M. W. & Seidenberg, M. S. Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review,* (in press).

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning Memory and Cognition, 20,* 1051-1062.

Joordens, S., Piercey, C. D., & Azarbehi, R. (2003). From word recognition to lexical decision: A random walk along the road of harmony. In *International Conference on Cognitive Modeling*.

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception & Performance, 20,* 1233-1247.

Masson, M. E. J., & Borowsky, R. (1995). Unsettling questions about semantic ambiguity in connectionist models: Comment on Joordens and Besner (1994). *Journal of Experimental Psychology: Learning Memory and Cognition, 21,* 509-514.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724-760.

Medler, D. A. & McClelland, J. L. (2001). Improving the performance of Symmetric Diffusion Networks via biologically inspired constraints. In K. Marko & P. Werbos (Eds.), *IJCNN'01: Proceedings of the INNS-IEEE International Joint Conference on Neural Networks* (pp. 400-405). Washington, DC: IEEE Press.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with Symmetric Diffusion Networks. *Cognitive Science, 17,* 463-496.

Piercey, C. D. (2002). *The Referent Model of Lexical Decision.* Unpublished doctoral dissertation, University of Alberta, Edmonton, Alberta, Canada,

Piercey, C. D., & Joordens, S. (2000). Turning and advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition, 28,* 657-666.

Piercey, C. D., Medler, D. A., & Hebert, B. E. (Eds.). (2003). Ambiguity Effects in Lexical Access: Do Blends Exist? (Vol. 8).

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107,* 786-823.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56-115.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14,* 191-201.

Rayner, K. & Duffy, S. A. (1987). Eye movements and lexical ambiguity. In J.K.O'Regan & A. Levy-Schoen (Eds.), (pp. 521-529). Amersterdam: North-Holland: Elsevier Science Publications.

Rueckl, J. G. (1995). Ambiguity and connectionist networks: Still settling into a solution: Comment on Joordens and Besner (1994). *Journal of Experimental Psychology: Learning Memory and Cognition, 21,* 501-508.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523-568.

When Input and Output Diverge: Mismatches in Gesture, Speech, and Image

Alissa Melinger (melinger@coli.uni-sb.de)
Department of Computational Psycholinguistics,
Saarland University, Saarbrücken, 66041, Germany

Sotaro Kita (sotaro.kita@bristol.ac.uk)
Department of Experimental Psychology, 8 Woodland Road
University of Bristol, Bristol BS8 1TN, United Kingdom

Abstract

The goal of the current paper is to investigate the behavior of gesture when the information conveyed by speech and the information conveyed by the image being described conflict as a result of perspective taking. To construct a corpus of speech-image mismatches, we designed a picture description elicitation procedure using path-like networks of colored circles. The results of our analysis demonstrate that gestures can be mismatched to both speech, as has been previously observed, and to the image, which has not been previously reported. The results provide insights into the nature of the representations that give rise to gestures.

The Origin of Gesture

This paper investigates the underlying cognitive processes involved in relating spatial information from a visual input to two separate output modalities, namely speech and gesture. Three theoretical possibilities have been proposed for how these three modalities of representation, visual-spatial, verbal and gestural, are related: The Lexical Semantic Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984), the Free Imagery Hypothesis (Krauss, Chen, & Chawla, 1996; Krauss, Chen, & Gottesman, 2000; but see de Ruiter, 1998, 2000 for another version of this hypothesis), and the Interface Hypothesis (Kita & Özyürek, 2003).

The Lexical Semantic Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984) proposes that gestures are generated from the semantics of the lexical items chosen to express the desired message. It predicts that gestures should always correspond to the meaning expressed by specific lexical items. In contrast, the Free Imagery Hypothesis (Krauss et al., 1996, 2000) claims that gestures are generated on the basis of pre-linguistic non-propositional representations; the strong reading of this proposal implies that the information conveyed by gesture should be unaffected by the specific lexical items selected during formulation and by the 'thinking for speaking' (Slobin, 1987, 1996) processes that convert the imagistic representation into propositional content (however, see below for alternative readings of this proposal). The Interface Hypothesis (Kita & Özyürek, 2003) claims that gestures originate from a mediating representation connecting spatio-motoric representations in memory and linguistic representations. According to this view, gestures are generated from the imagistic representation, but they can also be influenced by 'thinking for speaking' operations on this representation.

Discriminating between these theoretical alternatives is difficult because there is usually a close isomorphism between the semantic content of speech and the imagistic content of the representation speech is describing.

Bearing on this discussion are recent studies demonstrating that gestures can convey complementary information to what is expressed in speech. For example, when describing their solutions to the Tower of Hanoi problem, speakers' gestures sometimes corresponded to possible strategies that were not mentioned in the concurrent speech rather than to the strategy that was mentioned in speech. (Garber & Goldin-Meadow, 2002). The non-isomorphism between the content of speech and gesture has been referred to as *speech-gesture mismatches*. High rates of speech-gesture mismatches have also been reported for children who are in the transitional stage of acquiring the ability to correctly respond to the Piagetian conservation task (Church & Goldin-Meadow, 1986).

Speech-gesture mismatches appear to contradict the claims of the Lexical Semantic Hypothesis in that they express information not included in speech. However, the information expressed by gesture in the speech-gesture mismatches does not actually conflict with either the linguistic or the imagistic representation; instead they provide complementary information. Thus, they do not provide a strong test of the competing theories. In this paper we employed perspective taking to create situations in which what was said *conflicted* with what was seen. Examining the behavior of gesture in these cases should discriminate between the competing hypotheses regarding gesture generation. The Lexical Semantics Hypothesis predicts that gestures will always align with the speech. The Free Imagery Hypothesis predicts that the gestures will always align with the image. The Interface Hypothesis predicts that gesture alignment will be influenced by 'thinking for speaking' processes and therefore the alignment of gesture could be to either or both representations, depending on the specific situation.

Perspective Taking

Perspective taking is a critical step required to express spatial relations in speech (cf. Miller & Johnson-Laird,

1976). Spatial representations, which are inherently relative, must be grounded to some referent in a scene. The choice of the grounding referent impacts the linguistic terms that can be selected to express the relationship. Thus, perspective taking necessarily precedes linguistic formulation. It forms part of the 'thinking for speaking' conceptualizing process (cf. Levelt, Roelofs & Meyer, 1999) that abstracts away from the visual imagery and maps the relations onto propositional representations. Choice of perspective can be influenced by, among other things, language/culture specific resources (Levinson, 2003), the specific task at hand (Tversky, 1991) and/or pragmatic concerns (Levelt, 1996).

Consider the image in Figure 1. In describing the relationship between the ball and the car, the speaker can select himself as the grounding referent, describing the relationship from his own personal orientation, as in (1). This perspective will be referred to as the *deictic* perspective. Alternatively, he can select one of the objects in the scene as the grounding referent, such as the car, and describe the relation with respect to the car's inherent orientation, producing the *intrinsic* description in (2).

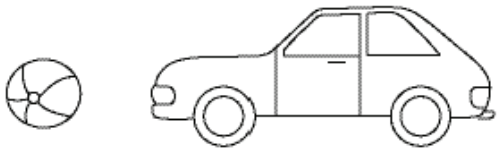


Figure 1

(1) The ball is to the left of the car.
(2) The ball is in front of the car.

When speakers choose to describe the relationship between the ball and car as in (2), a special situation arises; namely, the characteristics of the visual input, pre-abstraction, do not match the linguistic terms used to describe them. On the two dimensional representation of the image, nothing is *in front* of the car; the notion of *front* used in (2) is only relevant with respect to the orientation of the car — only within the perspectivized mental representation of the image. This contrast between the pre-abstraction visual input and the perspectivized mental representation provides the gesture researcher with the opportunity to contrast the content of the image with the content of speech in a unique way. Specifically, the content of the input imagistic representation and the output linguistic representation can be pitted against each other. How gesture behaves when the input and output representations conflict will reveal the underlying representation from which gesture was generated, thus discriminating between the three hypotheses.

Mismatch Corpus

To compile a corpus of speech-image-gesture mismatches, we presented speakers with networks of colored circles arrayed along a path. The images were very similar to networks used previously by Levelt (1996) to investigate perspective taking in speech production. As with other spatial relations, adopting different linguistic perspectives to describe an image, such as in Figure 2, results in the use of different linguistic terms to express the same spatial relations, as seen in examples (3a) and (3b).

Deictic Sample descriptions:
(3a) You begin with a yellow circle. *Above* that you see a blue circle. To the *right* you see a red circle and *above* the red circle you see another red circle. *Right* of the second red circle is the yellow circle and *right* of that is a blue circle.

Intrinsic Sample descriptions:
(3b) You begin with a yellow circle. You go *straight ahead* to a blue circle. Then you go to the *right* to a red circle and then *left* to another red circle. From the second red circle, go to the *right* again to a yellow circle and then *straight ahead* to a blue circle.

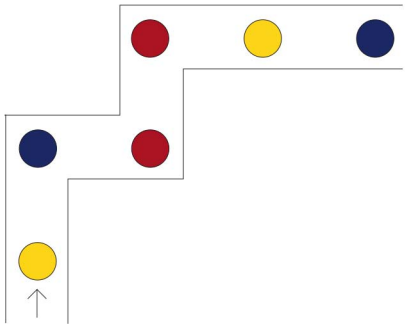


Figure 2

Notice that the term *straight ahead* in description (3b) is used to refer to two different directions of transition. First, it is used for the vertical transition from the first (yellow) circle to the second (blue) circle. Later, it is used again to refer to the lateral transition from the second from the last (yellow) circle to the last (blue) circle. In contrast, the terms used in description (3a) hold a constant relationship to a particular axis on the paper. For the deictic description, there is perfect isomorphism between the input image and the output description. In contrast, for the intrinsic descriptions, there is non-isomorphism, which allows for a further investigation of how gesture is related to the two representations.

If gesture is generated on a faithful memory representation of the image, as proposed by the strong version of the Free Imagery Hypothesis, then, in cases of speech-image mismatch, gesture should align to the image and conflict with speech. If gesture is generated from the lexical semantics of the words used to encode the message,

as suggested by the Lexical Semantics Hypothesis, then gesture should always match the speech and conflict with the image. If gesture is generated from an interface representation that results from 'thinking for speaking' processes, as suggested by the Interface Hypothesis, then gestures may match preferentially either the image or the speech, depending on the needs of the speaker at any given moment. In this case, characteristics of the image, the lexical item, or the situation could affect to which representation the gesture is aligned.

Constructing the Corpus

Speakers. Sixteen native speakers of Dutch from the Max Planck Institute for Psycholinguistics' subject pool were paid for their participation.

Pictures. Sixteen path-like images depicting networks of colored circles were constructed. Each image consisted of an explicit start point as well as red, yellow, and blue circles arrayed along a path. Half of the pictures had branching paths while the other half did not. All speakers saw all pictures in the same presentation order. In sum, 256 picture descriptions were collected.

Procedure. Speakers were seated across from their interlocutor separated by a visual block. Their task was to describe the pictures to the interlocutor, who was, in fact, a confederate.

Speakers were given approximately 15 seconds to study the image, which was placed on the table by the experimenter. After this memorization period, the picture was removed and the speaker began to describe the image. Speakers were free to describe the routes in any way that was natural to them; they were not given any linguistic examples to bias their description strategy. The listener was instructed not to ask any specific questions that might bias the content of the descriptions. She was free, however, to ask the speaker to repeat portions or even the entire description of an image. All sessions were video recorded.

Coding system. A native Dutch speaker familiar with gesture transcription systems but blind to the hypotheses under investigation used the videotapes to create a transcription of the speech as well as a record of all gestures. Several types of linguistic information were identified, including directional information (e.g., *right*, *left*, *straight ahead*), destination information (e.g., *a red circle*, *a blue circle*), landmark information (e.g., *you arrive at an intersection*), and shape information (e.g., *you will travel in a big circle*). In this paper, we will focus exclusively on directional information and accompanying directional gestures.

Gestures were either produced with the head or hands. Both were coded for several features, most crucially for the direction of the stroke but also for handedness. Once speech and gesture were fully transcribed, they were coded for three binary features: Speech matches image, gesture matches speech, and gesture matches image. These codes, together with codes for which directional term was produced and unperspectivized direction of transition, form the bases of the mismatch analysis.

Speech-image mismatches were identified as any transition in the network for which the verbal description provided in the intrinsic perspective did not match the actual direction of the transition in the network. For example, any transition labeled *right* that did not progress rightward on the page was a mismatch. Likewise, any use of *straight ahead* that did not correspond to an upward transition was coded as a mismatch. (Note that as the image was placed on the table in front of the speaker, the upward transition in the image was in the forward direction for the speaker, for which *straight ahead* is felicitous.)

Every picture provided multiple mismatch opportunities. For example, 11 networks included an upwards transition, similar to the transition from circle 3 to 4 in Figure 2, which intrinsic speakers described as *right* or *left*. Eight networks included lateral transitions, similar to the final movement in Figure 2, linguistically described as *straight ahead.* Four networks included downward transitions, linguistically described as *right* and three networks including lateral transitions that followed downward transitions. These transitions leftwards or rightwards were described with the opposite directional term, namely, rightward turns were described as *left* and vise versa.

Corpus Analysis

In this section we first give some descriptive details of the corpus before we turn to the crucial questions under investigation.

Characteristics of speech and gesture varied greatly between speakers. Six speakers produced almost no gestures at all. Of the ten gesturers, three produced predominantly deictic descriptions and seven produced predominantly intrinsic descriptions. While the deictic speakers are generally orthogonal to the issue of speech-image mismatch, two produced some mixed perspective descriptions, producing mismatch opportunities. Only speakers who adopted the intrinsic frame of reference AND who gestured are of relevance to our investigation.

In total, the corpus of directional terms consisted of 1440 directional tokens, 389 of which were produced with a co-expressive gesture. Table 1 presents lexical, gesture, and mismatch frequencies for each of the directional terms found in our corpus.[1]

[1] We will present English translations for the Dutch directional terms found in our corpus. With respect to the description of our images, there are no critical differences in how directions and spatial relations are lexicalized.

Table 1. For each directional term, the total number of tokens, the percentage of tokens produced with a gesture, and the number of speech-image mismatches.

Lexemes	Total Tokens	% With Gesture	Speech- Image Mismatches
Right	494	27%	100
Left	317	26%	79
Straight ahead	321	15%	62
Up	120	8%	--
Down	19	10%	--
Back	106	37%	--
Further	63	13%	--

The three directional terms that are relevant for mismatches are *right, left,* and *straight ahead*. In 241 instances, these words did not match the direction in the image. 58 of these speech-image mismatches were produced with a gesture that could either align with the linguistic term or the direction in the image. The terms *up* and *down* were only used by deictic speakers and therefore always matched the input image. The terms *back* and *further* can only be interpreted in the context of prior movements, and therefore the question of whether they match the picture is not applicable

The term *back* received the highest proportion of co-expressive gestures. The terms *left* and *right* were each produced with co-expressive gestures over 25% of the time, *further* and *straight* were produced with intermediate gesture rates and *up* and *down* had the lowest gesture rates.

We now turn to the central question of the paper. The crucial data from the corpus are gestures produced when speech and image are mismatched. The critical question is whether these gestures reflect the direction represented in the image, in speech, or both. The number of gestures that matched the image or the speech for each directional term is presented in Table 2.

Table 2. Number of speech-image mismatches for which the gesture matches either the speech or the image, for the three relevant directional terms.

Lexeme	Gesture = Image	Gesture = Speech
Right	8	15
Left	9	10
Straight ahead	13	3

The distribution of cases where the gesture matches the image compared to when it matches speech is different for the three directional terms, *right, left,* and *straight ahead*, χ^2 (2) = 7.13, $p < .05$. Two by-speakers comparisons were also carried out to assess this relationship. First, for the cases in which gesture aligned with speech, the proportion of speech-image mismatches for each of the three lexemes was calculated for each speaker by dividing the number of speech-image mismatches for the lexeme divided by the total speech-image mismatches for all three lexemes. The proportions differed significantly between the lexemes, Friedman's χ^2 (2) = 7.3, N=8, $p < .05$. Second, for the cases in which gesture aligned with image, the proportion of speech-image mismatches for each of the three lexemes was calculated for each speaker in the analogous way to the previous analysis. There is no evidence that proportions differed between the lexemes, Friedman's χ^2 (2) = 0.9, N=8, $p > .1$.

Table 2 shows that gesture alignment patterned differently for different lexemes. Table 3 further breaks down the information for different directions of transition within the network.

Table 3. Number of speech-image mismatches in descriptions of either upwards, downwards or lateral transition in which gesture aligned to either the image or to speech.

Transition Direction	Lexeme	Gesture = Image	Gesture = Speech
Up	Right or left	5	12
Down	Right or left	5	11
Laterally	Right or left	7	2
Laterally	Straight ahead	13	3

The alignment pattern for vertical (up and down) transitions was significantly different compared to lateral transitions, χ^2 (1) = 12.15, $p < .001$. Speakers preferred to align with the image when the transition was lateral but preferred to align with speech when the transition was vertical. One possible interpretation of this pattern is that speakers generally prefer to gesture laterally.

In speech-image mismatch cases, gestures sometimes aligned with the image and sometimes with the speech. This split was quite even for gestures produced for the lexemes *left* and *right* but not for *straight ahead*. In the latter case, speakers preferred to align their gestures with the image. The different alignment patterns to different lexical items may also be interpreted in terms of a general preference to gesture laterally rather than vertically.

What the data from the corpus clearly indicate, however, is that there is no strong tendency to align gestures to the image at the expense of speech or vise versa. When the information conveyed in speech conflicts with the information presented in the visual input, gesture can align with either. The decision as to whether a gesture aligns with the con-current speech is mediated by a spatial factor (lateral vs. vertical transitions). This result was not predicted by the Free Imagery Hypothesis or the Lexical Semantic Hypothesis, as we will discuss in more details in the next section. The result is, however, compatible with the Interface Hypothesis.

Discussion

By using images consisting of path-like networks of circles, we succeeded in constructing a corpus of picture descriptions in which the content of speech and the content of the to-be-described image often conflicted. Our aim was

to see whether gestures produced in these instances would be co-expressive with the lexical affiliate, as predicted by the Lexical Semantics Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984), with the characteristics of the image, as predicted by the strong reading of the Free Imagery Hypothesis (Krauss et al., 1996, 2000), or whether the alignment to one representation or another would be influenced by 'thinking for speaking' processes, as proposed by the Interface Hypothesis (Kita & Özyürek, 2003).

What our corpus analysis reveals is that gesture alignment behavior in speech-image mismatches was not driven solely by either the characteristics of the input image or the characteristics of speech. Rather, the gestural content seemed to be co-determined by the lexeme choice and the type of spatial representation. Specifically, when the lexemes *left* and *right* were used to express the spatial representations *upwards* and *downward*, gesture tended to align with speech rather than with the spatial representation. When the lexeme *straight ahead* was used to express the spatial concepts *leftwards* and *rightwards*, gestures tended to align with the spatial representation of the image. When the lexemes *left* and *right* were used to express the spatial representations *rightwards* and *leftwards*, respectively, gesture again tended to align with the spatial representation.

The fact that the gestural content was determined by the interplay between both lexical and spatial representations makes it difficult to maintain either the Lexical Semantics Hypothesis, which holds that gestures are generated from the semantic representations of lexical items that have been selected for speaking, or the strong version of the Free Imagery Hypothesis, which holds that gestures are generated from pre-linguistically generated imagery.

However, we need to recognize that there are different versions of the Free Imagery Hypothesis, which make different assumptions. In de Ruiter's (2000) version of the Free Imagery Hypothesis, gestures are generated in the Conceptualizer in Levelt's (1989) sense, which generates the (pre-linguistic) proposition to be linguistically formulated in the next utterance. According to de Ruiter, both gestural and linguistic perspectives are determined in the Conceptualizer. Similarly to the strong version of the Free Imagery Hypothesis, "the shape of the gesture [iconic gesture] will be largely determined by the content of the imagery" (de Ruiter, 2000: 293). As such, the results from the present study are problematic not only to the strong version of the Free Imagery Hypothesis, but also to de Ruiter's version. However, because in de Ruiter's model the shape of a gesture is determined in the Conceptualizer, which in principle has access to (pre-linguistic) propositions to be linguistically formulated, it might be possible to modify the model to account for the present results.

The gestural content is determined by the interplay between lexical choice and directions of the transition in the image. This result could be accounted for by the Interface Hypothesis (Kita & Özyürek, 2003), which proposes that gestures are generated from an interface representation, namely, a spatio-motoric representation that is in the process of being prepared for speech. According to this hypothesis, there is a general tendency for an interface representation to converge with the linguistic representation in the utterance being planned. The degree of convergence is determined by various contextual factors (Kita, 2000). In the case of this study, when the spatial representation of the transition is confusable, that is, when the transition is lateral (i.e., *leftwards* or *rightwards*), the convergence to the linguistic representation is weak, and thus gesture tends to match the spatial representation of the transition, rather than the linguistic representation. When the spatial representation of the transition is not confusable, that is, when the transition is vertical (i.e., *upwards* or *downwards*), the interface representation converges strongly to the linguistic representation. Note further that the idea that gestures help distinguish confusable spatial representations is compatible with theories of self-oriented functions, in particular, the theory that gestures help organize spatio-motoric information for speaking (Kita, 2000; Alibali, Kita, Yong, 2000; Kita, 2003).

The data also rule out the possibility that gestures can be randomly generated from either the input imagistic representation or the output lexical representations, alternating randomly between these two sources. This possibility, previously discussed in Kita and Özyürek (2003) predicts that speech-gesture mismatches should randomly align to the input or to speech, without a discernable pattern. This is not the observed pattern, as seen in Table 3.

One could object to our definition of speech-image mismatch. Consider for example the possibility that speakers mentally rotate the image in memory in order to calculate the correct directional term for the intrinsic perspective. In this case, apparent speech-image mismatches would in fact be matches. However, if this were true we would not have expected gesture alignment to ever conflict with speech, since speech would always match the perspectivized internal memory representation of the image for the speaker at that moment. This is not consistent with the observe data pattern, as seen in Table 1.

The gesture-speech mismatches reported in this study are of a different type from what have been attested in previous research. Gesture-speech mismatch has been observed in children's explanations for Piagetian conservation tasks (Church & Goldin-Meadow, 1986) and children's explanations for equivalence of an equation (Perry, Church, & Goldin-Meadow, 1988), and adult and children's description of the solution to the Tower of Hanoi puzzle (Garber & Goldin-Meadow, 2002). In these studies, gesture and speech refer to two distinct referents that are both relevant to the current goal of discourse, or two alternative strategies or solutions that might apply to the problem at hand. For example, in the explanation for a Piagetian conservation task, speech may indicate the height of a glass, *this one is tall*, and gesture may indicate the width of the same glass. By contrast, mismatches that result from perspective taking can be called same-referent mismatches. Speech and gesture have the same referent, namely a motion

vector, but they map the vector to a gestural body movement under different perspectives. Using these same-referent mismatches allows the predictions of the competing theories to be properly tested.

To conclude, we have introduced a new source of evidence into the field of gesture research, namely *speech-image* mismatches with concomitant gestures. These speech-image mismatches allow the content of the linguistic and the imagistic representations to be separated and contrasted. An analysis of the behavior of these gestures revealed that gestures cannot be generated from a purely linguistic or purely imagistic representation. Rather, gestural content was determined by the interplay between the lexical items used in the description and the type of directional information in spatial representation of the transitions. Many issues in gesture research have had difficulty in finding clear evidence for or against specific proposals exactly because it is generally difficult to disentangle the independent contributions of linguistic and imagistic representations. The present paper uses perspective taking to avoid this problem. The present study is also significant in that the methodology affords reliable elicitation of same-referent mismatches from normal adult speakers.

Acknowledgments

We acknowledge useful comments from the members of the Gesture Project at the Max Plank Institute for Psycholinguistics. We greatly appreciated comments on an earlier version of this paper from Joana Cholin, Asifa Majid and Andrea Weber. We also benefited greatly from insightful comments by anonymous reviewers. Special thanks also go to Anne-Marie van Hoof and Esther Vrinzen for acting convincingly as our confederate and for transcribing the descriptions.

References

Alibali, M. W., Kita, S, & Young, A. J. (2000). Gesture and the process of speech production: we think, therefore we gesture. *Language and Cognitive Processes, 15,* 593-613.

Butterworth, B. & Hadar, U. (1989). Gesture, speech and computational stages: A reply to McNeill. *Psychological Review, 96*, 167-174.

Cassell, J., McNeill, D., & McCullough, K. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*, 1-33.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition 23*, 43-71.

de Ruiter, J.-P. (1998). *Gesture and Speech Production*. Ph.D. Dissertation, Max Planck Institute for Psycholinguistics, The Netherlands.

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284-311). Cambridge: Cambridge University Press.

Garber, P., & Goldin-Meadow, S. (2002). Gesture offers insight into problem solving in adults and children. *Cognitive Science, 26*, 817-831.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining match: Gesturing lightens the load. *Psychological Science, 12 (6)*, 516-522.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162-185). Cambridge: Cambridge University Press.

Kita, S. (2003). Interplay of gaze, hand, torso orientation and language in pointing. In S. Kita (Ed.), *Pointing: where language, culture, and cognition meet* (pp.307-328). Mahwah, NJ: Lawrence Erlbaum

Kita, S. & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*, 16-32.

Krauss, R., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology, 28*, 389-450.

Krauss, R., Chen, Y., & Gottesman, R. (2000). Lexical gestures and lexical access: a process model. In D. McNeill (Ed.), *Language and gesture: Window into thought and action.* Cambridge, UK: Cambridge University Press.

Levelt, W. (1996). Perspective taking and ellipsis in spatial descriptions. In P. Bloom, M. A. Peterson, M. F. Garrett, & L. Nadel (Eds.), *Language and space* (pp. 77-107). Cambridge: MIT Press.

Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1-38.

Levinson, S. C. (2003). *Space in language and cognition: Exploration in cognitive diversity.* Cambridge: Cambridge University Press.

Melinger, A. & Kita, S. (Submitted). Conceptual load triggers gesture production.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and Perception.* Cambridge, MA: Harvard University Press.

Perry, M., Church, R.B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359-400.

Schegloff, E. A. (1984). On some gestures' relation to speech. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversational analysis.* Cambridge: Cambridge University Press.

Slobin, D. I. (1987). Thinking for speaking. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Proceedings of the 13th annual meeting of the Berkeley Linguistic Society* (pp. 435-445).

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96). Cambridge: Cambridge University Press.

Tversky, B. (1991) Spatial mental models. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory, Vol. 27* (pp.109-146). New York: Academic Press.

Cognition in Jazz Improvisation: An Exploratory Study

David Mendonça (mendonca@njit.edu)
Information Systems Department, 323 Martin Luther King, Jr. Boulevard
Newark, NJ 07102 USA

William A. Wallace (wallaw@rpi.edu)
Department of Decision Sciences and Engineering Systems, 110 8th Street
Troy, NY 12180 USA

Abstract

This research investigates thinking processes of duos of jazz improvisers in performance. Of particular interest are cognitive processes related to creativity and to reasoning about time, since both activities are fundamental to improvisation. Data sources are the group members' retrospective verbal protocols, collected after the performance of each tune. One result of this work is that cognition related to reasoning about time and creativity varied little either within or across groups, regardless of the type of tune being played. This result is further investigated by examining some of the statements from the protocols themselves.

Introduction

This research examines the thinking processes of duos of jazz improvisers in performance. Of particular interest are cognitive processes related to creativity and to reasoning about time, since both activities are fundamental to improvisation. To produce the data used in this study, each member of the duo watched and listened to a tape of the duo's performance and recalled out loud what he had been thinking during it. The study addresses a gap in prior research by presenting an analysis of cognition during improvisation as reflected in verbal protocol data. Since this study is thought to be the first of its kind, an exploratory approach is taken.

A brief review of related work is followed by a presentation of exploratory questions concerning how performers reason about time and how they produce ideas for performance. The results of the study are then presented, followed by an analysis of the contents of selected statements in the protocols. The paper concludes with a discussion of implications for current theory and directions for future work.

Related Prior Research

Improvisation in jazz is said to involve "reworking precomposed material and design in relation to unanticipated ideas conceived, shaped, and transformed under the special conditions of performance, thereby adding unique features to every creation" (Berliner, 1994). While improvising has been compared to "real-time composing" (Kernfeld, 1988), the two differ in salient ways (Nettl, 1974). Composition refers to "the discontinuous process of creation and iteration (usually through notation) of musical ideas" (Sarath, 1996). Improvisation, by contrast, is a continuous and serial process. Composing involves distributing musical elements (such as notes) over a score that is to be played serially: the composer may add to, delete or edit any part of the composition at any time before its performance. Performance of a composition involves interpreting and articulating a written or memorized score. Performance of an improvisation involves *conceiving, articulating* and *remembering* an unwritten, evolving score (Berliner, 1994). While a misplaced note in a composition can be erased and rewritten; a misplayed note in improvisation cannot. Errors in improvisation therefore "must be accepted as part of the irrevocable chain of acoustical events, and contextually justified after the fact by reinforcement or development" (Pressing, 1984). As stated by Pressing (1984), "If erasing, painting over, or non-real-time editing exist, improvisation does not."

Temporal Cognitive Processes

Following Berliner's (1994) comments, improvisers must reason about time in order to *conceptualize* what is to be *articulated* in light of what they *remember* has been played (see Sarath, 1996 for further discussion). In comparison with a tune that is being composed as it is being played (i.e., a free tune), the performance of a well-learned tune (such as a jazz standard) may place fewer demands on remembering, since much of what needs to be recalled (e.g., chord changes) is easily accessible from long-term memory (Johnson-Laird, 2002). Similarly, conceptualization may also be easier for a jazz standard, since the path in front of the improviser is better known. A tentative hypothesis, then, is that players will spend more cognitive effort on remembering and planning ahead for a free tune than on a standard.

Creative Cognitive Processes

Creativity—the production of new ideas—is fundamental to improvisation, since it is not enough for improvisers to produce music that has already been composed: they must produce something that, to them at least, is new. There have been numerous proposed models of the cognitive processes involved in creative thinking (Sternberg, 1999) and on the factors that influence creative thinking (Welsh, 1973). These

theories typically include convergent and divergent processes, along with some mechanism that governs switching between them (Newell, 1962). The Geneplore model (Finke, Ward, & Smith, 1992) describes creative thinking as entailing divergent processes of generation and subsequent exploration of ideas (see Ward, Smith, & Finke, 1999 for a discussion). Evaluation (a convergent process) is discussed in Geneplore in terms of constraint satisfaction.

A jazz standard may afford more opportunities for divergent thinking than a free improvisation. Jazz standards are tunes with familiar structures and a long history of performance which players routinely draw upon. Their performance may enable the improviser to spend less time trying to recall the tune or, in the case of free improvisation, trying to determine the structure and content of the tune. Similarly, evaluation may be easier to accomplish for a jazz standard, since the player can easily recall what has been played and speculate reliably what is to be played.

Analytic Framework

The analytic framework for this study is used to define a set of temporal and creative processes. A scheme for classifying the contents of the protocols based on these definitions was developed so that independent coders could identify these processes in the protocols. This section provides the definitions of the processes; the method of their application is discussed in the subsequent section on "Study Design."

An improviser in performance must reason about past, present and likely or possible future events, resulting in three different processes related to *temporal* cognition. *Orientation* is the process of considering a current performance event. An example of orientation is the statement "The time is in 4," since the speaker is referring to the present moment. *Retrospection* is the process of recalling a previous performance event. The statement "He had just played in three so I did too" is an example of retrospection. *Prospection* (a term coined for this research) is the process of looking ahead; that is, of predicting or speculating about a future event in the performance. An example of prospection is the statement "I knew I was coming to the end of my solo so I looked up."

Three different types of *creative* cognition are considered here. Idea *generation* is said to occur when a musical idea (i.e., one that pertains to the performance) is recalled or created. An example is the statement "I was thinking about playing an open figure there." Idea *development* is said to occur when a player further develops a musical idea which has already been generated, either by the speaker or the other member of the group. An example is the statement "I was thinking of inverting the figure I played previously." Both idea generation and development are regarded as divergent processes. Finally, idea *evaluation* is said to occur when the speaker makes a statement about the value or worth of a musical idea. An example is the statement, "I remember liking what I played there." Idea evaluation is a process that leads to decisions about whether or not to pursue ideas, and is therefore convergent.

Research Questions The research questions concern the frequencies of occurrence of temporal and creative cognition within and among groups during the performances of various jazz tunes. Although some tentative hypotheses have been discussed, a broad range of questions are addressed. This decision is due to the exploratory nature of the study, one goal of which is to provide suggestions for further lines of research. The research questions explore (i) the defensibility of the assumption of between-group homogeneity; they also examine possible differences in creative and temporal cognition (ii) within a group for a particular tune and (iii) across tunes by the same group. Data from performances of a jazz standard and a free improvisation are used. An example of question (i) is, do the proportions of occurrence of the various types of temporal processes differ across the groups for the performance of a jazz standard? An example of question (ii) is, do the proportions of occurrence of the various types of temporal processes differ between the trumpet and bass player in the performance of a jazz standard? An example of question (iii) is, do the proportions of occurrence of the various types of temporal processes differ between the performance of a jazz standard and of a free improvisation for a particular group?

Study Design

The study employs the preceding three research questions to investigate the impact of different types of tunes on how members of professional jazz duos reason about time and think creatively.

Tune Choice The tunes which the duos were asked to perform were intended to vary in difficulty and familiarity. The first tune played, "I Got Rhythm," has been extensively recorded and is the origin of the so-called "Rhythm Changes," a set of widely-used chord changes in jazz (Berliner, 1994; Kernfeld, 1988). All participants were expected to be very familiar with this tune and comfortable in improvising on it. The second tune, "Willow Weep for Me," was chosen since it is not as familiar as "I Got Rhythm" and because it is typically played as a blues ballad. In case a duo did not know this tune well enough to play it without sheet music, a backup tune— "Blue Train"—was used. The third tune, "Giant Steps," is known for being a difficult tune, since the chord progressions are highly idiosyncratic and the tempo is fast. The backup tune for "Giant Steps" was "Cherokee," which is in part challenging because it is also typically played at a quick tempo. The fourth tune was a free improvisation. A free improvisation has no pre-determined structure other than the one which performers create.

Solicitation of Participants Groups of two players each were solicited through word-of-mouth contact with professional musicians working in the Albany, New York area. (Though the duo is a common configuration in jazz, the use of duos was also due to practical concerns of cost

and the availability of appropriate recording space.) The player of the lead instrument—a trumpet player in all cases—was first secured. The lead player then suggested the second member of the duo, someone with whom he played often. The groups were intended to be reasonably homogeneous in terms of musical experience and experience improvising. Accordingly, all participants were asked to describe their backgrounds. Because they would be asked to verbalize their thoughts, only players who had experience in teaching—and therefore talking about—improvisation were asked to take part in the study. Finally, they were then told that, during the study, they would be given three or four tunes to play, videotaped during the performance of each tune and then asked to recall their thinking while watching the videotape.

Procedure

The study was conducted at a local professional recording studio. Once a duo arrived, study personnel reviewed the study protocol with them. The duo's members next signed consent and contract forms, tuned their instruments and performed a sound check.

Each member of the duo then entered a vocal (i.e., isolation) booth to practice giving concurrent and retrospective verbal protocols, using two tasks from the literature (Ericsson & Simon, 1993). In the third task they whistled or sang a short tune then recalled their thinking during the performance of the tune. Once these practice tasks had been completed to the player's and the experimenter's satisfaction, the players practiced operating the VCR that would assist them in giving the retrospective verbal protocol. The main part of the study then began.

The group first played "I Got Rhythm." They were asked to keep the performance to less than ten minutes and to take one solo each. A maximum length of ten minutes was chosen since that is the maximum length recommended in guidelines for conducting a retrospective verbal protocol (Ericsson & Simon, 1993). Once any questions had been answered, all personnel except two video camera operators left the main studio and the duo performed the tune. The videotape recorded the images and sound of the two performers, along with a display of the duration of the performance.

After the performance, each participant went to a vocal booth in order to watch and listen to a videotape of the performance and to deliver the protocol. (About two to three minutes usually elapsed between the end of the performance and the beginning of the protocol.) As is commonly done in verbal protocol-based studies, one experimenter remained with the participant in order to reiterate the instructions and to ask the participant to "keep talking" whenever there was silence for more than about ten seconds (Ericsson & Simon, 1993). While giving a protocol, each participant could control the videotape as necessary. All protocols were audio- and video-taped, so that it would be possible to hear and see what was on the videotape while the participant was speaking.

Once the protocol had been delivered, the participants returned to the performance area and prepared for the next performance. The above procedure was repeated for the tunes "Willow Weep for Me" and "Giant Steps." If the participants decided not to play a tune, the backup tune was played. For the fourth and final tune, participants were asked "to work out the composition of the tune as you play it." Again they were asked to keep the total length of the performance to less than 10 minutes. They were not asked to take solos, since doing so would have helped determine the structure of the performance.

Once the think-aloud protocol for the final tune had been given, participants were asked a series of questions about their background and their participation in the study. They were then paid and invited to discuss the study further in a relaxed atmosphere without being recorded. The total duration of each session was approximately two hours.

Results

Data are taken from the verbal protocols associated with the performances of "I Got Rhythm" (IGR) and a free improvisation (Free). Performances of these two tunes were chosen for initial analysis since they represent opposite ends of the spectrum of jazz performance. All the groups played IGR first and Free fourth.

Participants' protocols were first transcribed and segmented (Ericsson & Simon, 1993). All references to study participants in the protocols were masked so that it would not be possible for a reader to determine which protocol corresponded to which session or player. Segments pertaining to cognition during the actual performance of the tune were coded using the definitions for the types of temporal and creative cognition that are given above in the "Analytic Framework" section. Coders were provided with (i) the segmented and masked protocols and (ii) instructions on how to use the above definitions to code the protocols. Coding was done by two independent coders unfamiliar with the objectives of the research. The coders were trained first to identify creative processes, then applied the instructions to the protocols. The same procedure was then followed for temporal processes. A second coder coded approximately 10% of the data, and reliability as measured by Cohen's kappa (Cohen, 1960) was approximately 87%.

Counts of the various types of temporal and creative processes were then entered into contingency tables. For creative cognition, it was immediately obvious that there were too few instances of idea development to justify the use of the appropriate statistical test, the Chi-squared test for differences in proportions (Conover, 1999). All instances of idea development were therefore recoded as instances of idea generation, since, as discussed previously, both are processes of divergent thinking. All statistical tests were then performed on the tables at a 0.05 significance level. The observed significance level of a test is denoted p. Some of the contingency tables are shown below, with the following symbols used: for temporal cognition, O=orientation, P=prospection and R=retrospection; for

creative cognition, *G*=idea generation and *E*=idea evaluation.

Between-group Differences

A reasonable degree of homogeneity was desired among the groups in order to minimize the possibility of between-group confounding effects. Information on participants' backgrounds was collected, as discussed previously. Also, differences in groups' temporal and creative thinking processes for each tune were investigated using question (i), with the following results. For temporal thinking, no significant between-group differences were found among the groups for either IGR or Free.

For creative cognition, no significant differences were found among the groups for IGR, but a significant difference (p=0.0045) was found for Free. This result may be stated by saying that at least two of the proportions in some column were not equal to each other. Table 1 shows the data associated with Free.

Table 1: Creative Cognition, by Group for *Free*.

Group	G	E
One	33	37
Two	13	9
Three	68	26

The assumption of homogeneity may therefore be seen as reasonable for IGR but not for Free. Accordingly, the analysis will consider individual groups rather than pooling data across groups.

Description of Performances

I Got Rhythm All groups structured their performances of IGR in approximately the same way, with the head (i.e., the introduction and first AABA chorus) and ending (i.e., the last AABA chorus) played more or less as discussed by Kernfeld (1995). In Groups One and Two, solos were two choruses long; in Group Three they were four choruses long. All performances were less than ten minutes long.

Free Improvisation For the free improvisation, participants were asked to "work out the composition of the tune as you play it." Groups in sessions one and two asked for some additional guidance but were given nothing more than a key and/or time signature. All performances were less than ten minutes long and had a stable time signature. It should be noted that, although all free tunes were spontaneously composed, all were clearly in the idiom of bebop jazz.

Temporal Processes

For question (ii), no significant differences in temporal cognition were evident between the participants in each group for IGR or Free. Table 2 shows the question (ii) data associated with Group One for IGR.

Table 2: Temporal Cognition for IGR, by Player in Group One.

Player	O	P	R
Trumpet	25	13	11
Bass	17	10	4

Similarly, for question (iii), no significant differences were evident across the two tunes as performed by each group. Table 3 shows the question (iii) data associated with Group One.

Table 3: Temporal Cognition for Group One, by Tune.

Tune	O	P	R
IGR	42	23	15
Free	30	21	21

Creative Processes

No significant differences were evident in creative cognition between the participants in each group for IGR or Free. So, in Group One, the proportion of segments from the trumpet player reflecting idea generation is not significantly different from the corresponding figure for the bass player. Table 4 shows the question (ii) data associated with Group One for IGR.

Table 4: Creative Cognition for IGR, by Player in Group One.

Player	G	E
Trumpet	18	16
Bass	8	6

No significant differences were evident across the two tunes as performed by each group. Table 5 shows the question (iii) data associated with Group One.

Table 5: Creative Cognition for Group One, by Tune.

Tune	G	E
IGR	26	22
Free	37	33

For both temporal and creative processes, the results suggest that—contrary to expectation—the same proportion of segments reflected each type of temporal or creative process. The same was true for differences within the same group across the two tunes. This result is particularly surprising, since groups were expected to approach IGR and Free quite differently; indeed, the recorded performances of the tunes by any given group, while sharing certain elements (e.g., stable key signature

and meter within each performance) nonetheless sound quite different, particularly in Group Three and to a lesser extent in Group One.

Content Analysis

The contents of the protocols from the performance of IGR by Group One are now analyzed in order to provide further insight into individual- and group-level creative and temporal cognition. These protocols were chosen because they are richer in content than those of the other groups and because it is appropriate to begin with an analysis of a simpler tune before moving to more complex ones. The statements examined here were those thought to provide the most insight into processes of collaboration, temporal reasoning and creativity. To aid the discussion, statements are labeled with their segment number from the protocols.

Collaboration Statements by one player about himself, the other player in the group and the group itself suggest how the duo collaborated. In IGR, the trumpet player (JH) explicitly mentioned trying to fit the melody and rhythm of his playing with that of the bass player (PT), as follows:

19. I'm just ah, I'm hearing melodies, I'm trying to play them, that I know will fit with what PT is playing.
24. I'm trying to keep, trying to keep a steady rhythm with PT, trying to make my eighth notes very steady.

JH also engaged in active listening following the completion of his own solo and the onset of PT's solo:

32. I'm kind of ah thinking after the thought and reacting to what he's playing.

Segments 19, 24 and 32 reflect JH trying to solve two types of problems related to collaboration. The first type, as in S24, is reasonably well-posed and technical. In this case, it involves keeping the rhythm of the tune. In contrast, the second requires the generation of new melodies: here, ones that "fit with" PT's playing. Segment 19 (S19) suggests that generation of the melodies occurred closely in time to their evaluation and performance. The use of the phrase "hearing melodies" would seem to indicate that the process is more one of retrieval rather than on-the-spot composition. Computational approaches to this type of thinking in jazz improvisation have sometimes involved retrieval and use of fragments or motifs (e.g., Ramalho, Rolland, & Ganascia, 1999) S32 is similar in spirit to S19, the distinction being that JH was not actually playing anything at that moment. The statements show that JH was actively listening to PT, analyzing PT's playing and attempting to use the results of this analysis to guide his own playing.

Temporal Processes A number of statements by both players reflect reasoning about time, particularly about the group's movement through the structure of the tune. Once they were asked to play IGR, one of the things they discussed was whether or not to play the tag (an optional ending to the melody). JH decided that they would play the tag at the very end of the tune. As they neared the end of the tune (i.e., the last A section), JH recalled thinking

47. I'm, I'm thinking about how we're going to resolve the tune, how we're going to end it here.
48. I know we're going to put the tag on, which is what we're doing right now.

Similarly, at about the same point PT recalled thinking

41. Ah, we're coming up on the last eight, eight bars.
42. And we had talked about putting the the tag at the end, so I'm actually thinking, yeah we're going to put that tag on the end.

Successful completion of the tune was therefore in part dependent on both players recalling the need to play the tag at the end of the tune. Additionally, JH was thinking about how to resolve the tune, given the need to include the tag.

At numerous points in the performance, JH and PT each speculated about what might be played by the other person. For example, after the first chorus of JH's solo, PT thought

22. Now I'm, right now I tell you I'm thinking, is he taking another chorus?
23. There, so, he's taking two choruses.

These segments suggest that the *de facto* structure of the tune was in part determined during performance, thus requiring the performers to think explicitly about past, present and future events. For example, because JH had taken two choruses in his solo, PT did the same.

Creative Processes An interesting exchange occurred at the very end of the group's performance of IGR. As shown in S47-48 and S41-S42 (above), JH and PT recalled the agreement to play the tag and planned for it. JH made the following statements immediately after S48:

49. Now, I didn't, I also extended the tag.
50. I could have made that, I could have made that tag a few bars, uh one or two bars shorter. By not extending it, I I kind of doubled the time.
52. Ok, as I said I I I doubled the length here, just to see what PT would do, how he'd react.
53. And I held that note out because that gave PT an opportunity to decide for both of us how exactly that was going to end.
54. Um and then I just threw that little tag of those couple notes on the end uh expecting he might react off of that,...

JH therefore elaborated upon the tag by doubling the time, holding out a note and adding a couple of notes to the tag. Each generated idea was intended to result in PT generating an idea in reaction to it, which would of course require evaluation. PT's reaction was as follows:

43. And right there I'm thinking, should I put a tag after his little ending there, but I decided to just let him have the final word.

S54 (above) concludes with JH saying "and he chose not to. And that was his choice." This example therefore shows cycle of idea generation (by JH), evaluation of those ideas by (PT), and finally JH's evaluation of PT's evaluation.

Discussion and Conclusions

The statistical analysis suggests a great degree of stability in creative and temporal cognition across the various conditions. Groups may therefore have applied similar cognitive strategies, regardless of the conditions of performance. The analysis of the "I Got Rhythm" protocols for Group One provides additional insight into how improvisers collaborate while simultaneously abiding by constraints of an evolving musical structure and generating, evaluating and executing new ideas.

Temporal cognition is necessary when a tune's structure evolves in real time. A theory of improvisation, even for the performance of standard tunes, should therefore include explicit modeling of temporal reasoning (see Johnson-Laird, 2002 and Ramalho et al., 1999 for discussions on both sides of this issue). A key consideration is range of planning (Palmer, 1997), since the current study and others (Sarath, 1996) have suggested that improvisers engage in contingency-based reasoning during performance. This study used prompted retrospective verbal protocols as primary data sources. Data on physical movements (Palmer, 1997) and cues and communications (Brinner, 1995) of performers may be useful in triangulating the results.

Further work is needed in understanding the role of knowledge and experience in the production of new musical ideas. A large body of work (see Pressing, 1984) shows that skilled improvisers draw upon and adapt highly resilient motifs during performance. Some evidence (Berliner, 1994) suggests that the use of these motifs can be conscious, though this claim has not been rigorously tested.

The results show that some creative and temporal processes may themselves be highly collaborative. Indeed, as suggested by Pressing (1984), the nature of improvised performance demands that the all "acoustical events" must be folded into the performance. The protocols contained evidence of both routine and non-routine problems arising out of collaboration between duos.

Finally, as discussed by Johnson-Laird (1991, 2002), additional work is needed in expressing theories of improvisation as computer programs. The current study has provided some evidence that such programs should include mechanisms for reasoning about (i) evolving conceptions of musical structure (and therefore time), (ii) processes of creativity and (iii) how dependencies among group members are negotiated (e.g., Bongers, 1999; Walker, 1997) in order to deal with temporal constraints while thinking creatively (Ramalho et al., 1999). Such an approach ought to result in a theory of improvisation that seeks to explain how it occurs in a wide variety of domains.

Acknowledgements

This research was supported through National Science Foundation Grant CMS-9872699. Additional support was provided by the iEar Department at Rensselaer Polytechnic Institute. We thank the musicians who participated in this study as well as Neil Rolnick and the personnel at Max Trax Recording Studios.

References

Berliner, P. F. (1994). *Thinking in jazz*. Chicago: University of Chicago Press.

Bongers, B. (1999). *Exploring novel ways of interaction in music performance.* Paper presented at the ACM Creativity and Cognition Conference, Loughborough, UK.

Brinner, B. (1995) *Knowing music, making music*, Chicago: University of Chicago Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: John Wiley & Sons, Inc.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis* (Revised ed.). Cambridge, MA: The MIT Press.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research and applications*. Cambridge, MA: MIT Press.

Johnson-Laird, P. N. (1991). Jazz improvisation: A theory at the computational level. In P. Howell & R. West & I. Cross (Eds.), *Representing musical structure*. New York: Academic Press.

Johnson-Laird, P. N. (2002). How jazz musicians improvise. *Music Perception, 19*(3), 415-442.

Kernfeld, B. (1995). *What to listen for in jazz*. New Haven: Yale University Press.

Kernfeld, B. (Ed.). (1988). *The new Grove dictionary of jazz* (Vol. 1). New York: Macmillan Press Ltd.

Nettl, B. (1974). Thoughts on improvisation: A comparative approach. *The Musical Quarterly, 60*(1), 1-17.

Newell, A., Shaw, J. C. & Simon, H. A. (1962). The processes of creative thinking. In Gruber (Ed.), *Contemporary approaches to creative thinking*. New York: Atherton Press.

Palmer, C. (1997). Music performance. *Annual Review of Psychology, 48*, 115-138.

Pressing, J. (1984). Cognitive processes in improvisation. In R. Crozier & A. Chapman (Eds.), *Cognitive processes in the perception of art*. Amsterdam: North Holland.

Ramalho, G. L., Rolland, P., & Ganascia, J. (1999). An artificially intelligent jazz performer. *Journal of New Music Research, 28*(2), 105-129.

Sarath, E. (1996). A new look at improvisation. *Journal of Music Theory, 40*(1), 1-38.

Sternberg, R. J. (1999). *Handbook of creativity*. New York: Cambridge University Press.

Walker, W. F. (1997). *A computer participant in musical improvisation.* Paper presented at the Computer-Human Interaction, Atlanta, GA.

Ward, T. B., Smith, S. M., & Finke, R. A. (1999). Creative cognition. In R. J. Sternberg (Ed.), *Handbook of creativity*. New York: Cambridge University Press.

Welsh, G. S. (1973). Perspectives in the study of creativity. *Journal Of Creative Behavior, 7*(4), 231-246.

Help-seeking with a computer coach in problem-based learning : Its interaction with the knowledge structure of the learning domain and the tasks' cognitive demands

Julien Mercier (jmercier@cgocable.ca)
Applied Cognitive Science Laboratory, McGill University
3700 McTavish St., Montreal, Canada, H3A 1Y2

Carl H. Frederiksen (carl.frederiksen@mcgill.ca)
Applied Cognitive Science Laboratory, McGill University
3700 McTavish St., Montreal, Canada, H3A 1Y2

The Problem and its Context

Research on tutoring has shown that the student's interaction with the tutor heavily determines the learning outcomes. In human tutoring, the responsibility of the interaction is shared between the tutor and the student (Chi, 2001). In the case of a computer coach such as the McGill Statistics Tutor, the control of the interaction is put entirely in the hands of the learners. Learners' ability to interact with the system productively therefore represents a critical aspect affecting the learning outcomes. This ability of help seeking (Nelson-LeGall, 1981) has not been well researched from a cognitive science point of view in the context of computer-supported learning (Aleven et al., 2003).

The aims of the present work are to elaborate a cognitive model of help seeking and to examine its interaction with critical aspects of the learning situation. Two studies using discourse analysis methodology are conducted using a formal model of the learning domain.

Methodology

First-level Participants are 20 graduate students from a faculty of Education of a Canadian university. The seven-hour experiment involves working in pairs to solve a very challenging statistics problem (a two-way analysis of variance) for which students don't have sufficient background. A computer coach based on human tutoring, the McGill Statistics Tutor, is available to provide help with every aspect of the task.

Data consist of three complementary sources. The dialogue between the pair of participants as they work on the statistics problem using the computer coach. The interaction with the computer coach is also recorded, in two forms. First, the display of the computer is recorded using a special device. Second, the computer coach keeps a log of some characteristics of every help request made by the students. The students solutions to the problem are also integrated in the database.

Data analysis consists of complementary strategies. Trace analyses of the task performance and the help seeking process were elaborated. Statistical analyses were also performed..

Results and Discussion

Results show that a help seeking model based on information processing theory is reflected in the data. The components of the model are (1) recognize an impasse, (2) diagnose the impasse, (3) establish a specific need for help, (4) find appropriate help, (5) comprehend help, and (6) evaluate help.

Help seeking interacts with the performance of the task and with the structure of the domain knowledge. Help seeking is intertwined with problem solving ; help is sought to fill gaps in students' knowledge in order to solve the problem. However, student's use of the computer coach is not optimal since they tend to select help at higher levels in the hierarchical knowledge structure while they tend to problem solve at lower levels.

Conclusion

These results have implications for the design of computer coaches and instructional situations. These results help characterize the contribution of the learners to the emergence of more or less contingent tutorial interactions. In addition, identifying key skills that students use in problem-based learning situations is a first step in training and assessing those skills.

References

Aleven, V., Stahl, E., Schworm, S., Fisher, F., & Wallace, R. (2003). Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research 73, 3*, 277-320.

Chi, M.T.H., Siler, S.A., Jeong, H., Yamauchi, T., & Hausmann, R.G. (2001). Learning from Human Tutoring. *Cognitive Science, 25*, 471-533.

Nelson – Le gall, S. (1981). Help seeking : An understudied problem-solving skill in children. *Developmental Review, 1,* 224-246.

What distributional information is useful and usable for language acquisition?

Padraic Monaghan (pjm21@york.ac.uk)
Department of Psychology, University of York
York, YO23 1ED, UK
Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853, USA

Abstract

Numerous theories of language acquisition have indicated that distributional information is extremely valuable for assisting the child to learn syntactic categories, yet these theories differ over the type of information that is proposed as useful in acquisition. Mintz (2003) has proposed that children utilize the previous word and the following word (AxB frames) for acquiring categories, whereas Monaghan, Chater, and Christiansen (submitted) have suggested that information about the previous word alone provides a rich source of data for categorization. In three modeling experiments we found that bigrams were better than fixed AxB frames for learning syntactic categories in a corpus of child-directed speech. However, presentation of the preceding and succeeding words when these can be processed separately resulted in better learning than presenting the preceding word alone, and also improved performance over presenting the previous two words.

Introduction

What sort of information does the child use to develop an understanding of their language? The rational analysis approach answers this question by assessing what sort of information is *useful* for learning the language. If a particular source of information proves to be rich and reliable then a computational system (of which the child is a very special case) will exploit it. The child learns a sense of syntactic categories early in language development. In order to understand speech and relate it to the world, the child must know which part of speech refers to an action, and which to objects, and which words modify relations between objects. "Look at the cow mooing" elicits many possibilities for relations between words and the world, for example, whether the animal in question is referred to by the word "cow", "look", or "mooing". Constraints *within* the language, restricting which words in the sentence can refer to objects, for example, greatly limit the number of possibilities for relating words to the world.

But what sort of information is useful for constructing syntactic categories? A variety of different types of information have been proposed as useful for categorization, including gestural, semantic, phonological, and distributional information. Combining more than one type of information has indicated improvements in categorization (Reali, Christiansen, & Monaghan, 2003), and it may indeed be the case that combining multiple sources is necessary for categorization to take place (Braine, 1987).

This paper focuses on distributional information as a cue for syntactic categorization, and questions what type of information is most useful and thus usable by the child. Theories of the use of distributional information in language acquisition have suggested different analyses of the context in which a word (category) occurs, but no empirical comparisons of these competing accounts have been made. We present a series of computational models that compare the extent to which accurate syntactic categorization of language directed to the child can be made on the basis of different sources of distributional information.

Sources of distributional information

Theories of distributional information in language acquisition have tended to focus on demonstrating that such information can contribute significantly toward categorization, rather than proposing that the particular implementation is psychologically realistic. Redington, Chater, and Finch (1998) produced context vectors based on the two preceding words and the two words following the target word from the CHILDES (MacWhinney, 2000) database of child-directed speech. The resulting vectors for the most frequent 1000 words in the database clustered together with a high correspondence to syntactic categories. Redington et al. (1998) also assessed vectors resulting from using different context words. They found that good results were also obtained for the one preceding and one following word, and also for the two preceding words, and for the two succeeding words (with better performance for preceding words than succeeding words). Yet, using only the immediately preceding word also resulted in good performance, though addition of richer contextual information improved performance.

An alternative approach is the proposal that particular sequences of words are useful for determining syntactic category. Fries (1952) produced a set of "frames" in which only words of a certain category could appear. For example, only a noun could appear in "The __ is/was/are good". Similarly, Maratsos and Chalkley (1980) proposed that there were local constraints on the occurrence of particular word categories, such as that only a verb can occur before the inflection *–ed*.

Mintz (2003) provided an empirical test of this local source of information, by analyzing corpora of child-directed speech for the occurrence of frames of the preceding and the succeeding words. We refer to these as AxB frames, where A and B are fixed, and x indicates the intervening word. For example, for the frame "you __ to", "go" and "have" both occur as "x" words in the frame.

Mintz selected the 45 most frequent frames involving the preceding and succeeding word, and then grouped the words that occurred within each of these frames. In the above example, "go" and "have" would be grouped together in the analysis. Accuracy was assessed by counting the number of times that words of the same category were grouped together, and dividing this by the number of pairings of all words within the groups. Completeness was determined by counting the number of pairings of words of the same category within the group, and dividing this by the number of pairings of words of the same category occurring in any of the groupings.

The 45 most frequent frames resulted in high accuracy but low completeness, indicating that these frequent AxB frames grouped together words of the same category, but that many words of the same category tended to occur in different groups. Relatedly, Mintz (2002) found that people categorized words together when they occurred in AxB frames in an artificial language learning task, and consequently claimed that such AxB frames were a source of distributional information that children used to acquire syntactic categories.

An alternative proposal is that a frame involving only the preceding word – an Ax frame – is required in order to produce effective categorization (e.g., Valian & Coulson, 1988). Monaghan, Chater, and Christiansen (submitted) found that categorizations of child-directed speech based on the association between the 20 most frequent preceding words and the target word resulted in accurate classification of words of different categories, but critically, also resulted in a large proportion of words being classified. Additionally, Monaghan et al. showed that, in an artificial language learning task, participants could group words on the basis of Ax frame information alone.

Both AxB and Ax frames can therefore be exploited in learning artificial languages, but which source of information is most useful to the child learning their language? AxB frames result in high accuracy, but low completeness, whereas Ax frames produce high completeness at the expense of some accuracy. Should a learning system select accuracy over completeness, or vice versa?

A comparison of different sources of distributional information requires that alternative methods are subjected to the same analyses. In addition, an empirical test of whether accuracy or completeness is a priority in acquisition is necessary. We now present a series of modeling experiments that test the extent to which different types of distributional information lead to successful categorization of words in child-directed language. Experiment 1 replicated Mintz's (2003) analysis of AxB frames in child-directed speech, and directly compared the resulting classification to an Ax analysis. Experiment 2 assessed whether a neural network model learned to categorise words more accurately on the basis of AxB information or Ax information alone. Finally, Experiment 3 tested a neural network model learning from AxB information when the relationship between A and x and B and x can also contribute separately towards categorization, and compared performance to a model with information about the two preceding words.

Experiment 1

Method

Corpus preparation From the CHILDES database, we selected a corpus of speech directed towards a child of age 0-2;6 years (anne01a-anne23b, Theakston, Lieven, Pine, & Rowland, 2001). This was one of the corpora used by Mintz (2003). We replaced all pauses and turn-taking with utterance boundary markers, and the resulting corpus contained 93,269 word tokens in 30,365 utterances (mean utterance length = 3.072 words). There were 2,760 word types, and the syntactic category for these words was taken from the CELEX database (Baayen, Pipenbrock, & Gulikers, 1995), according to the most frequent category usage for each word. Some interjections, alternative spellings, and proper nouns were hand-coded. There were 12 syntactic categories: noun, adjective, numeral, verb, article, pronoun, adverb, conjunction, preposition, interjection, wh-words (e.g., *why*, *who*), and proper noun.

Analysis In accordance with Mintz (2003), we selected the 45 most frequent AxB frames from the corpus, and determined the words that occurred in the x position within each frame. Each AxB frame thus resulted in a cluster of words. Accuracy and completeness were assessed in the same way as for Mintz (2003), described above. An additional method for assessing completeness was taken as the total number of word types that were classified in (at least) one frame.

For the Ax analysis, the 45 most frequent words were selected from the corpus, and co-occurrence with these frequent words formed the clusters in the bigram analysis. Accuracy and completeness were assessed in the same way as for the AxB co-occurrence analysis.

Results

As an example of the resulting classification, Table 1 shows a summary of the words that were classified into the 5 most frequent AxB and Ax frames. For these most frequent AxB frames, two frames clustered verbs together, and two clustered only pronouns. For the Ax classifications, the results are noisier, but have far higher numbers of words classified. The most frequent Ax frame – "the x" – classifies 623 nouns, and very few verbs, whereas the next most frequent Ax frame – "you x" – classifies 210 verbs, and only 26 nouns. The accuracy and completeness results are shown in Table 2, together with those from Mintz (2003)[1]. In parentheses are the random baseline values. We closely replicated Mintz's (2003) results indicating the high accuracy of the AxB frames, though, as noted in the

[1] Data are shown from Mintz's analysis of the anne corpus, with standard labeling and word-type analyses.

Table 1. Classifications based on the 5 most frequent Ax and AxB frames.

Ax						
Ax	noun	verb	pronoun	adjective	preposition	other
a	335	33	2	56	0	11
it	37	69	12	29	13	43
to	76	107	16	6	1	9
you	26	210	15	27	8	39
the	623	23	9	38	5	14
AxB						
Ax**B**	noun	verb	pronoun	adjective	preposition	other
do_think	0	0	1	0	0	0
do_want	0	0	6	0	0	0
are_going	0	0	5	0	0	0
what_you	0	10	0	0	1	0
you_to	0	19	2	1	1	1

Table 2. Completeness and accuracy of classifications for the Ax and the AxB co-occurrence models.

	CO-OCCURRENCE MODEL		
	MINTZ	**AX**	**AXB**
Accuracy	0.94 (0.41)	0.57 (0.22)	0.88 (0.26)
Completeness	0.09 (0.04)	0.07 (0.04)	0.06 (0.03)
Words classified	405, 14.7%	1930, 69.9%	394, 14.3%

Introduction, there was very low completeness for this classification. The Ax analysis also resulted in high accuracy, and slightly higher completeness according to Mintz's definition. However, a striking difference between the AxB and the Ax analyses is the overall number of words from the corpus that were categorized. Clustering based on bigrams resulted in a classification of almost 5 times as many words as the trigram analysis. The small differences in completeness between the two analyses is therefore misleading, as this only considered words that were clustered – in the AxB case, completeness was assessed over only a fraction of the corpus considered in the Ax analysis.

Discussion

We successfully replicated Mintz's (2003) demonstration that classifications of syntactic category based on occurrence within the most frequent AxB frames resulted in impressively high accuracy. However, our prediction that high accuracy could also be achieved by the smaller, less specific Ax frame was supported. The Ax analysis had the additional advantage of enabling a classification of far more words from the child's environment than was possible using AxB frames. There is a pay-off between accuracy and completeness: a specific context will result in high accuracy, but low completeness, whereas a general context will result in lower accuracy but high completeness.

This raises the question as to whether categorization is best based on information that renders highly reliable classifications of only a few words, or whether learning would benefit from using information that classifies a larger proportion of the words in the environment, but with the possibility that such classifications may contain more errors.

One way to test this issue is to train a neural network to base predictions of the syntactic category of words based on either AxB frames, or Ax frames. After training, the neural network model's error on the predicted classifications reflects the extent to which the given source of information is beneficial for learning the syntactic categories of the language. If the model trained on AxB frames has lower error then learning is more effective when based on high accuracy but low completeness, whereas if the model trained on the Ax frames has lower error then high completeness at the expense of high accuracy is a better source of information for learning.

We were concerned with how effective the frame is in predicting the category of the x word, so we trained the models to predict the category of x without entering the identity of the x word at the input. In addition, we did not preselect the frames that were input into the model: the entire corpus was used for training and not just the 45 most frequent frames, as we were interested in whether the model would be able to pick up which frames were useful for categorisation. From Mintz's (2003) analysis, it is not clear whether the AxB frames are to be interpreted as non-compositional, or whether the relationship between A and x and between x and B may also contribute to categorization. Experiment 2 tests the non-compositional interpretation, whereas Experiment 3 assesses the compositional version of the AxB frames.

Experiment 2

We trained two neural network models to learn to predict the category of the target (x) word using the same corpus of child-directed speech as in Experiment 1. We compared the learning of models that were given either Ax or AxB information. The AxB model was designed to test whether the AxB frame was useful for learning when the frame is interpreted as a whole, i.e., the "A" and the "B" do not contribute separately toward classification.

Architecture

Ax model The model was a feed-forward network with a set of input units fully-connected to a hidden layer, which was fully-connected to an output layer. The model is shown in Figure 1. Each unit in the input layer represented one word type in the child-directed speech corpus (so there were 2,760 input units), and there was also a unit representing the utterance boundary, in accordance with other connectionist models of syntax learning (e.g., Elman, 1990) that provide this additional information to the simulated child learner. There were 10 units in the hidden layer. The output layer contained units representing the syntactic category of the next word in the corpus. The model was trained on all Ax bigrams in the corpus, with the first word in the bigram occurring in the input layer, and the category of the second word in the bigram as the target at the output layer.

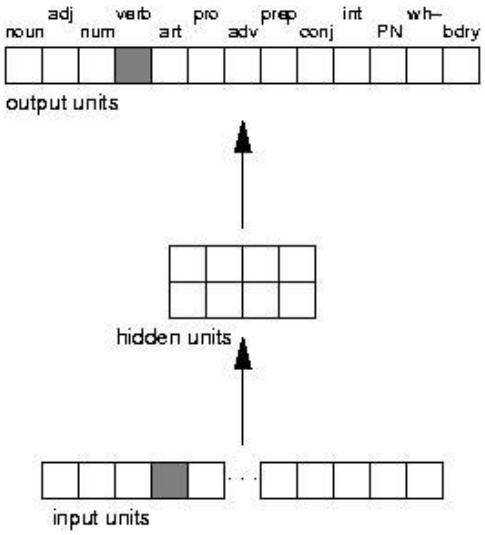


Figure 1. The feedforward neural network model of syntactic categorization. The active input unit represents either the A-word in the Ax model, or the AxB frame in the AxB model. The active output unit is the category of the x word, or the utterance boundary if x represents the end of the utterance. In the Figure, the output verb unit is active.

AxB model The AxB model was identical to that of the Ax model, except that in the input layer each unit represented one of the possible AxB frames. There were 36,607 such AxB frames, and so there were 36,607 input units in the model. The model was trained on all AxB frames in the corpus, with the A_B frame activating the appropriate unit in the input layer, and the syntactic category of the x word as the output layer target.

Training and testing

The models were trained using backpropagation with gradient descent with learning rate 0.01, and momentum 0.95. Before training, the weights between connections were randomized with mean 0 and standard deviation 0.1. We imposed a 0.1 error tolerance on the output units to prevent the development of very large weights on the connections. The models were trained on all Ax or AxB frames in the corpus, with each epoch being one pass through the corpus, and training was halted after 5 epochs, which was over 600,000 training events. As a baseline, we trained and tested the Ax model and the AxB model on a corpus where the frequency of words was maintained, but word-order was randomized. In the AxB randomized control model, there were 44,786 AxB frames and thus 44,786 input units in the model.

The models were tested after each epoch on the whole corpus, with the mean square error (MSE) across the output units taken as a measure of the ability of the model to learn to categorize words in the corpus on the basis of either the Ax or the AxB information. As an additional measure, we assessed whether the target unit – that is, the appropriate category of the x word – was the most highly activated for each pattern presentation.

Table 3. Percent correctly classified and MSE for the Ax and AxB models for each syntactic category in the corpus, with number of tokens (n) and *t*-test on MSE (all $p < 0.001$).

		% CORRECT		MSE		
CATEGORY	N	AX	AXB	AX	AXB	*t*
Nouns	12458	66.3	0	0.533	1.000	-116.316
Adjectives	4125	1.9	0	1.116	1.035	21.373
Numerals	1087	0	0	1.128	1.040	20.304
Verbs	23182	83.9	0	0.511	0.851	-145.602
Articles	7996	31.0	0	0.848	1.025	-52.371
Pronouns	18932	47.6	0	0.675	0.869	-71.369
Adverbs	5456	0	0	1.150	1.040	46.221
Prepositions	9491	31.3	0	0.865	1.016	-34.894
Conjunctions	1955	0	0	1.147	1.032	29.448
Interjections	3762	0	0	0.984	1.026	-24.608
Proper nouns	2104	0	0	1.149	1.032	28.642
Wh-words	3500	0	0	1.041	1.024	7.510
Boundary	30365	79.6	100	0.446	0.793	-147.391
TOTAL	**123634**	**52.4**	**22.9**	**0.680**	**0.911**	**-205.957**

Results

The Ax model performed better than the random baseline, MSE was 0.680 compared to 0.920, $t(247266) = -189.808$, $p < 0.001$. The model also classified more words correctly than the random baseline: 52.4% compared to 22.9%, $\chi^2 = 75{,}014.859$, $p < 0.001$.

The AxB model performed at a level similar to the random baseline. MSE was 0.911 which was slightly higher than the randomized version of 0.910, $t(247264) = 4.418$, $p < 0.001$. Classification was poor, with the model classifying all words as the utterance boundary, which was the single most frequent token in the input, This behavior was identical to the performance of the AxB model on the randomized corpus.

Table 3 shows the comparison between the Ax and the AxB models, for all words, and for each syntactic category. In terms of MSE, performance was better for the Ax model than the AxB model on all categories apart from adjectives, numerals, adverbs, conjunctions, proper nouns, and wh-words. However, performance was better for the large closed-class categories – pronouns and articles – and for nouns and verbs. Overall, the Ax model classified more words correctly than the AxB model, $\chi^2 = 75{,}014.011$, $p < 0.001$.

Discussion

The Ax model performed significantly better than chance in predicting the category of the x word from the preceding word. The AxB model performed at a chance level, and did not discriminate any word category. The better performance of the AxB model in terms of MSE on adjectives, numerals, adverbs, conjunctions, proper nouns and wh-words may have been due to a broader context serving these categories better: adverbs often occur after nouns in positions normally taken by verbs, and adjectives intervene between determiners and nouns. An enriched context would undoubtedly assist the categorization of these types. However, the better performance may merely have been due

to a lack of discrimination between any of the word types in the AxB model.

These simulations demonstrated that categorization of a large, entire corpus of child-directed speech was best achieved using information about the preceding word, rather than information about set frames comprised of the preceding and the following word. Greater coverage of the set of words, rather than greater accuracy in categorization, resulted in better performance.

The next experiment assessed whether a compositional treatment of the AxB frame may provide better information about the syntactic category of the target x word than the Ax frame alone, and compared it to a model with information about the two preceding words.

Experiment 3

We trained neural network models to learn to predict the category of the next word from the same corpus of child-directed speech as used in Experiments 1 and 2. We compared the learning of a model that was given information about the preceding and the following word in order to predict the category of the intervening word, but could operate on this information separately and combined. We call this the AxB-compositional (AxB-c) model. We also tested a model where information was given about the two preceding words: the ABx model. Note that these models embed the bigram information from the Ax model in the input. We predicted that both models would perform better than both the Ax model and the non-compositional AxB model from Experiment 2. We also predicted that the AxB-c model would outperform the ABx model, as proximity to the target word is most informative.

Architecture and training

The AxB-c model had the same architecture as the Ax model in Experiment 2, except that it had two banks of input units. In the first bank of units the unit corresponding to the A-word was activated, and in the second bank of units the B-word unit was activated. At the output layer, the model had to learn to predict the category of the x word. The same architecture was used for the ABx model, but it had as input the two words preceding the target word.

Training and testing was identical to that for the models in Experiment 2. Baselines for learning were determined by training and testing the models on the randomized corpus.

Results

For both models, performance was better than the random baseline in terms of accurate classifications and MSE. For the AxB-c model, accuracy was 69.4% (baseline 22.9%), χ^2 = 82422.148, $p < 0.001$, and MSE was 0.480 (baseline 0.920), $t(247266) = -329.487$, $p < 0.001$. For the ABx model, accuracy was 56.3% (22.9%), χ^2 = 60841.166, $p < 0.001$, and MSE was 0.628 (0.920), $t(247266) = -221.728$, $p < 0.001$.

As predicted, both the AxB-c and the ABx model performed with greater accuracy than the non-compositional AxB model from Experiment 2 for all syntactic categories: overall, $t(123633) < -300$, $p < 0.001$, for each individual syntactic category, all $t < -50$, all $p < 0.001$.

Table 4. Percent correctly classified and MSE for the AxB-c and ABx models. *T*-tests are computed on MSE (all $p < 0.001$, except † $p < 0.1$).

	% CORRECT		MSE		
CATEGORY	AXB-C	ABX	AXB-C	ABX	*t*
Nouns	73.7	68.0	0.408	0.509	-43.808
Adjectives	25.8	0	0.878	1.167	-44.306
Numerals	0	0	1.185	1.149	5.969
Verbs	85.4	86.6	0.289	0.466	-77.029
Articles	67.6	38.7	0.490	0.827	-72.861
Pronouns	80.5	53.5	0.361	0.585	-81.153
Adverbs	20.8	0	0.976	1.151	-33.207
Prepositions	59.0	37.8	0.592	0.807	-50.213
Conjunctions	0.5	0	1.140	1.148	$-1.409^{\dagger}$
Interjections	80.8	0	0.671	0.957	-71.643
Proper nouns	0.1	0	1.214	1.155	11.694
Wh-words	38.6	0	0.817	1.006	-23.613
Boundary	84.7	85.8	0.283	0.350	-26.769
TOTAL	**69.4**	**56.3**	**0.480**	**0.628**	**-147.470**

Compared to the Ax model in Experiment 2, the additional word information in the AxB-c and ABx models resulted in an increase in accurate classifications. For both models, classification was more accurate ($p < 0.001$), and resulted in lower error, both $t < -300$, $p < 0.001$. For the individual syntactic categories, the AxB-c and the ABx model performed better for all syntactic categories apart from numerals, all $t < -50$, all $p < 0.001$, though the difference for conjunctions was non-significant.

Table 4 compares the AxB-c model to the ABx model, indicating that accuracy was lower and MSE higher in the ABx model. The AxB-c model performed better on all syntactic categories apart from numerals and proper nouns.

Discussion

Providing decomposable information about the preceding and following word resulted in increased accuracy of performance in the model. The AxB-c model classified words of all syntactic categories better than the non-compositional AxB and the Ax models of Experiment 2. Accuracy across all the categories was high, though classifications of adjectives and adverbs was still inaccurate – these tended to be classified as nouns/pronouns and verbs, respectively. Adding information about the two preceding words also assisted in increasingly accurate classifications, though not to the same degree as providing the preceding and succeeding word.

General Discussion

Experiment 1 demonstrated, as predicted, that AxB information provides high accuracy at the expense of completeness, whereas Ax information results in slightly lower accuracy but much higher coverage of the language.

Experiment 2 tested the extent to which a computational model could utilize AxB frame information in categorizing the intervening word. The model trained on AxB frames performed at slightly below chance level, and well below the accuracy that could be achieved from categorizing on the basis of Ax information alone. The high completeness of Ax frames resulted in significantly better learning than the high accuracy but low-coverage of AxB information.

However, when the model is able to learn on the basis of AxB information when this information is compositional, i.e., the relationship between the preceding word and the target word and between the succeeding word and the target word can be computed separately, then a different picture emerges. The AxB-c model of Experiment 3 was more accurate than the Ax model of Experiment 2. Furthermore, this provided better classification results than the two preceding words (the ABx model), though this latter model also improved performance over a non-compositional AxB frame or just the single preceding word.

The simulations presented here suggest that learning is most effective when information about the preceding word and the succeeding word is available. However, this is only the case when the AxB frame is not computed as a whole. Learning must also be based in part on the relationship between A and x and between x and B. In the experiments presented in Mintz (2002), such a distinction is not made – the learning situation resembles that of the AxB-c model, where the participant has access not only to the AxB frame, but also to the Ax and the xB bigrams. Therefore, it is not yet possible to distinguish the contribution of bigram and trigram information in adult learning situations (though see Onnis et al., 2003).

The possibility remains that the requirement for category learning depends on establishing distinctions and similarities between only a few words in the language: it is not realistic or feasible to attempt to learn the whole language simultaneously. However, performance for the most frequent 100 words was poorer in the non-compositional AxB model than the Ax model, and even taking only those words that occurred in the most frequent 45 AxB frames resulted in poorer performance than for the 45 most frequent Ax frames.

The experiments presented in this paper require the models to learn pre-ordained syntactic categories. The task facing the child is more difficult: the child must also construct the categories. Yet, both tasks concern learning about which words co-occur. When the relationship between the occurrence of certain categories in particular distributional contexts is easy to learn then this demonstrates that the category itself is more clearly defined.

We have shown that AxB frames provide poor information about categorization unless this information is componential, such that Ax information is also available. We suggest that the distributional information that a neural network model finds most useful is more likely to be used by the child in acquiring syntactic categories.

Acknowledgments

This research was supported in part by a Human Frontiers Science Program Grant (RGP0177/2001-B).

References

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Braine, M.D.S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp.65-87). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Fries, C.C. (1952). *The Structure of English: An Introduction to the Construction of English Sentences.* New York: Harcourt, Brace & Co.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M.P. & Chalkley, M.A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K.E. Nelson (Ed.), *Children's Language* Volume 2, pp.127-214. New York: Gardner Press.

Mintz, T.H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30*, 678-686.

Mintz, T.H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91-117.

Monaghan, P., Chater, N., & Christiansen, M.H. (submitted). The differential contribution of phonological and distributional cues in grammatical categorization.

Onnis, L., Christiansen, M.H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. *Proceedings of the 25th Cognitive Science Society Conference* (pp. 887-891). Mahwah, NJ: Lawrence Erlbaum.

Reali, F., Christiansen, M.H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: Scaling-up the connectionist approach to multiple-cue integration. *Proceedings of the 25th Cognitive Science Society Conference* (pp. 970-975). Mahwah, NJ: Lawrence Erlbaum.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Theakston, A.L., Lieven, E.V.M., Pine, J.M., & Rowland, C.F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, *28*, 127-152.

Valian, V. & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*, 71-86.

Analogical retrieval from everyday experience: Analysis based on the MAC/FAC

Junya Morita (morita@cog.human.nagoya-u.ac.jp)
Graduate School of Human Informatics
Nagoya University, Furo-cho, Chigusa-ku, Nagoya City, Japan

Kazuhisa Miwa (miwa@cog.human.nagoya-u.ac.jp)
Graduate School of Information Science
Nagoya University, Furo-cho, Chigusa-ku, Nagoya City, Japan

Abstract

In order to investigate how analogs are retrieved from everyday experience, we conducted an experiment in which subjects were not presented with analogies by an experimenter, but presented with only one story as a retrieval cue. In our experiment, subjects were divided into four groups varying the cue stories, which were manipulated by their surface and structural features. In all the groups, the subjects were asked to report the cases that came to mind when they read the cue story. After retrieving, the subjects rated the inferential soundness (goodness as analogy) of each retrieved case. We computed similarities between each retrieved case and the cue stories, using a computational model MAC/FAC ("many are called but few are chosen"), which was developed for simulating two stages of analogy making (Forbus, Gentner, & Law, 1995). The results showed that (1) the retrieved cases were similar to the presented story in the surface features rather than in the structural features and (2) the structural similarity between the retrieved cases and the presented story increased with the rated scores of inferential soundness. These results confirmed that, as the results of prior controlled experiments suggested, the surface similarity guides the retrieval of cases and the structural similarity guides the evaluation of the cases.

Introduction

Analogy making is a core component of higher-order cognition, such as problem solving (Gick and Holyoak, 1980), decision-making (Markman and Moreau, 2001), and creative generation (Smith, Word, and Schumacher, 1993). In the past two decades, many researchers have conducted controlled experiments and gained reliable findings on analogy making. However, there have been only a few studies to verify the findings in less controlled environments. Our goal here is to replicate previous findings on analogy making in extended laboratory settings.

Prior to presenting our experiment, we briefly review a framework developed in the area of analogy research. First, in analogy research, a representation of a novel situation is called a target, and a past case that is similar to the target is called a base. The process of analogy making is comprised of two main components: the retrieval of the base and the mapping from the base to the target.

It has been pointed out that the analogy process is guided by similarities between the base and the target. Using propositional representations (predicate-argument formalism), Gentner (1983) distinguished three types of correspondence between the base and the target.

- Correspondence of *attributes*: e.g., The sun is round and yellow → The orange is round and yellow [sun (round) sun (yellow) → orange (round) orange (yellow)]
- Correspondence of *first-order relations*: e.g., The planets revolve around the sun. → The electrons revolve around the atom [revolve-around (planet, solar) → revolve-around (electron, atom)]
- Correspondence of *higher-order relations*: e.g., Because the sun attracts the planets, the planets revolve around the sun. → Because the atom attracts the electrons, the electrons revolve around the atom [cause (attract (solar, planet), revolve-around (planet, solar)) → cause (attract (atom, electron), revolve-around (electron, atom))]

The above discrimination was based on the types of predicates. The attribute is a predicate type that takes only a single argument. On the other hand, the first-order and higher-order relations are predicate types that take multiple arguments. There is no depth in the former, but there is in the later. Based on this discrimination, in analogy research, further discrimination of similarity has been proposed: i.e., surface similarity and structural similarity. The degree of surface similarity is roughly defined as the number of attributes shared between the base and the target. Contrary to the surface similarity, the degree of structural similarity is defined as the depth of structural mapping from the base to the target; so correspondence of higher-order relations is deeper than that of first-order relations (Gentner, 1983).

It has been demonstrated that two types of similarities take different roles in the analogy process (Holyoak, & Koh, 1987; Gentner, Rattermann, & Forbus, 1993; Wharton, Holyoak, Downing, Lange, Wickens, & Melz, 1994). For example, Gentner, Rattermann, & Forbus (1993) conducted experiments in which subjects learned several stories and then retrieved the learned stories when they read new cue stories. The cue stories were manipulated by two factors: surface and structural similarities to the learned stories. As a result, the subjects retrieved more often the surface similar stories than the structurally similar stories. However, once the subjects were presented with the learned stories with the cue stories, they rated the inferential soundness (goodness as analogy) of the structurally similar stories higher than that of the surface similar stories.

To explain these results, Forbus, Gentner, & Law (1995) proposed a computational model, called MAC/FAC ("many are called but few are chosen"), which simulates two stages of the analogy process. In the first stage of MAC/FAC, several potential bases are retrieved from a memory pool, computing the dot product of the target's content vector (CVector), which is a simple list of the predicates contained in the propositional representation, with the CVector of each case in the memory pool. In the next stage, the cases retrieved at the initial stage are further evaluated by using the Structure-mapping Engine (SME), which computes structural alignment and evaluation of the match between each set of cases and the target (Falkenhainer, Forbus, & Gentner, 1989). Finally, MAC/FAC selects the cases that have high structural evaluation scores (SES), which indicate the degree of depth and breadth of the common structure. In brief, MAC/FAC can discriminate the initial stage of retrieval that is guided by surface similarity from the evaluation stage that is guided by structural similarity. In the past, similar discrimination has been employed in many other models of analogy (Thagard, Holyoak, Nelson, & Gochfeld, 1990; Hummel & Holyoak, 1997).

Recently, however, limitations on the above finding have been pointed out. The limitations are derived from the fact that many analogy researchers have only dealt with cases created by the researchers themselves. In other words, the experiments have been conducted in closed laboratories, where the subjects retrieved cases created by researchers in advance. In real-world situations, it is impossible to predict the cases that will be retrieved or used. In real-world situations, the analogy is made from individuals' everyday experience. Therefore to extend the findings to realistic problems, it is necessary to investigate analogy making using cases that the subjects learn in their own everyday life.

From this viewpoint, Blanchette & Dunbar (2000) conducted experiments that examined analogy making in situations where the subjects were not provided analogies guided by an experimenter. In their experiments, the subjects were asked to generate analogies to the zero-deficit problem - the deficit that Canadian governments had to cut. The results showed that the subjects generated few analogies that have surface features in common with the target (the zero-deficit problem), but generated many analogies that shared deep structures with the target. Further, being asked to select the best analogy from the generated analogies, the subjects selected analogies that had deeper structural correspondence than the others.

These results indicate the strong effect of structural similarity on both the retrieval and the evaluation stages, contradicting the previous studies that showed different similarities involved in the two stages. Based on the results, Blanchette & Dunbar claimed that surface similarity has little effect on analogy making in situations where subjects use their own analogies.

Although we agree on the importance of their approach, which aims to combine naturalistic settings and controlled experiments (Dunbar & Blanchette, 2001), we think that further investigation is needed for their claim. Thus, we reexamined the similarity effects on the analogy process in a situation where the subjects retrieved cases that were learned in their own everyday life. The method of our experiment is similar to that of Blanchette & Dunbar, but there are three important differences, as follows.

First, we modified the instruction in which the subjects were asked to "generate analogies". Because many researchers argued that the term "analogy" commonly implies "the cases that have low surface similarity and high structural similarity" (e.g., Gentner, 1983), there exists a possible other account for Blanchette & Dunbar's results: The subjects might actually be reminded of surface similar cases, but would not report those cases. In order to test this possibility, we did not include the term "analogy" in our instruction.

Second, we constructed several controlled experimental conditions. Blanchette & Dunbar conducted the experiments without clear manipulations using surface or structural similarities. In such an experiment, it is difficult to exclude possible conjectured factors, such as the types of cases that subjects hold, or the frequency of using these cases in their everyday life. To control these factors, we divided subjects into several experimental groups and presented the targets whose surface and structural features were systematically changed.

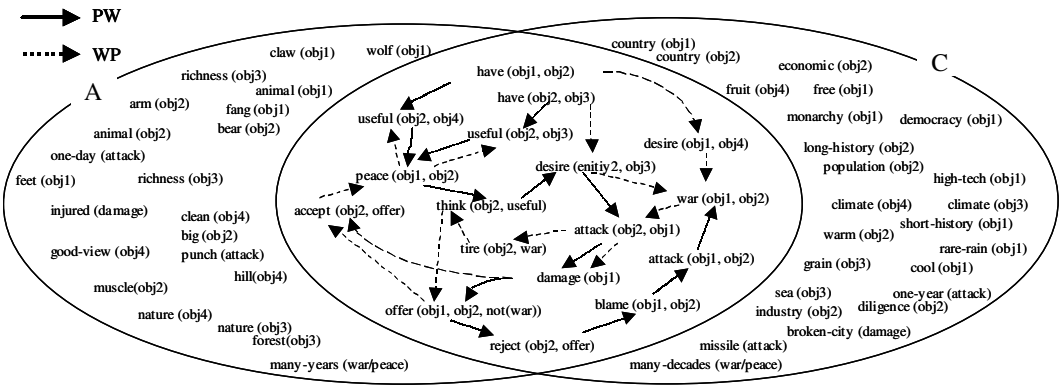
Third, we analyzed the data quantitatively. In Blanchette & Dunbar's study, the generated analogies were analyzed by categorizing their surface/structural features and comparing frequencies of categories. However, they did not show how much the generated analogies shared surface/structural features with the target. For quantitative analysis, we computed similarity scores for each retrieved case based on the algorithm assumed in the MAC/FAC (Forbus, Gentner, & Law, 1995).

Method

Materials

The experiment was conducted to investigate the effects of similarities on the analogy process in a situation where subjects retrieved cases that were learned in their own everyday life. In our experiment, the subjects were presented with a cue story and then were asked to report the cases that came to mind while they read the cue story. The cue story consisted of about 600 Japanese characters. In this paper, we call these stories the target stories.

The texts of the target stories were manipulated with their surface and structural features. The subjects were divided into four experimental groups varying target stories (the between-subjects factor). As the surface features, a set of attributes related to *animals* (A) and a set of attributes related to *countries* (C) were chosen. As the structural features, a story whose plot is a transition from *peace to war* (PW) and a story whose plot is a transition from *war to peace* (WP) were created. Combining the surface and structural features, four types of



CVector between targets

	A	C
A	29	0
C		29

SES between targets

	PW	WP
PW	28.8	12.2
WP		27.9

Figure 1: Propositions contained in the target stories.

target stories were prepared (A/PW, A/WP, C/PW, and C/WP).

Figure 1 shows the propositions converted from the texts in the target stories. Each of them was included in either set A or set C. Two complements $[(A \cap \bar{C})$ and $(\bar{A} \cap C)]$ contain attributes of objects and an intersection $(A \cap C)$ includes first-order relations between two objects. Each of the first-order relations was connected by two types of higher-order relations (PW/WP) represented by two types of arrows (solid/dotted).

Figure 1 also shows similarity scores calculated based on the algorithms of the MAC/FAC. The CVector, indicating surface similarity, was computed as a dot product of each pair of surface features (A vs. A, A vs. C, and C vs. C). In our study, the CVector was represented as a list of attributes, not containing first-order and higher-order relations. The SES, indicating structural similarity, was computed by inputting each pair of relational structures (PW vs. PW, PW vs. WP, and WP vs. WP) into the SME model. From the several matching rules of the SME package, we chose analogy rules that do not compute the match of attributes.[1]

In order to verify the above manipulation, we conducted a preliminary experiment. The subjects (n = 8) were presented with four target stories and then rated the inferential soundness of each pair of the target stories on a 1 ("low") – 5 ("high") scale. Similar to Gentner, Rattermann, & Forbus (1993), soundness was explained as "the degree to which inferences from one story would hold for the other".

[1]Our way of computing similarity was slightly modified from Forbus's method. Forbus treated the CVector as a list of all types of predicates including relations, and computed the SES using the literal-similarity rules that mapped all types of predicates including attributes (Forbus, Gentner, & Law, 1995). In our study, for clear discrimination between surface (attributes) and structural similarity (relations), we chose the above method.

The results showed that the manipulation is consistent with human feelings of soundness. Seven of eight subjects judged the structurally similar pairs (A/PW vs. C/PW and A/WP vs. C/WP) as having higher inferential soundness than the other pairs (A/PW vs. A/WP, CPW vs. C/WP, A/PW vs. C/WP, and A/WP vs. CPW).

Participants

Thirty-three undergraduate and graduate students participated in the experiment. They were divided into four groups: a group presented with A/PW (n = 8), a group with A/WP (n = 9), a group with C/PW (n = 8), and a group with C/WP (n = 8).

Procedure

The subjects participated in the experiment individually or in groups of two to four. The experiment was divided into the following three phases.

Retrieval phase In the first phase, the subjects reported the cases of which the target story reminded them. In explaining the task, we avoided using terms like "analogy" or "analogous". The subjects were simply told that "while reading the presented story, you should write out any cases that come to mind". After the instruction, the subjects were presented with one of the four targets and then they wrote down any reminded cases. This phase continued for twenty minutes.

Evaluation phase Following completion of the retrieval phase, the subjects were given a soundness rating task. The subjects rated the soundness of the match between each retrieved case and the presented target on a 1 – 5 scale.

Subjects' descriptions	Converted propositions
	((animal tiger1) :name prop3)
The story about two tigers.	((animal tiger1) :name animal2)
In a forest, two tigers lived.	((animal animal1) :name animal3)
Each of them has a turf. And	((have tiger1 turf1) :name have1)
they battled each other for	((have tiger2 turf2) :name have2)
the turf. One day, an animal	((desire tiger1 turf2) :name desire1)
that lived in the forest	((desire tiger2 turf1) :name desire2)
persuaded one of the tigers to	((war tiger1 tiger2) :name war1)
stop fighting. After this	((and desire1 desire2) :name and1)
persuasion, the relationship	((cause and1 war1) :name cause1)
between the two tigers	((not war1) :name not1)
became peaceful.	((offer animal1 tiger1 not1) :name off)
	((accept tiger1 off) :name accept1)
CVector (A) = 8	((cause offer1 accept1) :name cause2)
CVector (C) = 0	((cause war1 off) :name cause3)
SES (PW) = 7.59	((peace tiger1 tiger2) :name peace1)
SES (WP) = 11.75	((cause accept1 peace1) :name cause4)
	((many-tree turf1) :name prop1)
	((many-tree turf2) :name prop2)
The gallic war.	
In order to expand the	((country garia) :name country1)
national land, the Roman	((country other) :name country2)
Empire kept attacking other	((monarchy garia) :name prop1)
countries.	((have other land) :name have1)
	((attack gallia other) :name attack1)
CVector (A) = 0	((desire gallia land) :name desire2)
CVector (C) = 5	((war gallia other) :name war1)
SES (PW) = 6.18	((cause desire2 attack1) :name cause1)
SES (WP) = 6.10	((cause attack1 war1) :name cause2)

Figure 2: Examples of subjects' descriptions and propositions.

Explanation phase Finally, the subjects were asked to explain the retrieved cases in as much detail as possible.

Coding

The retrieved cases were coded using propositional representations. The subjects' descriptions were segmented by the appearance of a predicate. Then a coder judged whether each segmented sentence could be represented as a proposition by using predicates contained in the targets (the predicates in Figure 1). If possible, a proposition would be constructed by complementing for proper arguments. Examples of the coding are shown in Figure 2.

Results and Discussion

The total number of cases retrieved by the subjects was 266. There was no significant difference on the number of cases among the four experimental groups [$\chi^2(3) = 6.15, ns.$]. Thus, we treated each retrieved case as an individual datum for statistical tests.

In order to examine the relationship between the surface/structural similarities and the retrieval/evaluation stages of analogy making, we tested (1) whether the retrieved cases were similar to the presented target in the surface/structural features, and (2) whether the surface/structural similarity between the retrieved cases and the presented target increased with the degree of soundness rating.

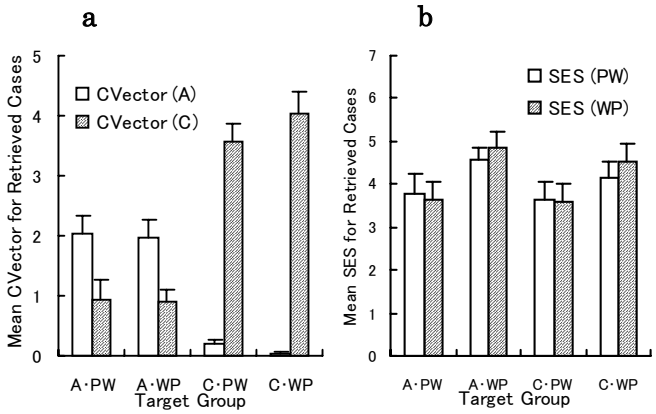


Figure 3: (a) Mean CVector for four groups. (b) Mean SES for four groups. *Note.* Error bars represent one standard error of mean.

1. Effects of Similarities on Retrieval

For investigation of the effects of similarities on retrieval, four types of similarity scores were computed based on the algorithm assumed in the MAC/FAC.

- *CVector (A)* was computed as the dot product between each retrieved case and the surface feature A ($A \cap \bar{C}$ in Figure 1).
- *CVector (C)* was computed as the dot product between each retrieved case and the surface feature C ($\bar{A} \cap C$) in Figure 1).
- *SES (PW)* was computed by inputting each retrieved case and the structural feature PW (the solid lines in Figure 1) into the SME model.
- *SES (WP)* was computed by inputting each retrieved case and the structural feature WP (the dotted lines in Figure 1) into the SME model.

We conducted two ANOVAs to investigate interaction between the above similarity scores and the experimental groups. If the surface/structure features affected the case retrieval, the retrieved cases would be similar to the presented target rather than the targets that were not presented for each group.

Effects of Surface Similarity on Retrieval Figure 3a shows the mean CVector for each group. A 2 × 2 × 2 surface features of targets (between) × structural features of targets (between) × types of CVector (within) ANOVA revealed significant interaction between the surface features of targets and the types of CVectors [$F(1, 262) = 118.21, p < .05$], indicating CVector (A) was higher than CVector (C) in the group A [$F(1, 262) = 14.22, p < .05$], and CVector (C) was higher than CVector (A) in the group C [$F(1, 262) = 134.67, p < .05$].

Both in group A and group C, the retrieved cases were more similar to the target that was presented for each group than the targets that were not presented. A strong effect of surface similarity on retrieval contradicts the results of Blanchette & Dunbar (2000), but is consistent with the findings of the previous controlled experiments (Holyoak, & Koh, 1987; Gentner, Rattermann, & Forbus, 1993; Wharton et. al, 1994).

Effects of Structure Similarity on Retrieval Figure 3b shows the mean SES for each group. A 2 × 2 × 2 surface features of targets (between) × structural features of targets (between) × types of SES (within) ANOVA revealed significant interaction between the structural features of the target and the types of SES $[F(1, 262) = 8.01, p < .05]$. However, simple main effects were significant only in group WP $[F(1, 262) = 7.50, p < .05]$. There was no significant difference of types of SES in group PW $[F(1, 262) = 1.60, ns.]$.

These results suggest that structural similarity has more restricted effects on retrieval than surface similarity. Again, this result contradicts the study by Blanchette & Dunbar, but is consistent with the findings of the previous controlled experiments (Holyoak, & Koh, 1987; Gentner, Rattermann, & Forbus, 1993; Wharton et. al, 1994).

2. Effects of Similarities on Evaluation

In order to investigate the effects of structural similarity on the evaluation stage, we treated the subject groups as counterbalance conditions, and reduced the number of factors for the ANOVA. Therefore, we investigated four types of similarity scores as follows:

- *CVector (presented)* was computed by combining CVector (A) in group A and CVector (C) in group C. This score indicates the degree of surface similarity, meaning how many attributes were shared between each retrieved case and the target that was presented for the subjects.

- *CVector (not presented)* was computed by combining CVector (A) in group C and CVector (C) in group A. This score indicates how many attributes were shared between each retrieved case and the target that was not presented for the subjects. Because two types of surface features (A and C) share no attributes, this score indicates surface dissimilarity.

- *SES (presented)* was computed by combining SES (PW) in group PW and SES (WP) in group WP. This score indicates the depth of structural mapping from each retrieved case to the presented target. Thus, this score indicates structural similarity that reflects higher-order relations.

- *SES (not presented)* was computed by combining SES (PW) in group WP and SES (WP) in group PW. This score indicates the depth of structural mapping from each retrieved case to the target that was not presented for the subjects. Since two structural features (PW and WP) share no higher-order relations but only first-order relations, this score indicates the degree of overlap of the first-order relations.

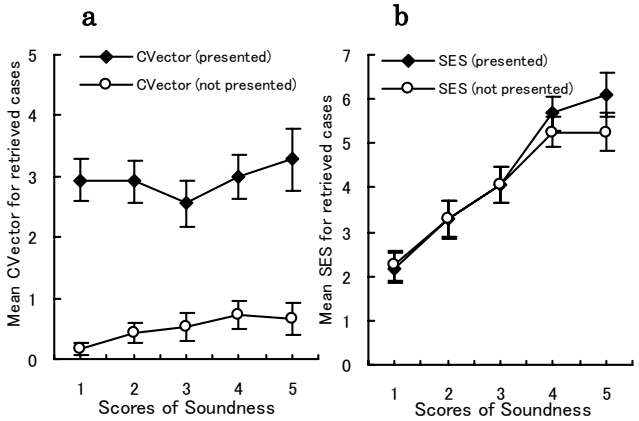


Figure 4: (a) Mean CVector for rating scores of soundness. (b) Mean SES for rating scores of soundness. *Note.* Error bars represent one standard error of mean.

We computed two ANOVAs for the retrieved cases by using the above similarity scores. Each ANOVA tested whether the similarity scores increased with the rated scores of soundness (1 – 5). If the structural similarity had a strong effect and the surface similarity had only a little effect in the evaluation stage, as suggested by the previous studies (Gentner, Rattermann, & Forbus, 1993; Blanchette & Dunbar, 2000), the CVector (presented/not presented) would not increase with the soundness rating, but the SES (presented/not presented) would increase with the soundness rating. Further, the SES (presented), which reflects the higher-order relations, would be related to the soundness ratings more than the SES (not presented), which reflects only the first-order relations.

Effects of Surface Similarities on Evaluation Figure 4a shows the mean CVector for each score of soundness (1 – 5). A 5 × 2 soundness scores (between) × CVector types (within) ANOVA detected a significant main effect of CVector types $[F(1, 261) = 121.17, p < .05.]$. However, a main effect of soundness $[F(4, 261) = 1.91, ns.]$ and an interaction between CVector types and soundness scores $[F(4, 261) = 0.40, ns.])$ were not significant. These results indicate an advantage of surface similarity over surface dissimilarity regardless of soundness ratings. Thus, as the previous studies indicated, the results suggest that there is no effect of surface similarity or dissimilarity on the evaluation stage.

Effects of Structural Similarity on Evaluation Figure 4b shows the mean SES for each score of soundness (1 – 5). A 5 × 2 soundness scores (between) × SES types (within) ANOVA revealed a significant interaction between soundness scores and SES types

$[F(4, 261) = 7.52, p < .05]$. Simple main effects of soundness scores were significant for both SES (presented) $[F(4, 261) = 15.46, p < .05]$ and SES (not presented) $[F(4, 261) = 11.09, p < .05]$. Further it was confirmed that SES (presented) was higher than SES (not presented) at the soundness scores 4 $[F(1, 261) = 8.76, p < .05]$, and 5 $[F(1, 261) = 35.73, p < .05]$.

Significant main effects of soundness on both the SES (presented) and the SES (not presented) indicate that the subjects' evaluation was positively correlated to the degree of commonalities in the first-order relations. The fact that the SES (presented) was higher than the SES (not presented) in the cases in which the subjects rated high soundness (rate scores 4 and 5) implies that commonalities in the higher-order relations are more strongly related to soundness than mere first-order relations. In summary, these results are consistent with prior studies (Gentner, Rattermann, & Forbus, 1993; Blanchette & Dunbar, 2000) showing the strong effects of structural similarity on evaluation.

General Discussion

Influence of Similarities on Analogy Process

In this study, we investigated types of similarities influencing the analogy process, conducting an experiment in which the subjects retrieved cases that they had learned in their own everyday life. As with the results of previous controlled experiments, our results also suggest that different types of similarities are responsible for the retrieval and evaluation stages of the analogy process. The results were different from those of Blanchette & Dunbar (2000), which showed little effect of surface similarity on the retrieval phase. The difference between our results and Blanchette & Dunbar's results could be explained by differences in the instructions. The subjects who participated in the experiments of Blanchette & Dunbar may have filtered out the surface similar cases because they were instructed to "generate analogies". Since there exist other differences between our experiment and Blanchette & Dunbar's experiments, such as the reality of the tasks used and the method of analysis, we must conduct further experiments controlling for these factors.

In addtion, our study obtained results indicating a strong effects of structural similarity on the evaluation stage. The results were clearer than Blanchette & Dunbar's results. Blanchette & Dunbar's analysis was based on counting elements shared with the base and the target, without any consideration of relational structures. In contrast, our analysis was based on a computational model that computes structural alignment and structural evaluation. Our analysis showed that the degree of shared attributes did not increase with the soundness rating, whereas the degree of shared relations increased with the soundness rating. Further, sharing higher-order relations made the relation to the soundness rating stronger. These results are important for the extension of the systematicity principle, proposed by Gentner (1983), which predicted that deeper structural mapping would be preferred in an analogical inference.

Investigation based on a Computational Model

The above results imply the benefit of using a computational model for analysis of psychological data. In the past, few studies have used a computational model to analyze data obtained from psychological experiments. However, without a sufficient computational model, it would be impossible to investigate complex cognitive conceptual products such as structural similarity.

Recently, in the community of cognitive science, the connection between theory and experimental data has been stressed. Usage of a computational model for analysis, demonstrated in this paper, could open a new way of directly licensing these two entities that play the most important role in cognitive scientific studies.

References

Blanchette, I., & Dunbar, K., (2000). How analogies are generated: The role of structural and superficial similarity. *Memory & Cognition*, *28*, 108–124.

Dunbar, K., & Blanchette, I, (2001). The in vivo/in vitro approach to cognition: the case of analogy. *Trends in cognitive science*, *5*, 334–339.

Falkenhainer, B., Forbus, K. D., & Gentner, D., (1989). The Structure-Mapping Engine: Algorithm and Example. *Artificial Intelligence*, *41*, 1–63.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

Gentner, D. (1983). Structure-Mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., Rattermann, M., & Forbus, K. D. (1993). The role of similarity in transfer: Separating retrievability for inferential soundness. *Cognitive Psychology*, *25*, 524–575.

Holyoak, K. J., & Koh, K. (1987). Surface and Structural Similarity in Analogical Transfer. *Memory & Cognition*, *15*, 332–340.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Markman, A. B., & Moreau, C. P. (2001). Analogy and analogical comparison in choice. In D. Gentner, K. Holyoak & B. N. Kokinov (Eds.), *Analogical Mind: Perspectives from cognitive science.* Cambridge, MA: The MIT press

Smith, S. M., Ward, T. B., & Schumacher, J. S. (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition*, *21*, 837–845.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, *46*, 259-310.

Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. E., & Melz, E. R., (1994). Below the surface: Analogical similarity and retrieval competition in reminding *Cognitive Psychology*, *26*, 64–101.

Using Testing to Enhance Learning: A Comparison of Two Hypotheses

Michael C. Mozer
Department of Computer Science & Institute of Cognitive Science
University of Colorado
Boulder, CO 80309 USA

Michael Howe
Department of Computer Science & Institute of Cognitive Science
University of Colorado
Boulder, CO 80309 USA

Harold Pashler
Department of Psychology
University of California at San Diego
La Jolla, CA 92093 USA

Abstract

Students learning facts such as foreign language vocabulary often rely on a self-testing procedure in which they cue themselves with the English word and try to recall the foreign language target, instead of simply memorizing cue-target pairs. The value of this strategy has been empirically verified by a long history of research, yet existing computational models of human learning do not address the enhancing-learning-through-testing phenomenon. Using a simple, well studied model—a feedforward neural net with no hidden units—we propose two different hypotheses for characterizing the phenomenon. Hypothesis 1 is that self-testing generates a target which is used for additional training. Hypothesis 2 is that self-testing produces a more reliable error signal for training than rote memorization. Through simulation studies, we find that hypothesis 2 readily explains the phenomenon whereas hypothesis 1 does not. Further, hypothesis 2 makes predictions worthy of further empirical study, and can be viewed as a natural consequence of temporal difference learning.

When learning foreign language vocabulary and other facts, students often study using index cards that have an English vocabulary word (or *cue*) on one side and a foreign language vocabulary word (or *target*) on the other. The intuition is that by testing oneself, the associations are better learned and retained.

This intuition has been supported by a long history of empirical demonstrations (e.g., Izawa, 1966; Young, 1971). For example, Bartlett and Tulving (1974) asked participants to learn a list of paired associates (the *study phase*), and later tested retention of the pairs using free recall or recognition (the *final test*). Before the final test, subjects were given a cued-recall test (a *self test*) of some of the paired associates. Retention was better on the final test for those items that received the self test.

In this paradigm, it is unclear whether the benefit of the self test is attributable to attempting retrieval per se, or to the fact that successful retrieval of an associate also results in a re-presentation of the pair—an additional training trial.

An obvious strategy for examining the effect of retrieval is to conduct and experiment with, in addition to the initial study phase and the final test, an intervening phase in which participants are given either a self test or an experiment-provided re-presentation of the paired associate (which we'll refer to as *study only*). In this paradigm, the outcome is ambiguous (Carrier & Pashler, 1992): self testing outperforms study in some experiments (e.g. Hogan & Kintsch, 1971), but not others (e.g., McDaniel & Masson, 1985). One explanation for the inconsistency is that the rate of retrieval success on self test trials varies among experiments, and the mechanisms of learning are likely to be dependent on retrieval success. The experiments have other problems, including different amounts of time for study-only and self-test conditions, and failure to control the time spent on individual items (Carrier & Pashler, 1992).

To overcome these metholological difficulties, Carrier and Pashler (1992) compared a *study-only* or *SO* condition in which each cue-target pair was presented for ten seconds to a *test/study* or *TS* condition in which the cue was presented alone for five seconds and then the target appeared for the final five seconds. In TS trials, participants were supposed to use the cue to retrieve the target, but even if retrieval failed, the trial still had value due to the presentation of cue and target together for five seconds. Consequently, the dependence on retrieval success rate is minimized. Also, the paradigm matches the total time per item in SO and TS conditions. If anything, self testing is at a disadvantage because the total viewing time for cue plus target was lower.

In Experiments 1 and 2 of Carrier and Pashler, 40 cue-target pairs were used, half each assigned to the SO and TS conditions. The experiment began with a study only phase in which participants viewed each of the 40 pairs once for ten seconds. Then two more passes were made through the pairs, presented in the manner designated for that pair—SO or TS. For both conditions, participants were instructed to say aloud the target. In the TS condition, this instruction required that participants recall the target, or if they failed to recall, to wait until the target appeared. Following the three presentations of each pair, a final test phase evaluated performance in the two conditions via cued recall.

In Experiment 1, the cues were consonant-vowel-consonant trigrams and the targets were two digit numbers. For the sake of ecological validity, Experiment 2 used a language learning task with English language word cues and the corresponding Siberian Eskimo Yupik language translation targets. Table 1 shows the percentage of error responses. In both Experiments 1 and 2, performance was better in the TS condition than

TABLE 1. Performance in Enhancing-Learning-Through-Testing Experiments

	Human Data (% Error)		Simulation (Mean Squared Error)	
	Study Only	**Test Then Study**	**Study Only**	**Test Then Study**
Carrier & Pashler, Expt. 1	42.0%	36.0%	.389	.321
Carrier & Pashler, Expt. 2	43.0%	36.0%		
Carrier & Pashler, Expt. 3	40.0%	32.7%	.356	.258

~~the SO~~ condition. These results indicate roughly 10% fewer errors with testing, therefore, having to retrieve the target is more effective than simply studying the target, when all else is controlled for.

Carrier and Pashler conducted a further experiment to rule out an alternative explanation of their results. In Experiments 1 and 2, participants may have used the first retrieval attempt in the TS condition to determine which items were difficult, and then increased their encoding effort for the difficult items on the second retrieval attempt, thereby learning the items better. Experiment 3 ruled out this explanation by giving participants only a single pass through the items in either the TS or SO conditions, following two passes through the items as study-only trials. The total number of items was reduced to 30. The results were similar to Experiments 1 and 2 (see Table 1), suggesting that the effect of attempting retrieval on later retention does not depend on strategic allocation of encoding effort.

Mechanism Underlying Enhanced Learning Through Testing

Why does testing oneself—i.e., attempting to retrieve a target from memory—have beneficial effects for later retention, above and beyond the effects due to mere study? A variety of explanations have been proposed for the self-testing benefit.

- Landauer and Bjork (1978) considered that retrieval attempts provide a general sort of practice or context that boosts performance at a future time. However, this account predicts that the benefits would not be item specific, i.e., SO and TS items would benefit equally in an experiment where the item types were mixed within subject.
- Mandler (1979) suggested that cued recall might strengthen the structural, integrative information about a cue-target pair. Cooper and Monk (1976) proposed that retrieval requires neural activity that consolidates the representation of the target in memory. However, both of these accounts do not provide a strong explanation for why TS should be better than SO, because both conditions involve simultaneous activation of cue and target.
- Bjork (1975) hypothesized that the act of retrieval may strengthen existing retrieval routes to the target representation, or may create new routes. Although interesting and consistent with the data, it is unclear what this hypothesis corresponds to in computational terms, and seems as if it might require novel, custom learning mechanisms.

This paper explores two alternative hypotheses concerning the enhancement of learning through testing, and we evaluate their plausibility via simulation studies. In proposing hypotheses, our aim was to determine whether an existing, well-accepted model could explain the basic phenomenon without requiring additional assumptions. A model is not convincing if two novel assumptions are needed to explain two data points. Further, an existing model is already constrained and therefore has the power to make strong predictions, which can guide the design of behavioral experiments.

Our hypotheses lie within the framework of neural network models. We explore the simplest architecture that might be capable of explaining the phenomenon: an associative network consisting of a pool of n_I input units fully connected to a pool of n_O output units. The activity of output unit j, y_j, is simply a weighted sum of the inputs, x_i, passed through a sigmoid squashing function that limits the output in the range [–1, +1]:

$$y_j = \tanh\left(\sum_{i=1}^{n_I} w_{ji} x_i\right).$$

A training set consists of n_L paired associates to be learned, $\{(\boldsymbol{x}^1, \boldsymbol{d}^1), \ldots, (\boldsymbol{x}^{n_L}, \boldsymbol{d}^{n_L})\}$, where the superscript is the index over pairs in the training set, and $\boldsymbol{x}$ and $\boldsymbol{d}$ are the activity vectors of the cue and target of the pair, respectively. To reflect the fact that items to be learned in the behavioral studies are arbitrary, make little contact with existing knowledge, and have no systematic similarity to one another, we assume that the cue and target activity vectors are random. (Further details in the methodology section that follows.)

In neural net models of cognition, the training of the model is often viewed as an abstract procedure for loading knowledge into a network, and as having no direct correspondence to the sequence of episodes a human learner experiences. In contrast, we commit to a one-to-one correspondence: An SO trial in a behavioral experiment is modeled as one weight update in the neural network. For many neural net architectures and learning procedures, this correspondence is implausible; training the network requires dozens if not hundreds of passes through the training examples, and training on one example can result in catastrophic interference with other examples. We avoid these problems in two ways. First, our architecture has direct connections from input

units to output units, in contrast to strictly layered architectures with hidden units. Second, we endow our architecture with as many inputs as training examples, i.e., $n_L = n_I$; consequently, cues are approximately orthogonal to one another, and interference among examples is minimal. Due to the architecture, the model can learn associations with roughly the same number of exposures as a human participant in a paired-associate experiment.

We use the standard supervised learning procedure for associative networks, a generalization of the Widrow-Hoff or LMS learning algorithm (Widrow & Hoff, 1960) to nonlinear outputs. Following presentation of a cue $\boldsymbol{x}$ to be paired with target $\boldsymbol{d}$, a weight update is performed:

$$\Delta w_{ji} = \varepsilon(d_j - y_j)x_i(1 + y_j)(1 - y_j)$$

where ε is a step size (learning rate).

Having described the general class of models we consider, we turn to two specific hypotheses concerning the nature of learning via self testing.

Hypothesis 1: Self-generated training

One hypothesis is based on the notion of Guthrie (1952) that one learns what one does. That is, when individuals test themselves, they generate a candidate response, and then learn the association between the cue and the candidate response, whether it is correct or incorrect. If the candidate is correct, existing connections are strengthened and are therefore more resilient to decay or interference; if the candidate is incorrect, the wrong association is reinforced, making it more difficult to unlearn.

This interpretation of self testing suggests that testing should benefit an individual only if the material is already somewhat familiar. Some evidence indeed suggests that testing on novel paired associates—when individuals cannot possibly know the correct response—is detrimental to learning (Cunningham & Anderson, 1965).

By this hypothesis, a TS trial involves the following steps: (1) The cue is presented and a candidate response is generated. (2) The LMS weight update is computed for the candidate response. (3) When the target is eventually presented, the LMS weight update is computed for the experiment-provided target. In contrast, an SO trial involves only the third step. In a TS trial, two weight updates are generated; the weight updates are added together and performed at the end of the trial.

How does the model generate a candidate response? It might produce an output and then deterministically select the nearest *well formed state*, defined as a state which has a meaning in the domain (e.g., the set of all targets used for training, plus some distractor alternatives, plus a null or "no response" state). However, individuals are essentially guessing at early stages of learning, and the deterministic rule implies an ability to find the best response among a set of barely-known alternatives. Instead, one might wish to introduce a stochastic selection rule. A standard stochastic procedure for reading out from a neural network is to use a Luce choice or Boltzmann rule (Luce, 1959). By this rule, the distance between each possible response, $\boldsymbol{r}^i$, and the network output, $\boldsymbol{y}$, is computed, $v_i = \|\boldsymbol{r}^i - \boldsymbol{y}\|^2$, and the probability of choosing response i is

$$p_i = \exp(-\beta v_i) / \sum_j \exp(-\beta v_j),$$

where large β achieves a more deterministic selection.

Rather than treating β as a free parameter, we chose β such that the mean correct-response probability is 0.95 if the network produces the correct response on each trial. The model has other free parameters, though, including learning rates for supervised and self-generated targets, the number of distractor vectors considered as candidates for response selection, and the possibility of memory (weight) decay that introduces forgetting.

Hypothesis 2: Complete processing of cue

Carrier and Pashler (1992) speculated on an intriguing basis for the self-testing benefit. They reasoned that in neural net models that learn by error correction, which includes the LMS algorithm, learning requires a comparison between the desired output and the *actual* output—the output that the network produces given its current state of knowledge. If presentation of the target simultaneously with the cue "contaminates" (to use their term) production of the actual output, learning would be less efficient. Essentially, presentation of the target terminates ongoing processing and interferes with the estimation of error needed for learning.

An elegant instantiation of this hypothesis in the context of neural net models is via the incorporation of time into the neural net, specifically, the notion that units in a neural net are slow integrators of information and therefore require many time steps for information to propagate from the input layer to the output layer (McClelland, 1979). We can do this in the network by indexing its output by the (discrete) time step t, i.e., $y_i(t)$, and adding a time constant to the activation dynamics:

$$y_j(t) = (1 - \tau)y_j(t - 1) + \tau\tanh(\sum w_{ji}x_i),$$

where $y_i(0) = 0$ Asymptotically, the output is independent of the time constant τ, for $0 < \tau \leq 1$, but τ determines the rate at which convergence is achieved, i.e., how rapidly information is transmitted from the input to the output.

If we assume that activation dynamics freeze—equivalent to setting $\tau = 0$—when the target is presented, the model will produce a more accurate estimate of its output in the TS condition than in the SO condition, and the learning procedure will have a better estimate of the error. From another perspective, note that if activation dynamics freeze at $t = 1$, the actual output y will be zero, and the training procedure reduces to a form of Hebbian learning; at the other extreme, if the activation dynamics do not freeze and the asymptotic

value $\boldsymbol{y}(\infty)$ is used for training, the learning procedure is exactly the LMS gradient descent step. Because LMS is a more powerful procedure than Hebb, it should yield better performance.

For our simulations, we established a relatively coarse-grained correspondence between time steps in the neural net and real-world time by designating the duration of each time step to be 250 msec. Rather than leaving the time constant τ as a free parameter, we chose τ in advance such activation would reach halfway to asymptote by 2000 msec. To match the TS condition in the behavioral studies, 20 time steps (= 5 seconds) of processing was allowed before the onset of the target. For the SO condition, we had some freedom to determine the time step at which activation dynamics freeze. Although the cue and target appeared simultaneously in the behavioral experiments, participants may nonetheless have done some amount of processing of the cue before the target is processed. We experimented with 0, 1, and 2 time steps of processing in the SO condition, and all yielded similar results; we chose 1 time step in modeling Carrier and Pashler, because that was a sufficient amount of processing to ensure that with enough practice, the model could learn the items in the SO condition. For evaluation of the model during the final test, the activity at time step 80 was used.

Simulation Studies

General Methodology

In all simulations, we used networks with $n_L = n_I = n_O$. Cue and targets were random binary vectors in $\{-1, 1\}^{n_I}$. To model Experiments 1-3 of Carrier and Pashler, we used the same number of items as in their experiment, $n_L = 40$ for Experiments 1 and 2, and $n_L = 30$ for Experiment 3. Because Experiment 2 is essentially a replication of Experiment 1 with different stimulus materials, and the materials in both experiments were intended to be unfamiliar to participants, we capture both experiments with one simulation. Half of the items were assigned to the SO condition and half to the TS condition.

Weights in the neural network were initialized to zero. We also conducted stimulations in which the initial weights were chosen from a normal distribution with mean zero and standard deviation 0.001. However, because the variability of the weights had no systematic effect on the results, we simplified by eliminating this source of noise from the simulation.

The experiments of Carrier and Pashler each involved three *epochs* of training, followed by a final test. An epoch is a presentation of all items in the training set. Within an epoch, order of presentation was randomized, with the constraints that Carrier and Pashler imposed to ensure an intermixing of SO and TS items.

Epochs were of two sorts: in a *pure study* epoch, all items were studied without testing, regardless of whether they were SO or TS items; in an *experimental* epoch, presentation of an item depended on whether it was assigned to the SO or TS condition. In Experiments 1 and 2, the first epoch was pure study (denoted *S*), and epochs 2 and 3 were experimental (denoted *E*); we use the shorthand notation *SEE* for this design. In Experiment 3, the first two epochs were pure study and the third was experimental, i.e., an *SSE* design.

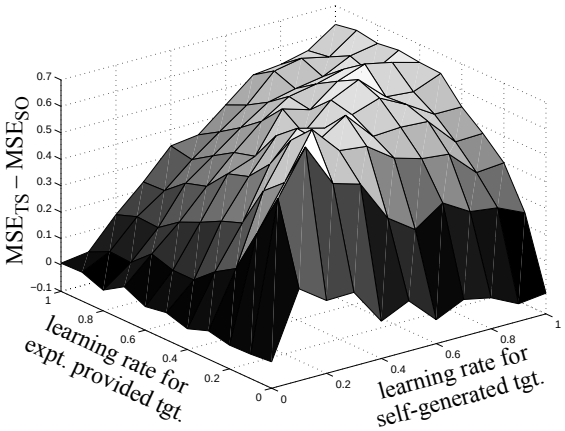


FIGURE 1. Testing error for TS minus SO as a function of the learning rates for self-generated and experiment-provided targets. The difference is nonnegative everywhere, indicating no enhancement through testing. For this simulation, no additional distractor states or weight decay are included.

All results reported are a mean computed from 1,000 independent simulations, where the simulations differ from one another in the choice of random training vectors and the randomization of items within an epoch.

We use mean squared-error (MSE) as a measure of performance of the model. With additional assumptions, we could classify a response as correct or incorrect (e.g., using the stochastic read out procedure that is built into Hypothesis 1), but there is little value in transforming a qualitative fit to a quantitative fit if several new assumptions are required. Consequently, we focus on obtaining qualitative measures of recall, and determining how manipulations of the model affect relative recall.

Hypothesis 1: Self-generated training

After a systematic exploration of the model parameter space, we failed to find any parameter settings that yielded an enhancement of learning by testing. Error was consistently higher in the TS condition than in the SO condition. The two conditions converge as the learning rate for the self-generated target approaches zero, where at the limit SO and TS become identical. Figure 1 illustrates one exploration of the parameter space.

In retrospect, the negative result should not have been surprising. After one or two epochs, the model—like people—is about as likely to make an error as to guess correctly. Consequently, the model will receive as much training from self-generated targets that steers it away from veridical recall as training that steers it toward veridical recall.

Hypothesis 2: Complete processing of cue

Fortunately, our second hypothesis yields more encouraging results. Consistent with Carrier and Pashler, the model produces an enhancement of learning by testing—a lower error for TS than SO—in simulations of Experiment 1/2 (one simulation for both experiments, since they are identical except for the stimulus materials) and Experiment 3 (right side of Table 1). In these simulations, we chose a learning rate that yielded the best possible performance, averaged over TS and SO items. However, the testing benefit was robust over the choice of learning rate.

Figure 2 facilitates a better understanding of the phenomenon in terms of the model. The Figure shows mean-squared error for TS and SO items for four different experimental designs. All designs involve three epochs of training, but they differ in how many epochs of pure study (*S*) precede the experimental (*E*) epochs. The designs range from all study (*SSS*) to all experimental (*EEE*). *SEE* and *SSE* correspond to Experiment 1/2 and Experiment 3, respectively.

The Figure shows that two testing trials helps more than one (*SEE* versus *SSE*). Interestingly, three testing trials shows little benefit over two (*EEE* versus *SEE*). This latter result was at first surprising to us, because it would seem that the more accurate error estimate that is obtained via testing would benefit epoch 1 as well as epochs 2 and 3. However, with an untrained net whose weights are all zero or close to zero, the initial output of the net is close to zero regardless of the number of time steps of activation dynamics.

Comparing SO items in *SSS* versus *EEE* designs, it appears that the SO items benefit from being in a context where testing is occurring; this is a bit surprising considering that the training of these items is identical and learning rates are identical across designs.The result also cannot be explained by virtue of generalization from the better-learned TS items to the SO items, because the items were generated with no systematic similarity structure. Instead, we suggest that the transfer from TS to SO is due to the TS items generating a more meaningful error signal—an error signal that reflects the sort of outputs the network is likely to produce if it is allowed to run to asymptote. Although the precise outputs will differ from one cue to another, the TS items provide information about the distribution of activity values for each output unit across items. This information can certainly be used to determine characteristics of the weight vector (e.g., its overall magnitude, and the sign of biases).

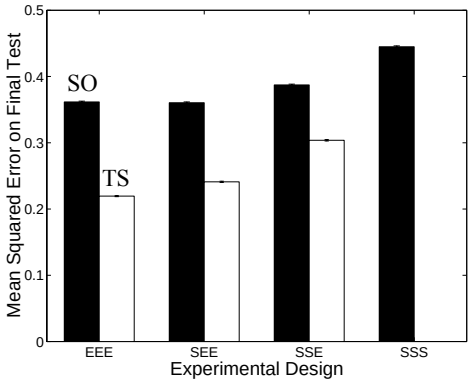


FIGURE 2. Mean-squared error in TS (white bar) and SO (black bar) conditions for experimental designs with three training epochs. +/–1 standard error of the mean is indicated.

We discuss the implication of these results next.

Discussion

In simulations of two models, we found that one hypothesis for the enhancement-of-learning-through testing effect—the hypothesis that self-generated responses are used as targets for further training—is not supported. Another hypothesis—that presentation of the target terminates processing of the cue—is consistent with the experimental data. In the remainder of the paper, we discuss predictions, extensions, and implications of the second hypothesis.

Predictions

- Our model predicts little difference between an *EEE* design and the *SEE* design used in Experiment 1/2. That there is no cost to testing on the first epoch runs against at least one experimental study (Cunningham & Anderson, 1965), but that study used a quite different methodology, and the finding of an initial-epoch-testing cost has not been widely reported in the literature.
- Our model predicts that an SO item should benefit from being embedded among TS items. If it is observed experimentally, a natural interpretation of this effect is that the greater effort on TS items spills over to the SO items. However, the model achieves this spillover without any notion of generalized "effort."
- Our model predicts the relative magnitude of the testing enhancement as a function of the cue-target asymmetry (CTA), i.e., the difference in time between the onset of the cue and the onset of the target. The Carrier and Pashler experiment used a CTA of 5 seconds (20 time steps in the model). One could conduct an experiment in which the CTA was longer or shorter. (Because the time scale of retrieval in the model was set arbitrarily, there is a degree of indeterminacy in the model's predictions. Nonetheless, with one free parameter tied down, the model can characterize the effect of shorter or longer CTAs) Figure 3 shows the model's performance as the CTA is varied. For small CTAs, there is little difference between SO and TS conditions; for large CTAs, the conditions are similar to those studied in the present simulations. Clearly, increasing the CTA has diminishing returns.

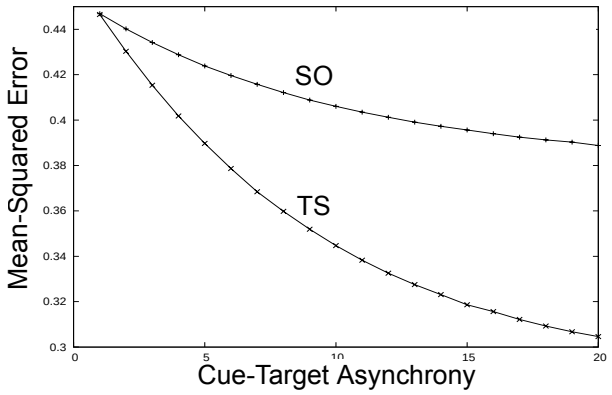


FIGURE 3. Test performance as the cue-target asynchrony in training is varied in the TS condition.

Extensions to the Model

In both SO and TS conditions, each simulation trial began by presenting the cue for T time steps—T being different for SO and TS conditions—at which point a weight update was performed to reduce the difference between the actual response, $\boldsymbol{y}(T)$, and the target. An alternative procedure involves updating the weights to reduce the difference between *each* of $\boldsymbol{y}(1)$, $\boldsymbol{y}(2)$, ..., $\boldsymbol{y}(T)$ and the target. This alternative procedure encourages the net to produce the target as rapidly as possible, and is equivalent to a form of *temporal difference* (TD) learning known as TD(1) (Sutton, 1988). Temporal difference learning is concerned with learning to predict the future given successively better information over time—exactly the situation experienced by the network with time constants, because the propagation of information occurs gradually. However, TD(1) is often not useful in practice because the earliest predictions, e.g., $\boldsymbol{y}(1)$, are treated as important as later, better predictions, e.g., $\boldsymbol{y}(T)$. To remedy this problem, Sutton proposed a family of algorithms, denoted TD(λ), for $0 \le \lambda \le 1$, where λ is roughly the emphasis on achieving correct early predictions. The λ that yields optimal performance depends on the domain. Although it would be interesting to discover how the self-testing benefit depends on λ, the deeper contribution of casting the learning procedure in the TD framework is that it offers a rationale for the termination of processing when the target is presented.

The TD framework is based on the notion that learning mechanisms are fundamentally concerned with predicting eventual outcomes at the earliest possible moment. The adaptive value of prediction is clear; accurate prediction can avoid danger and missteps. Considering the associative learning task in this manner, the cue is a predictor of the target, and TD learning aims to get from the cue to the target as rapidly as possible. However, once a target has been presented, nothing remains to be predicted. The TD framework has been valuable for explaining a broad range of data, from the animal conditioning literature (Sutton & Barto, 1981) to the neural basis of reward (Schultz, Dayan, & Montague, 1997). It seems a natural extension to the mechanisms of associative learning, although one must confront the finding that associative learning appears symmetric (Kahana, 2002).

REFERENCES

Bartlett, JC, & Tulving, E. (1974). Effects of temporal and semantic encoding in immediate recall upon subsequent retrieval. *JVLVB*, *13*, 297-309.

Bjork, R.A. (1975). Retrieval as a memory modified: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium.* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition,* 20*,* 632-642.

Cunningham, DJ, & Anderson, RC (1968). Effect of practice time within prompting and confirmation presentation procedures on paired associate learning. *JVLVB,* 7, 613-616.

Guthrie, E. (1952). *The Psychology of Learning (Rev. Edition)*. New York: Harper.

Hogan, RM, & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *JVLVB*, *10*, 562-567.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18*, 879-919.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*, 823–840.

Landauer, T. K. & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). London: Academic Press.

Luce, R.D. (1959) *Individual choice behavior: A theoretical analysis*. New York: John Wiley & Sons.

Mandler, G. (1979). Organization and repetition: Organizational principles with special reference to rote learning. In L.-G. Nilsson (Ed.), *Perspectives on memory research: Essays in honor of Uppsala University's 500th Anniversary* (pp. 293-327). Hillsdale, NJ: Erlbaum.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287-330.

McDaniel, MA, & Masson, MEJ. (1985). Altering memory representations through retrieval. *JEP:LMC*, *11*, 371-385.

Schultz, W., Dayan, P. & Montague, R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Sutton, R. (1988). Learning by the method of temporal differences. *Machine Learning, 3,* 9–44.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–170.

Widrow, B., & Hoff, M.E. (1960). Adaptive switching circuits. In *IRE WESCON Convention Record*, *pt. 4*, 96-104.

Young, JL (1971). Reinforcement-test intervals in paired associate learning. *J ournal. of Math. Psych.*, *8*, 58-81.

Control of Response Initiation: Mechanisms of Adaptation to Recent Experience

Michael C. Mozer
Department of Computer Science & Institute of Cognitive Science
University of Colorado

Sachiko Kinoshita
MACCS & Department of Psychology
Macquarie University

Colin J. Davis
MACCS
Macquarie University
Sydney, NSW 2109

Abstract

In most cognitive and motor tasks, speed-accuracy trade offs are observed: Individuals can respond slowly and accurately, or quickly yet be prone to errors. Control mechanisms governing the initiation of behavioral responses are sensitive not only to task instructions and the stimulus being processed, but also to the recent stimulus history: when stimuli can be characterized on an easy-hard dimension (e.g., word frequency in a naming task), an easy item is responded to more slowly when intermixed with hard items than when presented among other easy items; likewise, hard items are responded to more quickly when intermixed with easy items. We propose a mathematical theory with three components: a model of temporal dynamics of information processing, a decision criterion specifying when a response should be initiated, and a mechanism of adaptation to the stimulus history. Performance during the course of an experimental trial is cast in terms of a utility function that increases with accuracy and decreases with response time. We assume a decision criterion that initiates a response at the point in time that maximizes expected utility. We posit that the effect of the stimulus history arises because information concerning recent trial difficulty is incorporated into the utility estimate. We present further behavioral studies to validate predictions of the theory.

Consider a simple task in which you are asked to name the sum of two numbers, such as 14+8. Given sufficient time, you presumably produce the correct result; however, under speed pressure, mistakes can occur. In most all cognitive and motor tasks, such empirical *speed-accuracy trade offs* are observed: Individuals can respond slowly yet accurately, or quickly and be prone to errors. Speed-accuracy trade offs are due to the fact that evidence accumulates gradually in response systems over time (Rabbitt & Vyas, 1970). Responses initiated earlier in time will be based on lower quality information, and hence more likely to be incorrect. This paper addresses a simple yet fundamental form of cognitive control—the mechanism that governs the initiation of a behavioral response, and therefore, where an individual operates on the speed-versus-accuracy continuum. In the following section, we describe data that place constraints on the nature of control mechanisms. We describe shortcomings of existing theoretical frameworks that have tried to account for these data. We then present a framework that successfully explains key phenomena and makes further predictions which we have verified through additional behavioral studies.

The Blocking Effect

To understand the control mechanism that initiates responses, consider the variables that affect its operation. The mechanism is influenced by task instructions: individuals can choose to emphasize speed or accuracy. The mechanism is also influenced by recent performance: participants often slow down after producing an error (Rabbit & Vyas, 1970). Finally, even in the absence of errors, the mechanism is sensitive to the recent stimulus environment (Kiger & Glass, 1981): when items are presented in a sequence or *block*, reaction time (*RT*) and error rate to an item depends on the immediately preceding items.

This *blocking effect* is generally studied by manipulating item difficulty. Some items are intrinsically easier than others, e.g., 10+3 is easier than 5+8, whether due to practice or the number of cognitive operations required to determine the sum. By definition, individuals have faster RTs *and* lower error rates to easy problems. However, the RTs and error rates are modulated by the composition of a block. Consider an experimental paradigm consisting of three trial blocks: just easy items (*pure easy*), just hard items (*pure hard*), and a mixture of both in random order (*mixed*). When presented in a mixed block, easy items slow down relative to a pure block and hard items speed up. Thus, the control mechanism that initiates responses uses information not only from the current stimulus, but also adapts to the stimulus environment in which it is operating. Table shows a typical blocking result for a word reading task, where word frequency is used to manipulate difficulty. Based on our review of the blocking-effect literature (e.g., Lupker, Brown & Columbo, 1997; Lupker, Kinoshita, Coltheart, & Taylor, 2000; Taylor & Lupker, 2001), we summarize the central, robust phenomena as follows.

TABLE 1. RTs and Error Rates for Blocking study of Lupker, Brown, & Columbo (1997, Experiment 3)

	Pure Block	Mixed Block	Difference
Easy Item	488 ms (3.6%)	513 ms (1.8%)	+25 ms (–1.8%)
Hard Item	583 ms (12.0%)	559 ms (12.2%)	–24 ms (+0.2%)

(1) Easy items are faster *and* less error prone than hard.

(2) When intermixed, easy items slow down and hard items speed up. However, the convergence of RTs for easy and hard items in a mixed block is not complete. Thus, RT depends both on the stimulus type and the composition of the block.

(3) Speed-accuracy trade offs are observed: a drop in error rate accompanies easy-item slow down; a rise in error rate accompanies hard-item speed up.

(4) Blocking effects occur across diverse paradigms, including naming, arithmetic verification and calculation, target search, and lexical decision. They are obtained when stimulus or response characteristics alternate from trial to trial (Lupker et al., 2000). Thus, the blocking effect is not associated with a specific stimulus or response pathway, but rather is a general phenomenon of response initiation.

(5) Overt responses are necessary for obtaining blocking effects, but overt errors are not.

(6) A signature of the effect concerns the relative magnitudes of easy-item slow down and hard-item speed up. Significantly more speed up than slow down is never observed. The trend is that speed up is less than slow down—indeed, some studies show no reliable speed up—although equal magnitude effects are observed.

(7) The effects of stimulus history are local, i.e., the variance in RT on trial n due to trial $n–k$ decreases rapidly with k. Dependencies for $k>2$ are not reliable (Taylor & Lupker, 2001).

Explanations for the Blocking Effect

The blocking effect demonstrates that the response time depends not only on information accruing from the current stimulus, but also on recent stimuli in the trial history. Therefore, any explanation of the blocking effect must specify how control processes, which determine the point in time at which a response is initiated, are sensitive to the composition of a block. Various mechanisms of control adaptation have been proposed.

Domain specific mechanisms. Most of the proposed mechanisms are domain specific. For example, Rastle and Coltheart (1999) describe a model with two routes to naming, one lexical and one nonlexical, and claim that the composition of a block affects the emphasis that is placed on the output of one route or the other. Meyer, Roelofs, and Levelt (2003) manipulate word length and explain blocking effects in terms of a control process, sensitive to block composition, that decides when to initiate a naming response—either after the motor program for the first syllable has been generated, or after the motor program for the entire word has been generated. Because of the ubiquity of blocking effects across tasks, domain-specific accounts are not compelling. Parsimony is achieved only if the adaptation mechanism is localized to a stage of response initiation common across stimulus-response tasks.

Rate of processing. Kello and Plaut (2003) have proposed a *rate-of-processing* explanation, according to which control processes adjust a gain parameter on units in a dynamical connectionist model. The parameter determines the steepness of the sigmoid curve. Technically, the gain does not affect rate of processing, i.e., it does not simply rescale time. Increasing the gain does result in more rapid convergence, but it also yields a higher error rate; thus the account should more appropriately be framed in terms of adapting the *rate of convergence.* Simulations of this model have explained the basic blocking effect, but not the complete set of phenomena we listed previously. Of greater concern is the fact that the model predicts that naming *duration* decreases with increased speed pressure, which doesn't appear to be true (Damian, 2003; Kinoshita, unpublished).

Evidence criterion. A candidate mechanism with intuitive appeal is the trial-to-trial adjustment of an *evidence criterion*, which specifies the level of evidence that must accumulate in support of a decision before the response is initiated. Random walk and diffusion models have such a parameter, often called the response criterion (Ratcliff, 1978). According to this account, the evidence criterion is determined by recent trial history: if previous trials were easy, the criterion is set low, if previous trials were hard, the criterion is set high. Thus, the criterion would be lowest in a pure-easy block, intermediate in a mixed block, and highest in a pure-hard block. When the criterion is high, RTs are slower but error rates are lower, resulting in slow down of easy items and speed up of hard items in a mixed block.

Taylor and Lupker (2001) illustrate that adaptation of an evidence criterion can—at least in some models—yield incorrect predictions concerning the blocking effect. Strayer and Kramer (1994) attempted to model the blocking effect with an adaptive response criterion in the diffusion model. They managed to fit their blocking data but the account had two fatal shortcomings. First, they allowed different criteria for easy and hard items in a mixed block, which makes no sense because the trial type was not known in advance, and setting differential criteria depends on knowing the trial type. Second, they used a nonstandard blocking paradigm in which the trial difficulty depended on whether an item was presented in a pure or mixed block, easy items being more difficult and hard items being less difficult in a mixed block.

In spite of these problems, we were also convinced an evidence-criterion-adjustment explanation should be feasible. We used a somewhat different model of temporal dynamics and response initiation than Strayer and Kramer (to be described shortly), but like Strayer and Kramer, the model had an adaptive parameter that determined the trade off between speed and accuracy. After six frustrating months of exploration, we admitted defeat: The model was unable to obtain the right qualitative pattern of results; either the blocking effect was an order of magnitude smaller than that observed in experi-

ments, or went in the wrong direction. Several variants of the model came close, but were not robust; tiny changes to parameter values yielded qualitative effects on the pattern of results.

The failure of an evidence-criterion-adjustment account is not surprising from another perspective. On logical grounds, the relative importance of speed versus accuracy should be determined by task instructions and pay offs. Item difficulty is independent and unrelated factor. Consistent with this logical argument is the finding that manipulating instructions to emphasize speed versus accuracy does not produce the same pattern of effects as altering the composition of a block (Dorfman & Glanzer, 1988).

Adaptation to the statistics of the environment. Having ruled out three possible explanations, we sketch a fourth alternative, which is based on the premise that the goal of cognition is optimal and flexible performance across a variety of tasks and environments. In service of this goal, control mechanisms must be sensitive to the statistical structure of the environment, e.g., stimulus characteristics and configurations, response contingencies, etc. Previous models of control have exploited this assumption. For example, Treisman and Williams (1984) and Mozer, Colagrosso, and Huber (2002) considered a sequential choice task involving two response alternatives, and proposed that control mechanisms estimate prior probabilities of the two responses. If one response is more frequent, the larger prior induces a bias toward that response, which typically boosts performance.

Blocking effects can be explained via a related hypothesis. Because RTs depend on whether a trial is easy or hard, the control mechanisms responsible for response initiation must utilize an estimate of the item difficulty, or the quality of information available to response processes. If this estimate is unreliable (noisy), and if control mechanisms make the ecological assumption that the current trial is similar to recent trials, the estimate can be made more reliable by incorporating estimates from recent past trials. We elaborate this idea in a mathematical model of response initiation, and show that it can explain the key blocking phenomena listed earlier as well as other puzzling phenomena.

The ASE Model

We refer to our model as ASE, which stands for *Adaptation to the Statistics of the Environment*. Although the key claim of the model concerns the *mechanism of control adaptation* based on recent experience, we must make two additional sets of assumptions, one set concerning the *temporal dynamics* of information processing, and another set concerning the *decision criterion* for response initiation Although the specific assumptions we make are not critical, they must be made explicitly to fully flesh out the model.

Temporal dynamics. We need a way to characterize the temporal dynamics of information processing in tasks such as naming. The particular model of temporal dynamics is not critical, as long as it has the property that the quality of information available for responding increases gradually and monotonically over time.

We chose the probabilistic information transmission (PIT) model of Mozer, Colagrosso, and Huber (2002, 2003). To summarize the key properties relevant for the current work, the model consists of a cascaded series of processing *pathways* whose details are determined by the task being modeled. For example, to model a word naming task, we use a perceptual pathway that maps visual word forms to an internal semantic/lexical representation, and a response pathway that maps the internal representation to a distinct verbal naming response. Each pathway is a dynamic Bayesian network, and the conditional probability distributions in the model are specified by the nature of the mapping, the state of expertise being modeled, and the similarity structure among elements of representation. Given a stimulus presentation, the output of the model is a probability distribution over response alternatives as a function of time (Figure 1a). The response chosen at a particular time is a sample from the distribution (the model cannot choose the most probable response). The time course of processing depends on information transmission probabilities in the model. Easy, high frequency, and well practiced items have higher transmission probabilities, and hence are conveyed more rapidly.

This model is a generalization of random walk models and has several advantages. It provides a mathematically principled means of handling multiple alternative responses (necessary for naming) and similarity structure among elements of representation, and characterizes perceptual processing, not just decision making. The counter model (Ratcliff & McKoon, 1997) or connectionist integrator models (e.g., Usher & McClelland, 2001) could also serve us, although the PIT framework has an advantage in that it operates using a currency of probabilities—versus more arbitrary units of *count* or

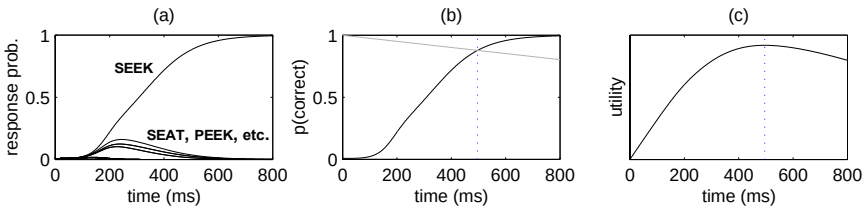


FIGURE 1. (a) Output of the probabilistic information-transmission model for presentation of the stimulus, e.g., SEEK, on a word naming task. (b) Treating the most probable output is an estimate of accuracy (light line is the time threshold). (c) Utility function based on most probable output (dashed line = time of maximum utility)

activation—which leads to explicit, interpretable decision criteria and adaptation mechanisms, and requires fewer additional assumptions to translate model output to predictions of experimental outcomes.

Decision criterion. To model blocking effects, we must make an explicit assumption concerning the decision criterion used for response initiation. A simple speed criterion (i.e., respond at α milliseconds following stimulus onset) or accuracy criterion (i.e., respond when the error rate is below α) is inadequate, because easy items are both faster and more accurate than hard items in pure blocks. Ratcliff's (1978) diffusion model uses an evidence threshold, which effectively yields an accuracy criterion that declines over time. We adopt this notion, as illustrated by the gray line in Figure 1b. The line is characterized by one free parameter, the slope κ. This criterion can be recast in an optimization framework: A response is initiated at the point in time that maximizes *utility*, where utility increases with expected accuracy and decreases with time (Figure 1c). Previous psychological theory has suggested that individuals can choose the optimal point at which to respond (Mozer et al., 2002; Rabbitt & Vyas, 1970; Triesman & Williams, 1984).

To summarize, we propose a theory premised on five key assumptions. (1) Transmission of stimulus information to response systems is gradual and accumulates over time. (2) Control mechanisms respond at the point in time that maximizes a utility measure that depends on both expected accuracy and time. (3) During ongoing processing, the system is able to compute an estimate of its response accuracy for the current stimulus. (4) This estimate is unreliable. (5) If control systems make the ecological assumption that the current trial is similar in difficulty to recent trials, the accuracy estimate can be made more reliable by incorporating estimates from recent trials.

An accuracy estimate can be obtained from the PIT dynamics by assuming that the most probable output at a point in time is correct (Figure 1b); we refer to this as curve as the *current accuracy trace (CAT)*. Given the response criterion (grey line, Figure 1b), a response initiation time can be determined (dashed line).

If the model's transmission probabilities are noisy, the CAT is a high-variance estimate of accuracy, because the assumption that the most probable response state is correct may be wrong. The suggestion of noise is not arbitrary, but rather is a central claim of the diffusion model, and has been key to explaining a variety of RT data. To overcome this noise source, it is sensible for control mechanisms to rely not solely on the CAT, but on accuracy traces from recent trials. We claim that the model maintains a *historical accuracy trace (HAT)*, and the trace used for estimating utility—the *mean accuracy trace (MAT)*—is a weighted average of CAT and HAT, i.e., $\mathrm{HAT}(n) = \lambda\mathrm{CAT}(n-1) + (1-\lambda)\mathrm{HAT}(n-1)$, where n is an index over trials, and $\mathrm{MAT}(n) = \theta\mathrm{CAT}(n) + (1-\theta)\mathrm{HAT}(n)$; λ and θ are averaging weights. Figure 2a depicts the CAT, HAT, and MAT. The two solid curves represent CATs for easy and hard trials, as well as the MATs for pure blocks. The dotted curve represents the expected HAT in a mixed block—an average of easy and hard CATs. The dashed curves represent the MATs for easy and hard trials in a mixed block, formed by averaging the HAT and corresponding CAT. Because the CAT and HAT are time-varying functions, the notion of averaging is ambiguous; possibilities include averaging the accuracy of points with the same time value and times of points with the same accuracy value. It turns out that the choice has no qualitative impact on the simulation results we present. The essential requirement is that the computation to determine response-initiation time can be performed in real time, including identification of the utility maximum.

Modeling Blocking Effects

Figure 2b provides an intuition concerning the model's ability to replicate the basic blocking effect. The mean RT for easy and hard items in a pure block is indicated by the point of intersection of the CAT with the time threshold. The mean RT for easy and hard items in a mixed block is indicated by the point of intersection of the MAT with the time threshold. The easy item slows down, the hard item speeds up. Because the rate of processing is not affected by the blocking manipulation, the error rate will necessarily drop for easy items and rise for hard items. Although the RTs for easy and hard items come together, the convergence is not complete as long as $\theta > 0$. The theory thus explains the first three phenomena of Section 1. The fourth phenomenon, that the effects occur across diverse paradigms, is consistent with the theory: the theory concerns the response curves, but not the stimulus or response modalities or domains that underlie the curves. Consequently, cross-task blocking effects are implied by the theory. The theory is consistent with the observation that blocking effects occur even in the absence of overt errors, because the theory is neutral with regard to error production, and only if response mechanisms are engaged (phenomenon 5). If responses are not produced, the

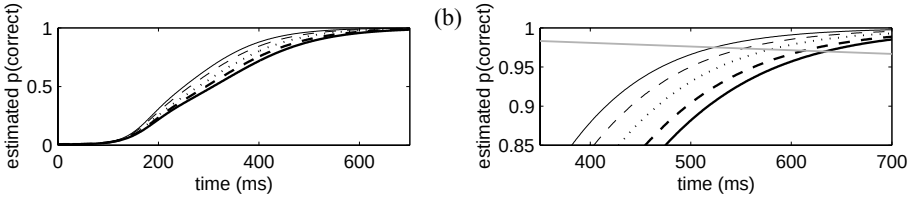


FIGURE 2. (a) easy and hard CAT (solid), mixed block HAT (dotted), and easy and hard MAT (dashed); (b) close up of the traces, along with the decision threshold

TABLE 2. Expt. 1 of Taylor & Lupker (2001): Human data and simulation

	Human Data			Simulation		
	Pure	**Mixed**	**Difference**	**Pure**	**Mixed**	**Difference**
Easy	519 ms (0.6%)	548 ms (0.7%)	29 ms (0.1%)	524 ms (2.4%)	555 ms (1.7%)	31 ms (–0.7%)
Hard	631 ms (2.9%)	610 ms (2.9%)	–21 ms (0.0%)	634 ms (3.0%)	613 ms (3.7%)	–21 ms (0.7%)

TABLE 3. Context experiment: Human data and simulation

	Human Data			Simulation		
	Same Context	**Diff. Context**	**Switch Effect**	**Same Context**	**Diff. Context**	**Switch Effect**
Easy	432 ms	488 ms	56 ms	437 ms	493 ms	56 ms
Hard	514 ms	467 ms	–47 ms	514 ms	470 ms	–44 ms

response-accuracy curves need not be generated, and the averaging process that underlies the effect cannot occur.

The fact that hard-item speed up is never greater than easy-item slow down (phenomenon 6) turns out to be a key diagnostic. Our initial candidate models tended to yield more speed up than slow down because the magnitude of RT change was proportional to the RT, and hard RTs are larger than easy RTs. Empirically, the error-averaging model we propose never yields more speed up than slow down. As shown in Figure 2b, the mixed-block MAT (dashed) hugs the pure-block MAT (solid) more tightly for hard than easy items. The asymmetry is due to the fact that the easy CAT reaches asymptote before the hard CAT. The model produces more symmetric blocking effects when responses are initiated at a point where both easy and hard CATs are ascending at the same rate (leading to high error rates, unlike the behavioral data). However, we were unable to find model parameters that produced the invalid pattern of more speed up than slow down.

Beyond providing qualitative explanations for key phenomena, the model fits specific experimental data. Taylor and Lupker (2001, Expt. 1) instructed participants to name high frequency words (easy items) and nonwords (hard items). Table 2 compares mean RTs and error rates for human participants and the simulation. One should not be concerned with the error-rate fit, because measuring errors in a naming task is difficult and subjective. (Over many experiments, error rates show a speed-accuracy trade off.) Taylor and Lupker further analyzed RTs in the mixed block conditional on the context—the 0, 1, and 2 preceding items. Figure 3 shows the RTs conditional on context. The model's fit is excellent. Trial *n* is most influenced by trial *n–1*, but trial *n–2* modulates behavior as well; this is well modeled by the exponentially decaying HAT.

Simulation details. Parameters of the PIT model were chosen to obtain pure-block mean RTs comparable to those obtained in the experiment and asymptotic accuracy of 100% for both easy and hard items. We added noise to the transmission rates to model item-to-item and trial-to-trial variability, but found that this did not affect the expected RTs and error rates. We fixed the HAT and MAT averaging terms, λ and θ, at 0.5, and picked κ to obtain error rates in the pure block of the right order. Thus, the degrees of freedom at our disposal were used for fitting pure block performance; the mixed block performance (Figure 3) emerged from the model.

Asymptotic Effect of Context

In the standard blocking paradigm, the target item is preceded by a context in which roughly half the items are of a different difficulty level. We conducted a behavioral study in which the context was maximally different from the target. Each target was preceded by a context of ten items of homogeneous difficulty, either the *same* or *different* difficulty as the target. This study allows us to examine the asymptotic effect of context switching. We performed this study for two reasons. First, Taylor and Lupker (2001) obtained results suggesting that a trial was influenced by only the previous two trials; our model predicts a cumulative effect of all context, but diminishing exponentially with lag. Second, several candidate models we explored predict that with a strong context, speed up of hard is significantly larger than slow down of easy; the model we've described does not.

The results are presented in Table 3. The model provides an excellent fit to the data. Significantly larger context effects are obtained than in the previous simulation (~50 ms in contrast to ~25 ms), and—given the strong context—the easy items become slower than the

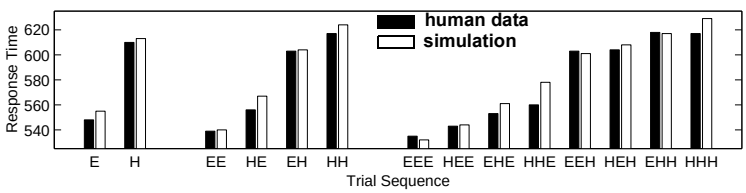


FIGURE 3. RTs from human subjects (black) and simulation (white) for easy and hard items in mixed block, conditional on 0, 1, and 2 previous item types. Last letter in a string indicates the current trial and first letters indicate context. Thus, "EHH" means a hard item preceded by another hard item preceded by an easy item.

TABLE 4. Simulation of validity-modulating masked priming effect

	Repetition-Prime Trial	Unrelated-Prime Trial	Priming Effect
20% valid	560 ms	585 ms	25 ms
80% valid	515 ms	580 ms	65 ms

hard (although this effect is not statistically reliable in the experimental data). Further, both data and model show more slow down that speed up, a result that allowed us to eliminate several competing models. For this simulation, we fit parameters of the PIT model to the same-context results. We also treated the MAT averaging constant, θ, as a free parameter on the rational argument that this parameter can be tuned to optimize performance: if there is not much variability among items in a block, there should be more benefit to suppressing noise in the CAT using the HAT, and hence θ should be smaller. We used 0.35 for this simulation, in contrast to 0.5 for the first simulation.

Reinterpreting Other Experimental Findings

In many studies, contrasts are made between experimental blocks whose composition varies in terms of the proportion of easy and hard items. In such cases, our model may provide an alternative interpretation of experimental results. Consider a subliminal priming study in which participants are asked to perform lexical decision on a target string preceded by a masked prime (Bodner & Masson, 2001). The prime and target could be identical or unrelated. Although the prime was subliminal—not accessible for report—a repetition priming effect is observed: lexical decision RT to a target is faster if the prime is identical to the target. Subliminal repetition priming effects are common in the literature, but what is surprising in this study is that *prime validity* influences priming: The priming effect is larger when the prime and target are identical on 80% of trials (high validity) than when they are identical on 20% of trials (low validity). Bodner and Masson suggest that "recruitment of the prime resource to assist target processing should be more likely when the...prime validity...is higher." (p. 618). This counterintuitive explanation implies that the prime is analyzed deeply: its match to the target is determined, prime validity is estimated, and the estimate is available for strategic control.

Our model offers an alternative account. The repetition prime makes a trial easy because the prime activation supports the target, and the unrelated prime makes a trial relatively hard. Low and high validity conditions are thus mixed blocks containing 20% and 80% easy trials, respectively. We ran a simulation to show that these mixtures yield a blocking effect consistent with the reduction of priming in the low validity condition (Table 4). In the model, the prime influences the time course of information transmission, which modulates the model's response-initiation criterion on future trials—a simpler, more elegant account than Bodner and Masson's.

Conclusions

Theories in cognitive science occasionally hand the problem of control to a homunculus. More commonly, control processes are left unspecified. And when implemented, control generally involves explicit, active, and sophisticated mechanisms. We have described a model that achieves an interesting sort of control—sequential adaptation of the speed-accuracy trade off. However, the mechanism that gives rise to this adaptation is passive and in a sense dumb; it essentially reestimates the statistical structure of the environment by updating an expectation of task difficulty. Our hope and belief is that many aspects of cognitive control can be explained away by such simple, passive mechanisms, eventually eliminating the homunculus from cognitive science.

REFERENCES

Bodner, GE, & Masson, ME (2001). Prime validity affects masked repetition priming: Evidence for an episodic resource account of priming. *Journal of Memory & Language, 45*, 616–647.

Damian, MF (2003). Articulatory duration in single word speech production. JEP: LMC, *29*, 416-431.

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decision and recognition memory. *J. Mem. & Lang.*, 27, 633–648.

Kello, CT & Plaut, DC (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language*, *48*, 207-232.

Kiger, JI, & Glass, AL (1981). Context effects in sentence verification. *JEP:HPP*, *7*, 688-700.

Lupker, SJ, Brown, P, & Colombo, L (1997). Strategic control in a naming task: Changing routes or changing deadlines? *JEP:LMC*, *23*, 570–590.

Lupker, S, Kinoshita, S, Taylor, T, & Coltheart, M. (Nov. 2000) Is a time criterion used when naming pictures and computing sums? *Annual meeting of the Psychonomic Society*, New Orleans.

Meyer, AS, Roelofs, A, & Levelt WJM (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, *48*, 131–147.

Mozer, MC, Colagrosso, MD, & Huber, DE (2002). A rational analysis of cognitive control in a speeded discrimination task. In NIPS XIV (pp. 51-57). Cambridge, MA: MIT Press.

Mozer, MC, Colagrosso, MD, & Huber, DE (2003). Mechanisms of long-term repetition priming and skill refinement: A probabilistic pathway model. In *Proceedings of the Twenty Fifth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum Assoc.

Rabbitt, PMA, & Vyas, SM (1970). An elementary preliminary taxonomy for some errors in laboratory choice RT tasks. *Acta Psych.*, *33*, 56–76.

Ratcliff, R. A theory of memory retrieval. *Psych. Rev.*, 1978, 85, 59-108.

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psych. Review*, *104*, 319–343.

Taylor, TE, & Lupker, SJ (2001). Sequential effects in naming: A time-criterion account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 117–138.

Treisman, M & Williams, TC (1984). A theory of criterion setting with an application to sequential dependencies. *Psych. Review*, *91*, 68–111.

Usher, M, & McClelland, JL (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550-592.

Numerically-Driven Inferencing in Instruction: The Relatively Broad Transfer of Estimation Skills

Edward L. Munnich (munnich@berkeley.edu)
Michael A. Ranney (ranney@cogsci.berkeley.edu)
Daniel M. Appel (dappel@berkeley.edu)
University of California, Graduate School of Education, 4533 Tolman Hall, Berkeley, CA 94720-1670

Abstract

What is the current U.S. immigration rate? Policy-makers, voters, and consumers should have a sense of quantities of this kind in order to help shape effective policies, and schools must prepare students for such roles. We examine the Numerically-Driven Inferencing paradigm (NDI), using a method in which participants: *Estimate* policy-relevant quantities, state *Preferences* for these, receive actual quantities as feedback to *Incorporate*, and offer preferences again to exhibit any policy *Changes* (EPIC). Past work has generally suggested rather poor estimation of such base rates, but there is potential for improvement as one carries out many estimates over various issues, and perhaps a benefit for taking a more analytic approach to estimation. Here we consider whether one can improve estimation skills broadly by using multiple perspectives in estimation problems, and by working out of conflicts that arise among multiple, locally coherent, numerical understandings. Using an NDI curriculum that emphasized disconfirmation, we found that estimation improved across a wide variety of questions.

What is the current annual U.S. immigration rate (including both legal and illegal immigration)? Please take a moment to estimate this quantity, and reflect on the kinds of skills you used to generate your estimate. One might assume that those who know about immigration issues are good at estimating immigration rates, while those who know about environmental issues are good at estimating per capita garbage production, but that there is no general skill for estimating across content domains. Research on estimation suggests that people can improve the accuracy of estimates in a variety of ways, including using category information (e.g., Huttenlocher, Hedges, & Prohaska, 1988), or learning relevant "seed" numbers (e.g., Brown & Siegler, 2001), but there is no indication that such benefits transfer broadly to estimation over a wide variety of quantities, to say nothing of problem solving skills more generally. However, we suggest that in domains ranging from estimation to physics problem-solving, it is important to learn to seek alternatives to initial conceptions of problems, which brings the possibility of disconfirming hypotheses. The potential value of such a strategy is illustrated Johnson-Laird and Hasson (2003), who have found that when some premises are consistent with an invalid conclusion, counterexamples are useful in rejecting the conclusion. The focus of the present paper is on the extent to which analytic estimation skills can transfer broadly, so that people might improve their estimates for quantities across a broad range of issues without specific instruction on those issues.

Theoretical Framework

This project builds on the Numerically-Driven Inferencing paradigm (NDI; Ranney, Cheng, Nelson, & Garcia de Osuna, 2001), which examines how understandings of relevant base rate information (e.g., the present U.S. immigration rate) affects people's attitudes on public policy issues (e.g., given the immigration rate, what would you *prefer* that rate to be?). With NDI's methods, people need not be asked whether they are for or against a particular issue, but rather what they would prefer the *numbers* to be. Indeed, it is not uncommon that those who consider themselves to be in favor of reducing immigration (e.g., believing the current base rate of a policy-relevant quantity to be 10%, one might prefer 5%) have more in common than they realize with those who claim to favor an increase (e.g., believing the rate to be 1%, but sharing a preference for 5%). However, if such people were only asked the extent to which they favor or oppose an issue, they would appear to be at odds. In contrast, NDI asserts that qualitative attitudes have some—albeit not necessarily direct—relationships with relevant quantities, and aims to explore the nature of the relationships. By focusing on numerical concepts, NDI can shed light on how these concepts interact with people's initial attitudes, and the extent to which learning actual values shapes subsequent attitudes: Do we maintain preferences for the same *absolute* rates, or for the same *proportions* relative to actual rates? To what extent do we shift our policy stances after surprising feedback (Munnich, Ranney, Nelson, Garcia de Osuna, & Brazil, 2003)?

NDI builds on research in many fields, such as attitude, conceptual change, mental models, and judgment and decision-making (although NDI deals directly with base rates—not through Bayesian analyses). In particular, NDI has drawn on work in scientific conceptual change including the Theory of Explanatory Coherence (TEC; Ranney & Thagard, 1988; Thagard, 1989), which describes change as spawned by incoherence and conflicts among ideas, such that people try to revise their beliefs to increase global coherence. In an illustration of this, Ranney, Schank, Mosmann, and Montoya (1993; based on a misconception noted by Keysar, 1990) found that most participants initially believed that Berlin lay *on* the East/West German border, but revised their beliefs as they incrementally received

information that could be used to disconfirm "on-border" hypotheses (e.g., they were told/reminded of the Berlin airlift, the Western Allies' agreement to halt their troops far west of Berlin, Berlin's location within united Germany, and northern and southern ends of the border). With each successive piece of evidence, participants moved toward a more accurate view of Berlin's location relative to the border, suggesting that they modified their belief networks to maintain coherence in the face of the new information.

According to TEC, evidence that is critical, germane, and credible carries considerable weight in our belief systems. Within NDI, we seek to understand when and how a particular kind of evidence that meets these criteria—numerical propositions—can catalyze knowledge-transforming effects. NDI asserts that estimates and numerical preferences are outputs of our belief systems—the tips of a "reasoning iceberg." One's understanding of an issue may be thought of as a network of ideas connected by personal experiences, media, religion, etc. When asked to estimate an immigration rate, few can simply recall it. Instead one activates various understandings about immigration that shape the estimate. Likewise, numerical preference is an output from an extensive belief network that lies below the surface of overt response. For example, one might believe the assumed immigration rate to be acceptable and simply reiterate one's estimate as one's preference (a status quo policy). However, if later surprised by the actual immigration rate, one's sense of reality is challenged, and one might come to the conclusion that prior reasoning was incorrect or incomplete.

In this conception, the iceberg's "bulk"—the belief network from which estimates and numerical preferences emerge—may be transformed by the impact of feedback. As such, NDI can offer rich, quantitative findings to cognitive scientists concerned with the dynamics of belief networks. In this paper, we consider *curricula* based on NDI, designed to facilitate the recruitment of multiple, locally coherent understandings that can mutually constrain one another. Just as feedback that conflicts with one's numerical understanding might lead to a transformation, when one spontaneously seeks to disconfirm one's own numerical hypotheses by bringing alternative numerical notions to bear, it may lead to revisions that bring one's belief network into closer alignment with facts of the world. Such a transition would be evidenced by improved estimation across a wide range of issues.

NDI Findings That Frame the Issues

To address NDI, Ranney and colleagues developed a variety of methods, including EPIC (Estimate-Prefer-Incorporate-Change), which is used in this paper: (1) Participants *estimate* a quantity that is relevant to an issue, as you did for the U.S. immigration rate at the beginning of this paper. (2) Participants indicate what they *prefer* the quantity to be; to familiarize yourself, please write down what you would prefer the U.S. immigration rate to be (including both legal and illegal immigration). (3) Participants receive correct base rate feedback to *incorporate*; now, please look at the actual immigration rate in the footnote below.[1] Finally, (4) participants indicate again what they prefer the quantity to be; has your preference *changed* now that you know the actual number? We have found that, to the extent feedback is surprising, it generally leads to nontrivial belief revision. So far, research on estimates within NDI has focused on a rather short period of time, but an obvious extension of this work is to consider (a) whether estimation skills can *improve* with targeted interventions, and (b) the extent to which there may be broad transfer.

Illustrations of the kinds of alternative conceptions that people can have comes from Munnich et al. (2003), who reported differential patterns of estimation for the same underlying question: One group was asked to estimate the number of abortions in the U.S. *per million live births,* while a second group drawn from the same undergraduate class estimated the number of abortions in the U.S. *per million fertile women* each year. The results showed a striking contrast in numerical understanding, depending on how the question was framed: For the *per-women* question the median response (10,000) was *half* the correct answer at that time, but for the *per births* question, the median estimate (10,000 as well, coincidentally) was *33.5 times* too low at the time the study was conducted.[2] Could people perhaps improve their estimates of abortions per live births by considering how many abortions there are per fertile women? More broadly, what might happen when people bring together alternative conceptions and resolve conflicts on their own, without external feedback? To address this issue, McGlothlen (2003) interviewed high school students as they produced estimates and numerical preferences for a variety of issues, and reported on their online reasoning processes. She coded responses as analytic—containing relevant numerical information and constraints—or holistic—based on a feeling or general sense of the issue. McGlothlen found that estimates reached through an analytic process were significantly more accurate than those reached through a holistic process. This leads us to the following hypothesis:

> An analytic approach invokes multiple locally coherent numerical representations that provide mutual constraints among themselves, leading to more refined, more globally coherent, and hence more accurate estimates than would be observed if only one representation were invoked.

To discover whether there is a causal relationship between invoking multiple representations and accuracy, we might manipulate the degree to which people take an analytic approach. Below, we discuss an experiment in which the analytic process is explicitly emphasized in the instruction given to one group of students, and the accuracy of this group's estimates, pre- and post-instruction, is compared to

[1] The U.S. Census Bureau reports that the annual U.S. immigration rate, including legal and illegal immigrants, is 0.4%.

[2] Garcia de Osuna, Ranney, and Nelson (2004) observed a median of 5,000, sixty-seven times too low.

a parallel group who received no such instruction. If those taught analytic strategies show greater estimation accuracy, it would provide causal evidence for the benefit of an analytic approach.

Previous Curricular Interventions

In several recent studies, our group has observed estimation accuracy benefits, arising from practice with forming estimates and generating preferences. These activities are unusual for math or science classes, for which problems are generally solvable in straightforward ways by applying formulas and principles. Our curricula illustrate the utility of mathematical and scientific reasoning through our use of problems about issues that students find interesting. We ask students for societally-relevant opinions, which is virtually unheard of in math classes. These factors motivate students in ways that standard curricula may not, and shows benefits for estimation ability with relatively little practice.

In one such intervention, Curley (2003) and Howard (2003) gave fifth-grade science camp students standard physics labs about the stopping distances of vehicles. An Experimental class received NDI problems to frame the labs, while a Control class did not. Both classes took a pretest with estimation and preference problems, then a posttest with a different set of items three days later.[3] In this between-subjects design, Curley and Howard observed improvement in estimation accuracy for *both* classes on items about U.S. household income and the number of alcohol-related automobile crashes. Notably, the only NDI experience that the Control class received was during the pretest, suggesting that exposure to such items alone might be sufficient to improve estimation abilities.

In a later study, Juan (2003) found similar effects among eighth-grade Algebra students. Her Experimental class received one NDI problem per day for three days, followed by graphing activities and class discussions on estimates and preferences. In contrast, the Control class received standard algebra instruction. All students took a pretest and a posttest, in which they estimated twelve quantities (six per test). Questions dealt with issues such as California's population and teachers' salaries. On each test, students estimated and offered preferences before and after feedback for two items, and simply estimated for the remaining four items. Experimental students showed a significant overall gain between pre- and posttest estimates, while the Control class showed only a marginally significant improvement. However, an additional sign test between groups showed no advantage for the Experimental class.

The studies discussed up to this point showed minimal benefit for the curricula themselves—while both Experimental classes improved, the Control classes may have also benefited just from their pretest experience with NDI. This raised the possibility that one may improve estimation merely by working on NDI problems. Before drawing this conclusion, we carried out the following experiment, which lasted much longer than past interventions, and focused not on content areas like physics or college costs, but on analytic techniques aimed at improving students' general estimation abilities.

[3] When we ask for preferences, the objective is to assess how students' numerical understandings affect their preferences, not to make a normative assessment.

A Focal Experiment

Method

Two high school geometry classes participated, each with 27 students. Both classes received a normal geometry curriculum, but the Experimental class spent 12% of their time over a ten-week period on activities centered on six NDI questions. Activities included discussions and written reflection aimed at promoting the analytic responses that McGlothlen (2003) found to correlate with successful estimation. In addition, explicit connections were made between logical argumentation about issue-relevant quantities and the argumentation required in geometric proof. Due to limited space, we omit discussions of possible benefits regarding student motivation and the transfer of argumentation skills from NDI to geometric problems.

Table 1: Pretest, Intervention, and Posttest Questions
(Group B received the pre- and posttest in reversed order)

Pretest (Group A)	**Intervention**	**Posttest** (Group A)
Average US age	CA population	Cars per driver
Athlete salary	College vs. H.S. grad earnings	College degrees (%)
College cost	H.S. dropout rate	Homes-with-computers
Miles driven/Year	Athlete salaries by gender	% Female teachers
Commute time	Poverty line	Garbage per person
Incarceration rate	US oil imports	Hours of sleep
Soda calories		Inflation
Homes-with-TVs		Voting percentage
US population		Teacher salary
		Car price

On Thursday of each week, students generated estimates and preferences for a given quantity as homework (see Munnich et al., 2003, for examples of how such items are worded). In class on Friday, they discussed their estimates in groups and generated a group estimate, in which they: (a) provided a consensus estimate, (b) explained their rationale for that number, (c) provided rationales for a number considerably higher than their estimate, and for a number considerably lower than their estimate. This was followed by a short class discussion to acquaint students with alternative approaches that their classmates had taken. Between hearing classmates' arguments and generating rationales for estimates and preferences other than their own, students were encouraged to engage in problems analytically, considering the strengths and weaknesses of various constraints that might be placed on the estimate.

The following Monday, students' original estimates and preferences were returned, along with the actual number as feedback. From Monday to Tuesday, they generated final

preferences, based on the feedback and any insights they gleaned from group discussions. Finally, on Tuesday, students discussed their preferences in groups and generated arguments that might be used by one who preferred (a) decreasing the quantity, (b) maintaining the status quo, and (c) increasing the quantity. As with estimation discussions, this was followed by a whole class discussion on preferences.

To measure the intervention's effects, both Experimental and Control classes were given ten NDI items as a pretest, and then, ten weeks later, ten different NDI questions as a posttest (an immigration rate item was excluded when it became clear that responses were bizarrely high in many cases; students also reported numerous misinterpretations). Questions were counterbalanced so that the items that half of each class saw on the pretest (Group A in Table 1) appeared on the posttest for the other half of each class, and vice versa. Students were asked to generate estimates and preferences, and were then handed a separate sheet of paper with the actual quantity, which also elicited a Likert surprise rating and their final preferences. Students received two items per day for five days (as past studies indicated that fatigue sets in when students receive many items on a single day). Each of the ten problems was presented in the same order to all students, minimizing any benefit for discussing items with classmates in unintended ways.

Results and Discussion

All estimates (for both classes, pre- and posttest) were ranked by proximity to the actual value for each item, in order to put data from questions with different scales onto one common scale (i.e., accuracy rankings). Between-group analyses on the rankings assessed whether there were differences between the classes, and whether each class improved from pre- to posttest. Mann-Whitney tests showed no reliable pretest difference between Experimental and Control classes (z=1.04, n.s.). On the posttest, however, the Experimental class estimated reliably more accurately than Controls (z=3.29, p<.001). Further, while there was no difference among Controls on pre- and posttests (z=0.41, n.s.), Experimentals showed a significant improvement (z=2.74, p=.003). These effects indicate that the intervention led to improved estimation of novel quantities (i.e., transfer).

To explore the effect's loci for the Experimental class, planned Mann-Whitney comparisons were performed separately on each item. Participants improved significantly on three items (U.S. population, z=2.49; cars per driver, z=1.97; hours of sleep, z=1.96; ps<.05), and marginally on three other items (college cost, z=1.48; teacher salary, z=1.36; miles driven/year, z=1.19; ps<.10). Among these items, we see patterns of both near and relatively far transfer from the intervention. The only one of these items that was directly related to one of the intervention questions was that on U.S. population (i.e., related to an question on California's population in the intervention). The other items range from those that seem to have only an indirect relationship with intervention items (e.g., "teacher's salary" may be related to "H.S. vs. college grad earnings," although teachers' incomes are closer to the incomes of high school graduates than to those of other college graduates), to items that have no obvious relationship with the intervention problems (e.g., hours of sleep the average person gets).

The results point to benefits from an intervention focused on analytic approaches to estimation. Looking more closely, we found both transfer among highly similar questions, as well as the relatively far transfer of general estimation skills to seemingly unrelated quantities. These findings are in line with the hypothesis that multiple numerical representations provide constraints on one another and can lead to more globally coherent estimates. The difference between the classes was that, although both had experience with estimation and giving preferences on the pretest, the Experimental class received a curriculum that engaged them in discussions of multiple perspectives in estimation and numerical preference. These results are rather surprising if one believes that estimation ability is not a broadly transferable skill. Given the variety of topics covered by items on the pre- and posttests, it is unlikely that the Experimental class could have learned the vast array of new facts about the world necessary to drive observed improvements. Rather, they appeared to use their extant numerical knowledge about the world more constructively than before.

Why did students improve broadly in estimation? McGlothlen (2003) found that those who invoked a richer repertoire of analytic tools estimated better than those who used a more holistic/feeling approach. With this in mind, one explanation for the Experimental students' improvement is that the intervention moved them towards a more comprehensive approach to estimation. Our lab is conducting ongoing research to examine other possible causes for students' improved performance. One such possibility is that Experimental students enjoyed the curriculum and were simply more motivated than Controls to complete the posttest exercises. If this were the source of improvement, we would expect Experimental students to spend more time on solutions, and report more interest in the task, but we would not expect to see greater richness in the strategies they employ. Another possibility is that Experimental students benefited from the recency of their practice with estimation during the intervention. If this caused the difference between the groups, then, again, although Experimentals gave more accurate estimates, we would not expect subsequent analyses to show that they used richer strategies than Controls. We cannot reject this possibility at present, but we find it highly unlikely, as estimation curricula are generally quite taxing for students: Our prior results indicate that without a particularly engaging curriculum, more recent practice leads to a performance decrement, presumably due to fatigue.

General Discussion

Many propositions inform our social preferences (e.g., Ranney & Schank, 1998), but to illustrate the role played by numbers, consider whether your immigration preference would change if you made an estimate that was highly inaccurate. What sort of numerical feedback might call your assumptions into question, leading you to a different preference? Preferences are central to human cognition, and *numerical* preferences provide useful sources of evidence regarding conceptual change. Numerical preference represents a concrete way in which mathematics is relevant to our lives, and contributes to discussions of quantitative literacy in math education. By ignoring base rates, voters or political candidates may take stands that conflict with what they would otherwise prefer. Of course, some people take *absolute stances* on particular issues, such as completely eliminating abortion; as such, they imply that the numbers are irrelevant to their beliefs on the issue, and we would not expect them to change their preferences after feedback very often (Ranney et al., 2001). For those who indicated nonzero preferences, Munnich et al. (2003) found two main patterns: First, those who were less surprised by base rates generally *proportionately rescaled* their preferences—those who preferred halving the abortion rate initially, still preferred halving the actual rate when it was revealed. This suggests the base rate was belief-relevant, but that it did not inspire dramatic revisions of belief networks. Second, those who were more surprised by feedback showed *policy shifs*—accommodative belief revisions—for instance, those who preferred halving the abortion rate initially, but were surprised by the actual rate, indicated final preferences notably more or less than half of that rate (see Garcia de Osuna, Ranney, & Nelson, 2004, for more discussion of the qualitative nature of such shifts).

Even when considering the same issue, people can arrive at markedly different estimates and policies, depending on how the issue is framed (cf. Schwarz, 1999). As noted earlier, when Munnich et al. (2003) asked for the number of abortions per live births, the median response was 33.5 times too high. With their estimates so far off, what happened with these people's preferences? After feedback, they showed a policy shift—a 64% more reductive policy than they had initially indicated. By contrast, when participants estimated the number of abortions per fertile women, the median estimate was much closer—half the actual number. Rather than shift policies, for the fertile-women variant, participants merely rescaled their preferences to adjust to their new understanding of the number. In other words, when a quantity (e.g., the number of abortions performed each year) is framed in different ways, people show vastly different abilities in estimating the quantity, and this strongly affects their preferences after they learn the actual numbers.

Our hypothesis in this paper focused on the estimation side of NDI, but there are also implications for preference. When an intervention successfully fosters estimation ability, what might we predict, regarding people's preferences? One possibility is that as estimates improve, feedback-driven surprise will abate, and policies will stabilize, producing *less* subsequent policy shift. However, it is also possible that when estimates improve, people might become more sensitive to numbers, and attach more importance to small errors, yielding *more* policy shift**.** Note that while some of our past studies showed framing effects, they did not focus on people who recruited relevant facts to frame issues in different ways for *themselves*. When an individual integrates multiple constraints without prompting, the effects may be quite different than what we see with more passive participants. In analyses of the preference data gathered along with the estimation data reported above, we find support for both possibilities—while some participants appear to shift policies less after intervention, others seem to be more sensitive to small changes in numbers, and thus shift less. In aggregate these effects largely cancel each other out. A more in-depth analysis of individuals' changes in estimation ability, surprise levels, and preferences is being conducted to determine how each phenomenon contributes to the overall pattern of results. One possible benefit of this research may be in teaching people to construct policies that are less susceptible to rhetoric. That is, as people adopt more analytic strategies (assuming this is why estimates improve in our curricula), when they hear a quantity in advertisements or on the news, they might think of the issue several different ways and generate a preference that is constrained by other numbers they have considered.

Beyond transfer to tasks involving numerical understanding, what other forms of transfer might exist? NDI problems can be considered examples of "Fermi Problems," after the physicist who famously posed queries such as "How many piano tuners are there in Chicago?" Few, if any, can simply *recall* answers to Fermi questions, but through successive approximations and drawing on other known quantities, one can approach the correct answer. When Fermi questions are posed—often by potential employers or as classroom exercises—the implicit assumption is that one's answers are indicative of general analytic ability and creativity in problem solving. It is not difficult to imagine that NDI-type interventions might benefit reasoning about the location of Berlin relative to the former East-West border: With analytic techniques, one could do for oneself what Ranney et al. (1993) did for their participants—foster the integration of multiple, mutually constraining, perspectives into a solution.

More broadly, was Fermi's physics problem-solving ability related to his ability to estimate the number of piano tuners in Chicago? Much of the problem solving literature indicates little *general* transfer of problem solving skill across divergent domains (Singley & Anderson, 1989), so this may initially seem unlikely. However, we note that one of Ranney and Thagard's (1988) participants ("Pat") reached a more sophisticated understanding of projectile motion through the same kinds of processes that we have argued to underlie strong numerical reasoning. Pat initially believed that a ball dropped by a walking person would fall

straight to the ground. Later on in her verbal protocol, she contemplated the motion of a ball thrown obliquely upwards, and decided that it would follow an arc-shaped trajectory. Upon realizing this, it occurred to her that, from the zenith of its trajectory to the ground, the ball would descend analogously to a ball dropped while walking. Accordingly, she concluded that the two trajectories must have a similar arc-shape. Pat thus revised her view of the path of the dropped ball to a (more accurate) curved trajectory. This example illustrates the potential generality of the analytic skills that are useful in numerical reasoning: In both physics and estimation, we seem to benefit from using alternative representations, and then resolving conflicts among them. The degree to which one skill transfers to another is a worthy topic for future research.

Summary

It is critical that citizens and consumers be able to make decisions on numerically laden issues. We found that people can improve their numerical understandings through activities emphasizing the consideration of multiple perspectives and the integration of mutual constraints, and we discussed possible implications of such findings for individuals' policy stances. We propose that improvements in estimation abilities arose from an analytic approach that this intervention cultivated, leading students to seek evidence that might disconfirm their initial hunches. Such an approach might have value beyond the numerical and policy realms, with respect to more general reasoning and problem solving skills. In these ways, classroom interventions that test aspects of the emerging theory around the Numerically-Driven Inferencing paradigm have the potential to answer questions of fundamental interest to both cognitive science and society.

Acknowledgements

We thank the students and teachers who participated in this study, as well as Mandy Bachman, Morgan Curley, Christine Diehl, Barbara Ditman, Karen Draney, Sujata Ganpule, Cirila Howard, Josette Juan, Florian Kaiser, Lilian McGlothlen, Michelle Million, Janek Nelson, Luke Rinne, Mirian Song, Mark Wilson, and the UCB Reasoning Group for their helpful comments. This work was funded by a UCB faculty research grant and an AERA/IES Postdoctoral Fellowship.

References

Brown, N. & Siegler, R. (2001). Seeds aren't anchors. *Memory & Cognition, 49,* 405-412.

Curley, M. (2003). *An EPIC curriculum: An examination of a curriculum to promote reasoning for conceptual change.* Unpublished Master's Project, University of California, Berkeley.

Garcia de Osuna, J., Ranney, M., & Nelson, J. (2004). Qualitative & quantitative effects of surprise: (Mis)estimates, rationales, & feedback-Induced preference changes while considering abortion. *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Howard, C. (2003). *An EPIC Quest for Justification: The Effects of a Numerically-Based Intervention on Students' Estimates and their Justifications*. Unpublished Master's Project, University of California, Berkeley.

Huttenlocher, J., Hedges, L., & Prohaska, V. (1988). Hierarchical organization in ordered domains: Estimating the dates of events. *Psychological Review, 95*, 471-488.

Juan, J. (2003). *An EPIC curriculum with attitude: The extension of a novel curriculum involving estimations and attitudes about higher education.* Unpublished Master's Project, University of California, Berkeley.

Johnson-Laird, P. & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory & Cognition, 31(7),* 1105-1113.

Keysar, B. (1990). *East meets west at the Berlin wall: Mental maps and the changing world order.* Unpublished data.

McGlothlen, L. (2003). *High school students reasoning with numbers: Interviews using the estimate, predict, incorporate, and change (EPIC) method*. Unpublished Master's Project, University of California, Berkeley.

Munnich, E.., Ranney, M., Nelson, J., Garcia de Osuna, J., and Brazil, N. (2003). Policy shift through Numerically-Driven Inferencing: An EPIC experiment about when base rates matter. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society.* (pp. 834-839). Mahwah, NJ: Erlbaum.

Ranney, M., Cheng, F., Nelson, J., and Garcia de Osuna, J. (2001). *Numerically driven inferencing: A new paradigm for examining judgments, decisions, and policies involving base rates.* Paper presented at the Annual Meeting of the Society for Judgment & Decision Making.

Ranney, M. and Schank, P. (1998). Toward an integration of the social and the scientific: Observing, modeling, and promoting the explanatory coherence of reasoning. In S. Read & L. Miller (Eds.), *Connectionist models of social reasoning and social behavior*. Mahwah, NJ: Erlbaum.

Ranney, M., Schank, P., Mosmann, A., & Montoya, G. (1993). Dynamic explanatory coherence with competing beliefs: Locally coherent reasoning and a proposed treatment. In T.-W. Chan (Ed.), *Proceedings of the International Conference on Computers in Education: Applications of Intelligent Computer Technologies* (pp. 101-106).

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426-432). Hillsdale, NJ: Erlbaum;

Schwarz, N. (1999). How the questions shape the answers. *American Psychologist*, *54*, 93-105.

Singley, M. & Anderson, J. (1989) *Transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-502.

Workload is Bad, Except when it's Not: The Case of Avoiding Attractive Distractors

Christopher W. Myers, Wayne D. Gray, & Michael J. Schoelles
Cognitive Science Department
Rensselaer Polytechnic Institute
[myersc, grayw, schoem] @rpi.edu

Abstract

Increased cognitive workload is typically considered to hinder task performance. The current study presents an example where increased workload *aided* a visual search task. Increased workload, via a secondary task, provided participants extra time to avoid distracting stimulus configurations. Furthermore, initial fixations on distracting densities occurred at higher frequencies when initial saccades lasted less–than 400 milliseconds. We conclude that the combination of the primary visual search task and the secondary task create an environment where the secondary task was *beneficial* to the visual search task.

Introduction

There is a rich literature demonstrating how visual stimuli affect visual search patterns (Findlay, 1982, 1997; He & Kowler, 1991; McCarley, Kramer, & Peterson, 2002; Pomplun, Reingold, & Shen, 2003; Rayner, Liversedge, White, & Vergilino-Perez, 2003; Wolfe, Cave, & Franzel, 1989; Zelinsky, 1996). However, few studies have focused on how *stimulus configurations* influence eye movements. An example of a stimulus configuration is differences in inter-stimulus distance, or *density*. Stimulus density can be easily manipulated. Increasing the inter-stimulus distance decreases density, and vice-versa. There is also little research describing the effects of increased workload on visual search. Do visual search strategies change as a function of workload? In this paper, we address workload and stimulus configuration effects on visual search.

Previous research suggests that saccades are programmed and targeted in an automatic, data-driven fashion. Data-driven processes shape overt behavior via environmental factors, and are typically considered unconscious processes. There are two striking examples that suggest data-driven processes determine saccadic endpoints. The first example is the *global effect* (Findlay, 1982, 1997). The global effect occurs when saccadic endpoints land at intermediate target positions during abrupt onset tasks containing at least two stimuli. That is, when two stimuli appear to the right or left of an initial fixation point, saccadic endpoints tend to be located between the stimuli. The global effect provides evidence that global target configurations influence saccadic amplitude. It appears that saccadic processes use stimulus attributes such as spatial properties in determining endpoints.

The second example is the *center–of–gravity effect* (He & Kowler, 1991). The center–of–gravity effect indicates that saccades directed toward a shape land at consistent locations near the center of the shape, and that the shape's contour information is all that is necessary for consistent saccades. The two effects taken together suggest that the saccadic mechanism relies on spatial properties of stimuli when determining saccadic endpoints.

The amount of influence deliberate, top-down strategies have on saccadic endpoint location is still unclear. However, it is unlikely that humans solely rely on purposeful, top-down strategies when determining saccadic endpoints. He and Kowler (1991) propose a serial, two-stage process for determining saccadic endpoints that incorporates both automatic processes and intentional strategies. The two-stage process involves an initial intentional target selection, followed by an automatic weighted averaging of the shape or stimuli to determine the saccadic endpoint.

Shen, Reingold, and Pomplun (2000) demonstrated that in a conjunctive search task visual search is also affected by the cost structure of the search environment. When few same-color distractors were present, saccadic selectivity was biased towards color. However, as the number of same-color distractors increased, saccadic selectivity shifted from same-color to same-shape stimuli. This suggests that visual search may be sensitive to the soft constraints of the search space. Hard constraints arise from the types of stimuli built into the search environment, and the types of interactive behavior permitted (such as searching by color or shape). Hence, hard constraints determine which microstrategies are possible (Gray & Boehm-Davis, 2000). In contrast, soft constraints determine which of the possible microstrategies are *most likely* to be selected (Gray & Fu, 2004). When selection is non-deliberate or automatic the least effort microstrategy is chosen. Searching same-color targets when they are the majority distractor leads to higher movement latencies, higher manual response times, and more fixations than searching the minority, same-shape distractors (Shen et al., 2000).

Our research has focused on where a participant is likely to initially fixate. Initial fixations are the dwells located at the endpoint of the initial saccade. This work has uncovered an effect of stimuli density on initial fixation locations, or the *pro-density effect* (Myers, Gray, & Schoelles, 2003, 2004). As the density of stimuli increases (inter-stimulus distances become smaller), the probability of initially fixating the dense group also increases. Our work in conjunction with Shen et al. (2000) makes it apparent that stimulus features are not the only aspects of the search space considered. Rather, we have found that stimulus configurations are also important. It is likely that the results of Shen et al. (2000) and Myers et al. (2003; 2004), are solely attributable to neither data–driven nor purposeful,

top–down search strategies, but are attributable to a combination of both processes.

If the determination of saccadic endpoints and initial fixation locations are not the exclusive result of top-down, purposeful processes, then taxing top-down processes in order to eliminate any purposeful search strategy will allow automatic, data-driven processes greater influence in overt behavior. Our research has demonstrated that the pro-density effect is heightened with increased workload. The pro–density effect *doubled* with an added auditory task (Myers et al., 2004). This result suggests a data–driven component when determining where to initially fixate or saccade.

In the studies conducted by Myers et al. (2003; 2004), target and density locations were completely orthogonal; as a result, dense clusters of stimuli provided no useful information of the target's whereabouts. Therefore, there was no incentive to *avoid* dense clusters of stimuli.

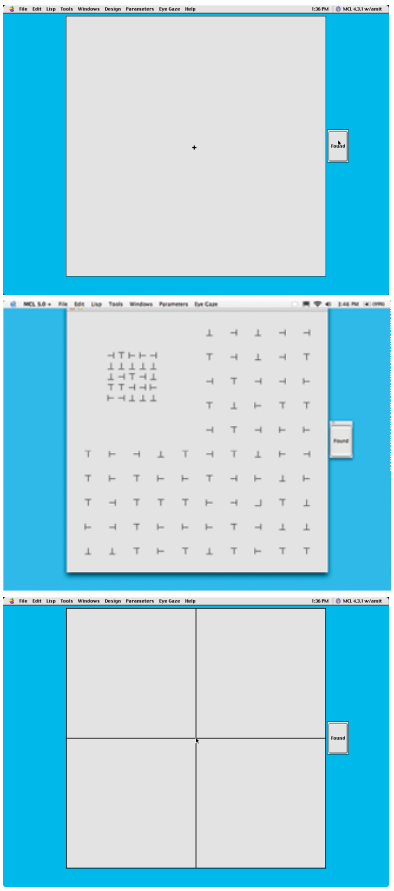
Having found a pro-density effect in previous work, the current study attempted to determine the robustness of this effect by establishing a negative correlation between dense clusters and the probability of a target being located in a dense cluster. If initial fixations still land on the dense cluster, this would suggest that the effect is determined by low–level, bottom–up process that are drawn to certain configural properties. On the other hand, if initial saccades resist the dense cluster or show an *aversion* to the dense cluster, this "anti-density" effect would suggest target location information provided by dense clusters might be incorporated into a conscious, top–down strategy such as deliberately avoiding the dense cluster. However, Myers et al. (2003; 2004) demonstrate that dense clusters are initially fixated more than chance and the number of initial fixations increases with the degree of density and added workload. Therefore, a dense cluster of stimuli is an *attractive distractor* in the current study. Workload was manipulated between participants as a dual task condition and a single task condition. Participants in the dual task condition performed two tasks simultaneously, thereby increasing cognitive workload. The single task group performed one task.

If participants were able to resist initially fixating a dense cluster, the pro-density effect would drop to at most chance levels in the single task group. This would suggest that deliberate processes are overriding the influence of unintentional, data-driven processes on overt behavior in the task environment. We also predicted no effect of degree of density (moderate vs. strong) in the single task group. For the dual task group, we predicted an increase in the pro-density effect as demonstrated in Myers et al. (2004). This would suggest that data–driven processes begin to peer through deliberate strategies in dual task, high load situations. We did not predict the pro-density effect to increase two–fold, rather that it would increase to levels significantly greater than chance. Finally, we predicted that there would be a significant effect of degree of density in the dual task condition, specifically that strong densities would be initially fixated more often than moderate densities.

Method

Participants

A total of thirty-three undergraduate students volunteered to participate. All participants had normal, or corrected–to–normal vision. Participants were randomly assigned to one of two groups. The single task group performed a visual search task, and the dual task group simultaneously performed the same visual search task and an auditory letter classification task. There were 16 participants in the single task condition and 17 participants in the dual task condition. The study lasted approximately 1 hour, and participants were run individually.



1. Fixation
2. Stimulus
3. Test

Figure 1. One visual search task trial, presented in order of from top to bottom.

Apparatus

The data collection apparatus consisted of a PowerMac G4 Apple computer running MacOS Jaguar, a 17-inch flat panel display with the resolution set to 1024 x 768, a chinrest, and an Eyegaze eye-tracking system developed by LC Technologies that measured gazepoint at a 60Hz rate.

Visual Search Task

The visual search task was composed of three different displays, each presented sequentially and in a fixed order. Each display was composed of a window and a *Found* button. An example of the task is depicted in Figure 1. The initial display was composed of a single cross hair located in the middle of the window. The cross hair was used as an initial fixation point for each trial (top, Figure 1). Once the eye tracking system determined cross hair fixation, the stimulus display appeared. The stimulus display (middle, Figure 1) contained a target (e.g., L) and distractors (e.g., T) that were randomly rotated about their axes on each trial. The stimulus display consisted of a 10x10 stimuli matrix, enabling the display to be divided into four equal quadrants of stimuli. The L was placed at the center of a randomly chosen quadrant that did not contain a dense cluster of stimuli.

The within-subject independent variable, stimuli density, varied on 3 levels: strong, moderate, and weak with an inter-stimulus distance of 0.54°, 0.97°, & 1.94° of visual angle, respectively. Each density level occurred on 33% of all trials, with quadrant location randomized for each trial. Quadrants not containing a dense cluster had an inter-stimulus distance equal to weak. (Hence, on weak density trials, all four quadrants were of equivalent density.) The L was *never* located in a dense cluster of stimuli.

Participants were instructed to find the target as quickly as possible and were aware there were only four possible target locations. On target discovery, participants clicked the 'found' button. After clicking 'found', the test display appeared (bottom, Figure 1). The test display was divided into four visible quadrants with the mouse pointer located at the quadrants' intersection. Participants were instructed to click on the quadrant where the target was discovered. Once the participant clicked on a quadrant, the pointer was automatically relocated to the 'found' button and the fixation display reappeared, beginning a new trial. Each participant performed 4 blocks of 48 trials.

Auditory Letter Classification Task

Participants in the dual task condition were acoustically presented random letters of the alphabet in four-second intervals via the speaking software *Victoria*, developed by Apple™. For each letter presented, the participant pressed X if the current letter came before the previous letter or C if it came after. For example, if the subject heard 'A' followed by 'G' she would press C signifying 'G' occurs *after* 'A' in the alphabet. If after four more seconds 'B' was presented, she would press X signifying 'B' occurs *before* 'G'. Letter presentation occurred every four seconds throughout each block of 48 trials. Participants were instructed to simultaneously perform both the visual search task and the letter classification task to the best of their ability. Each dual task participant received accuracy feedback on the letter classification task at the end of each block. No subject scored below 85% accuracy in the last three blocks. Single task participants did not participate in the letter classification task.

Dependent Measures

Our dependent measure was the initial fixation location for each trial, where the initial fixation is the second dwell on the stimuli display. The first dwell was a *residual fixation* resulting from fixating the cross hair on the fixation display. Fixations were determined using the eye tracking software's default fixation analysis. Initial saccades were defined as the eye movement from the residual fixation to the initial fixation.

Results

Comparisons are between the *actual* number of initial fixations on a dense cluster and the number expected by chance. Since there were four possible quadrants to fixate within, there is a 25% chance that an initial fixation would occur on the dense cluster. All t-tests reported are two-tailed and measured at a 0.05 significance level. The first block was removed to reduce any variance attributable to task familiarization. Trials in which all four quadrants were of equivalent density (weak density trials) were excluded from the analyses.

Single Task Condition

The dense cluster was initially fixated on 23.63% of the trials when a dense cluster was present. This rate of initial fixation does not differ from chance [$t(15) = -0.71$; $p = 0.485$]. The planned comparison of degree of density (moderate vs. strong) was not significant ($p = 0.95$). This supports our hypothesis that for the single task there would be no pro-density effect.

Dual Task Condition

The dense cluster was initially fixated on 17.44% of all trials. The rate of initial fixation significantly differs from chance [$t(16) = -2.88$; $p = 0.01$]. There was a marginally significant effect between degrees of density [$t(16) = 1.962$; $p = 0.067$] when comparing moderately dense clusters ($M = 15.14$, $SE = 2.51$) to strongly dense clusters ($M = 19.86$, $SE = 2.72$).

In the dual task condition, we predicted a positive effect of density on initial fixation locations. Instead, we found a negative effect. Initial fixations on dense clusters of stimuli, under dual task conditions, were less than would be expected by chance. We term this the *anti-density effect* and explore it in the following sections. We also found the strong density was initially fixated more often than the moderate density.

Initial Saccade Latencies (ISLs)

Initial Saccade Latencies were defined as the amount of time the participant continued to fixate the cross hair location *after* the stimuli display appeared. The eye tracker used in the study sampled the eye position every 16.67

milliseconds. Each sample in the residual fixation was counted and multiplied by 16.67 in order to determine each subject's ISL for each trial.

ISL Analyses

Before analyzing the ISL data we removed outliers from the data set. Outliers were identified for each group by calculating the mean and standard deviation for each group and removing any data point that exceeded the mean by +/– 3 standard deviations. This procedure resulted in removing 27 data points from the single task group and 41 data points from the dual task group. All blocks were included.

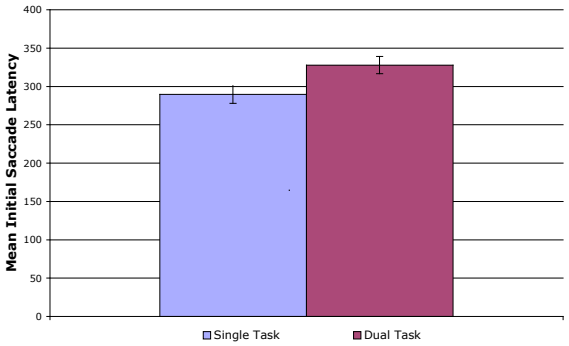


Figure 2. Comparison between dual and single task ISLs. The error bars represent standard error.

An independent groups t-test was performed between the dual and single task conditions on mean ISL. The dual task group had longer ISLs on average (M = 327.79, SE = 11.5) when compared to the single task group (M = 289.53, SE = 11.3), and this difference was significant, $t(31) = 2.37$; p = 0.024 (see Figure 2). This result signifies that there was a difference between the groups average ISL.

Discussion of Results

The single-task manipulation worked as predicted: the density effect occurred at chance levels. However, our dual-task manipulation exhibited an anti-density effect. This result is quite startling in the face of previous research that consistently demonstrated dense clusters *attracting* initial fixations (Myers et al., 2003, 2004). We also found a marginally significant effect of degree of density in the dual task condition.

When unintentional processes associated with dense clusters are not producing an effect, dense clusters should only be initially fixated at chance levels. However, if participants were using implicit information provided by the dense cluster (that the target was not located there), then participants should avoid dense clusters. However, 23.63% of initial fixations are located on a dense cluster of stimuli. The results do not support dense cluster avoidance for the single task. In the single task condition it is apparent that the unintentional attraction of dense clusters has been overridden by a different, possibly deliberate, strategy.

Perhaps participants learned to avoid the dense cluster in the single task condition, but were unable to reduce the effect below chance levels. However, in the dual task condition, the pro-density effect is reduced to below chance levels (17.44%, depicted in Figure 3). This was a significant reduction from chance, and suggested the dense cluster was avoided. It is apparent that something was aiding participants to avoid dense clusters in the dual task group. Initial saccade latencies provided a clue. Due to longer ISLs, participants might have more time to implement a conscious, deliberate strategy. We explore this possibility in the upcoming sections.

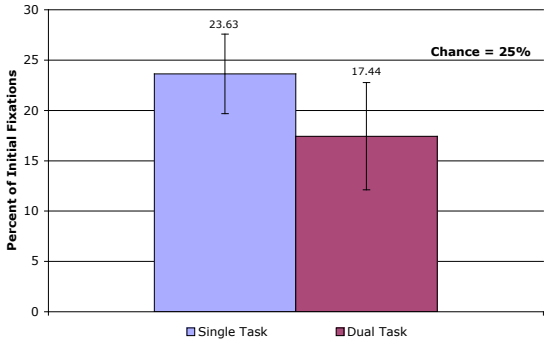


Figure 3. Effects of density by task compared to chance, error bars indicate 95% confidence interval.

Post-Hoc Analyses

Our initial analyses suggested the letter classification task aided dual task participants. Extra time might allow participants to avoid dense clusters. Added time could result from performing some aspect of the letter classification task at stimuli display onset, such as making a comparison, pressing a keyboard key, or even retrieving a memory of the previous letter presented. Extra time was apparent in participants' initial saccade latencies (ISLs). Specifically, dual task participants had significantly longer ISLs compared to single task participants. Further analyses were performed to determine if short ISLs led to a stronger pro-density effect and longer ISLs led to a weaker pro-density effect (anti-density effect). All four blocks were included in the analyses.

To determine if the pro-density and anti-density effects occurred at different rates for different ISLs, we divided our data into 5 bins. Each bin spanned 150 ms, and ranged from 100 ms to 550+ ms. The number of pro-density initial fixations was derived for each bin and divided by the total number of initial fixations for the same bin, creating a percent of pro-density fixations (see Figure 4).

Figure 4 shows a reduction in the pro-density effect as ISLs increase in duration. Figure 4 also shows that single task participants followed the same general trend. The separation between the single task curve and the dual task curve is a result of the single task having a greater number of pro-density fixations.

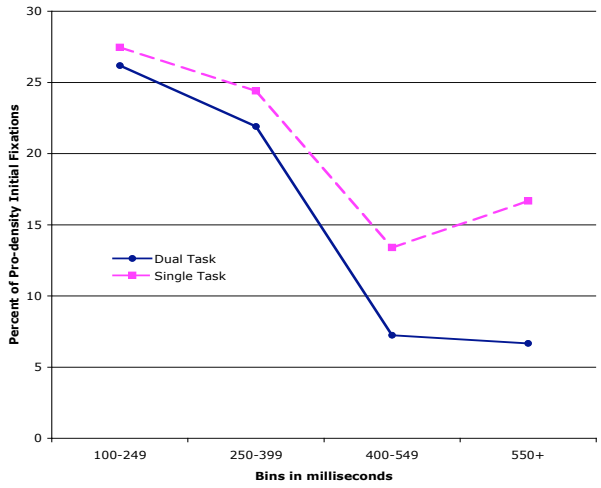


Figure 4. Pro-density effect as a function of initial saccade latency, by task.

It is important to note that the 550+ bin does not contain much data, especially for the single task condition. In fact these data points may be considered aberrant for the single task condition, however they fell within the outlier cutoff. Very few single task participants had ISLs that were 550 ms or greater, as demonstrated in Figure 5.

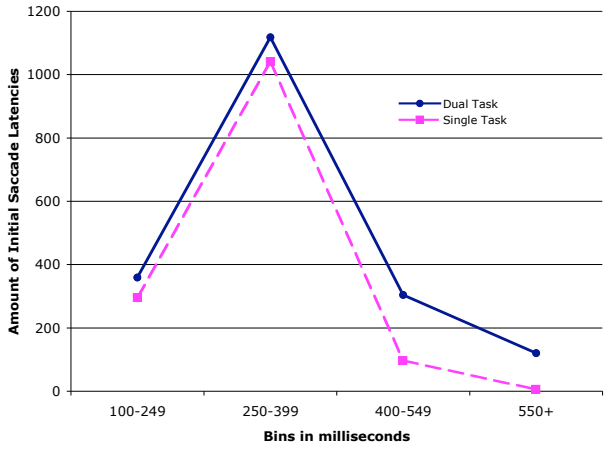


Figure 5. Frequency of initial saccade latencies for each bin, by task

Figures 2 and 4 provide evidence that there was a difference in mean ISLs between the dual and single task conditions. Figure 4 demonstrates that as ISLs increased, the likelihood of initially fixating a dense cluster reduced dramatically. In order to test for significance, a 2 (single task, dual task) x 3 (bins 100–249, 250–399, & 400–549) repeated measures ANOVA was performed. The 550+ bin was removed from the analysis due to insufficient data in the single task group, as demonstrated in Figure 5. Results indicate that there is a significant interaction [$F(1,2) = 3.292$; $p = 0.045$)] between the presence of the letter classification task and ISLs (see Figure 6). There was also a main effect of ISL on the percent of initial fixations on a dense cluster [$F(1,2) = 31.275$; $p < 0.0001$)].

Summary and Discussion

The results of our post-hoc analyses revealed a surprising effect. Participants were *aided* in the dual task condition via the auditory task. Generally, aid came in the form of not initially fixating distracting dense clusters. As a result of increased ISLs, the visio-cognitive system gained the opportunity to acquire and use information relevant to the task at hand. This analysis suggests that low-level strategies are chosen based on the soft constraints inherent in the task environment (Gray & Fu, 2004).

In previous research (Myers et al., 2003, 2004) dense clusters were *uninformative* and we found that dual task situations increased the pro-density effect. In the current study, dense clusters were made *informative* and the information was somehow used in a beneficial manner. When dense clusters are uninformative, they should always be considered as a possible target location. However, when a dense cluster provides target location information, then it becomes possible to reduce your search costs, and is akin to differences in saccadic selectivity as a function of distractor ratios discussed by Shen et al. (2000). Costs attributable to the current experiment's search space begin at very low levels. However, when the opportunity arose to reduce cost, providing benefit by reducing the search space, both dual *and* single task participants seized the opportunity. This occurred at greater rates as ISLs increased. It appears that the visio-cognitive system is extremely sensitive to cost-benefit tradeoffs, even when the cost is an average of one extra fixation. Our data suggests a limit: enough time must be provided in order to achieve a reduction.

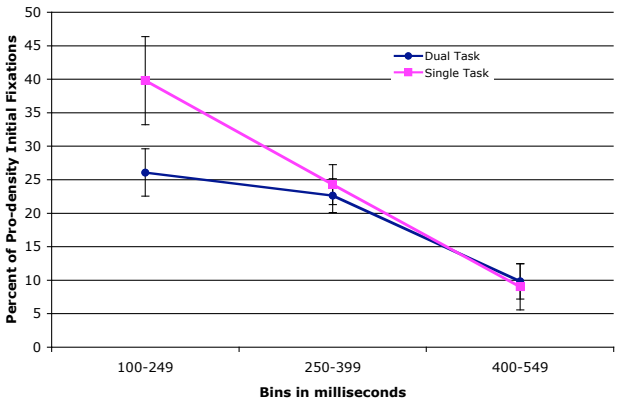


Figure 6. Interaction of initial saccade latency and task condition. Error bars represent standard error.

The reasons for and ways in which purposeful, top-down strategies interact with automatic, data-driven processes are unclear; however our analyses do shed some light. The data

suggest that the pro-density effect is an automatic, data–driven process. This is attributable to short ISLs leading to high percentages of initial fixations on dense clusters. This was observed in both task conditions when ISLs were relatively short (see Figure 6). Dense clusters were avoided on roughly 90% of all trials in both task conditions when ISLs were ≥ 400 ms. This suggests more time is necessary to impose deliberate, top–down strategies. In addition, dense clusters were initially fixated more than chance levels for short ISLs. This indicates that automatic, data–driven processes have more influence in overt behavior when ISLs were relatively short, and as ISLs increased the ability to impose top–down strategies on the search task increased. Soft constraints theory suggests that after initially adopting a least-cost strategy, such as avoiding the dense cluster, the number of initial fixations on dense clusters should be reduced as the new microstrategy gains in success over time. It is likely that participants do reduce initial fixations from block 1 to block 4 for all ISL bins, and this reduction would be an example of a learned, unconscious, data-driven strategy.

Further support comes from the planned comparisons between moderate and strong densities in the dual task condition. When top-down processes are sapped by added workload, there are differences between strong and moderate densities. However, there was no difference in the single task condition. When deliberate top–down strategies are taxed, it is likely that bottom-up processes have an opportunity to exert more control on overt behavior.

Pomplun, Reingold, & Shen (2003) have developed a computational model (Area Activation Model) that predicts where saccadic endpoints will be located. These locations are based on information in the search space such as color and shape. The current study points to areas in the model where more work is needed; namely, that stimuli configuration and cognitive workload are important aspects of visual search that must be considered when developing models of saccadic selectivity.

Although the experiment revealed surprising effects, we did not design the experiment with these effects in mind. We see this as a possible limitation in our study and feel that more studies such as the one presented here need to be completed to understand the true nature of the effects that stimuli configurations and high levels of cognitive workload have on visual search.

Acknowledgments

The work reported was supported by a grants from the Air Force Office of Scientific Research AFOSR #F49620-03-1-0143, as well as the Office of Naval Research ONR #N000140310046.

References

Findlay, J. M. (1982). Global visual processing for saccadic eye movements. *Vision Research, 22*, 1033-1045.

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research, 37*(5), 617-631.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6*(4), 322-335.

Gray, W. D., & Fu, W.-t. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28*(3).

He, P., & Kowler, E. (1991). Saccadic localization of eccentric forms. *Journal of Optical Society of America: A, 8*(2), 440-449.

McCarley, J. S., Kramer, A. F., & Peterson, M. S. (2002). Overt and covert object-based attention. *Psychonomic Bulletin & Review, 9*, 751-758.

Myers, C. W., Gray, W. D., & Schoelles, M. (2003). *This way or that: Determining where to look first.* Poster presented at the 25th Annual Meeting of the Cognitive Science Society, Boston, MA.

Myers, C. W., Gray, W. D., & Schoelles, M. (2004). *The effects of stimulus configuration and cognitive workload on saccadic selectivity*. Poster presented at the 4th Annual Meeting of the Vision Sciences Society, Sarasota, FL.

Pomplun, M., Reingold, E. M., & Shen, J. (2003). Area activation: a computational model of saccadic selectivity in visual search. *Cognitive Science, 27*, 299-312.

Rayner, K., Liversedge, S. P., White, S. J., & Vergilino-Perez, D. (2003). Reading disappearing text: cognitive control of eye movements. *Psychological Science, 14*(4), 385-388.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. *Perception, 29*, 241-250.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419-433.

Zelinsky, G. J. (1996). Using eye saccades to assess the selectivity of search movements. *Vision Research, 36*(14), 2177-2187.

Paying Attention to Attention: Perceptual Priming Effects on Word Order

Rebecca Nappa (nappa@sas.upenn.edu)
Department of Psychology, 3401 Walnut St.
Philadelphia, PA 19104 USA

David January (djanuary@sas.upenn.edu)
Department of Psychology, 3401 Walnut St.
Philadelphia, PA 19104 USA

Lila Gleitman (gleitman@sas.upenn.edu)
Department of Psychology, 3401 Walnut St.
Philadelphia, PA 19104 USA

John Trueswell (trueswel@sas.upenn.edu)
Department of Psychology, 3401 Walnut St.
Philadelphia, PA 19104 USA

Abstract

Two experiments are reported which examine how manipulations of visual attention affect adult speakers' linguistic choices regarding word order and verb use when describing simple visual scenes. Participants in Experiment 1 were presented with scenes designed to elicit the use of one of two perspective verbs (e.g., "A dog is chasing a man"/"A man is running from a dog"). Speakers' visual attention was manipulated by preceding the display with a crosshair positioned on one or the other character. Cross-hair position affected word order and verb choice in the expected direction. Experiment 2 replicated this effect with a subliminal attention-capture cue, and results were further extended to the order within conjoined noun phrases in sentential subjects ("A cat and dog are growling…"). The findings have important implications for incremental theories of sentence planning and suggest some specifics for how joint-attention might serve as a useful cue to children learning verbs.

Introduction

What makes people say what they say? This is a complex question, which has been the source of much investigation and dispute over the past several decades. Early on in the generative linguistic tradition, the emphasis on the productive and creative power of structural expression led many researchers to assume that properties of a visual stimulus can be related to a speaker's linguistic choices in only vague and theoretically uninteresting ways (e.g., Chomsky, 1957). Currently, though not disputing that one can say – or not say – many different things under the same environmental conditions, investigators doing experimental research on word order and structural choices in sentence production have concluded that some combination of perceptual, conceptual and linguistic accessibility contribute in a dynamic way to utterance planning. In particular, questions of word order have received much attention, and prompted much debate, as this issue must be richly intertwined with the planning of both an utterance's overarching message and the syntactic structure carrying that message. In advance of speaking, one must somehow decide where the upcoming utterance is to start, and much of how it is to proceed. A number of factors seem to contribute to this process.

First, studies have found a crucial role for preferred (i.e. primed or otherwise accessible) syntactic structures in the form a message ultimately takes (e.g. Bock & Loebell, 1990). Additionally, lexical/conceptual factors (e.g. accessibility, animacy) have been shown to affect word order materially, even at the expense of a preferred syntactic structure (Tversky, 1977; Bock, 1986; MacDonald, Bock & Kelly, 1993).

The role perceptual prominence plays in word and/or constituent order, however, seems a bit more nebulous. Within the literature on visual attention, it is quite clear that perceptual cues are involved in the interpretation of visual stimuli; research on perception of ambiguous figures (e.g. duck/rabbit, wife/mother-in-law) has shown that the perception of such stimuli can be driven by localizing eye gaze on critical features of a given interpretation (Georgiades & Harris, 1997). And perceptual factors (e.g. size, color) are clearly involved in ordering within simple conjoined noun phrases (e.g. *A bear and a dog*) (Osgood & Bock, 1977; Gleitman, Gleitman, Miller & Ostrin, 1996), but the role of perceptual prominence in constituent order remains unclear. Some find no relationship between initially fixated stimuli and subject role assignment (Griffin & Bock, 2000), while others find evidence supporting a role for attention (perceptual prominence) in constituent order (Tomlin, 1997; Forrest, 1996).

Some have interpreted these latter results as evidence for an incremental account of language production, in which a speaker builds an utterance as it is produced, and is apt to begin with whichever sentential elements are most salient at the time of speech onset. This account contrasts with more structuralist version of sentence planning, in which the underlying message of an utterance must be wholly planned prior to the onset of speech, and which accounts for the robust and reliable effects of syntactic priming (see Bock, in press, for discussion).

A troubling issue with all prior investigations into perceptual prominence and word ordering arises, however, if one examines the methodology. Manipulations have all been overt attention-getting devices (raising demand characteristic concerns), and have often had rigid task demands allowing for minimal generalization.

The current research investigates the question of perceptual contributions to word and constituent order, drawing on the attention and perception literature for more suitable methods. In two experiments, subjects' attention was directed subtly (Experiment 1) and then subliminally (Experiment 2) to scene participants, to determine whether perceptual cues under these covert conditions have any effect on the linguistic choices speakers must make. If such perceptual factors lead subjects to differing descriptions of identical scenes, a clear role can be established for attentional factors in sentence planning and constituent order. Such results may also provide evidence for an incremental approach to production, or perhaps, rather, to message planning. Finally, as we describe later, these effects may rebound on aspects of word learning.

Experiment 1

In the spirit of the afore-mentioned perceptual attention research on ambiguous figure resolution, our first investigation of attentional effects on event interpretation used a simple crosshair fixation point – prior to stimulus presentation – to direct a subject's eye gaze to a scene participant (analogous to directing gaze to a set of critical features in the ambiguous figure literature). Stimuli were designed to elicit one of two word order and verb choices on the part of the speaker, thereby making one or the other character in a scene the subject of the sentence. If initial visual attention subtly alters a speaker's perspective on the scene, we should expect that the speaker's choice of sentential subject and verb would be influenced by our attentional manipulation.

Methods

Norming and Stimuli Prior to initiating data collection on an attention-manipulating task, the specific stimuli to be used were normed, to identify baseline rates of verb selection for these particular items. Twenty-one monolingual English-speaking University of Pennsylvania Intro Psychology students participated for course credit. Subjects were presented with the 52 pictures to be used in experiment one, and asked to describe the event that was taking place in the scene using a simple sentence. No other manipulations or cues were introduced. Of these 52 pictures, twelve depicted pairs of so-called Perspective Verbs (e.g. chase/flee, see Figure 1), and these were the critical items (PVs).

Rates of verb use for these twelve items varied (see Table 1), but for each verb pair, subjects showed some degree of bias towards one interpretation and/or verb choice; there was a preferred verb and a dispreferred verb, and hence a corresponding preferred subject and dispreferred subject (passives were rare, occurring only 6 times across all 252 items). Overall, preferred subjects and verbs were used 69% of the time, dipsreferred subjects and verbs were used 27% of the time.

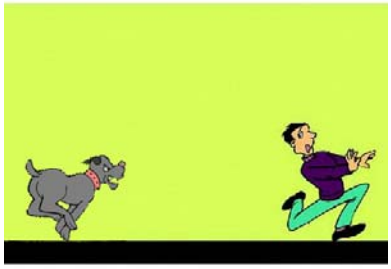


Figure 1: Sample Perspective Verb item from Experiment 1.

Table 1: Norming study baseline rates of verb usage in PV stimuli in Experiment 1. Percentage of total usage across all utterances in parentheses.

Item	Preferred Verb	Dispreferred Verb
Buy/sell	Sell (62)	Buy (38)
Chase/flee (dog/man)	Flee (57)	Chase (43)
Chase/flee (rabbit/elephant)	Chase (71)	Flee (29)
Eat/feed (puppies/dog)	Feed (76)	Eat (24)
Eat/feed (child/mother)	Feed (95)	Eat (5)
Give/receive	Give (71)	Receive (29)
Listen/talk (office)	Talk (76)	Listen (24)
Listen/talk (phone)	Talk (19)	Listen (29)
Perform/watch (singer)	Perform (67)	Watch (33)
Perform/watch (speaker)	Perform (86)	Watch (14)
Win/lose (boxing match)	Win (95)	Lose (5)
Win/lose (race)	Win (48)	Lose (24)

Participants and Design Eighteen monolingual English-speaking Introductory Psychology students at the University of Pennsylvania participated in this study for course credit. There were three conditions, defined by the location of the crosshair fixation point prior to scene presentation: Dispreferred (where the dispreferred subject would appear), Preferred (where the preferred subject would appear), and Middle (a neutral middle region, as a control). Manipulations were within-subjects, with each subject's gaze directed to the dispreferred subject on four of the twelve critical items, to the preferred subject on four of the

twelve critical items, and to a neutral middle region on the remaining four items.

Procedure Subjects in this experiment were presented with 52 scenes depicting participants engaged in a given activity (e.g. a picture of a boy swimming), including the twelve critical items, depicting perspective verb pairs. Subjects were instructed to describe each picture using one simple sentence, and subjects' utterances throughout the task were recorded.

A crosshair fixation point preceded presentation of each of the 52 scenes. This fixation point was presented on-screen for approximately 500 msec, then immediately followed by presentation of the scene (either filler or trial). (Earlier pilot work with an eyetracker confirmed that subjects followed directions and routinely fixated the cross prior to stimulus presentation.) Subjects in the current study were misled to believe that position of the crosshair was random and irrelevant to their task, so as to prevent their eyes from inspecting scenes in the same fashion on each trial. Position of the crosshair in fact corresponded directly to position of an upcoming scene participant. Although some subjects noted that the fixation marker frequently had been where an object appeared, no subject reported noticing the correlation between the location of scene participants and the crosshair. And in post-experimental interviews, most subjects who bothered to posit a guess as to the experiment's purpose speculated that it pertained to color brightness and/or interpersonal relationships of scene elements.

Results and Discussion

Rate of preferred verb usage was highly influenced by cue location in the expected direction (see Figure 2). In particular, when the preferred subject (e.g., the dog) was visually cued, speakers uttered on 77% of the trials sentences like "A dog is chasing a man." When the dispreferred subject was cued, however, speakers produced such utterances only 61% of the time (and showed a corresponding increase in utterances like "The man is running from the dog"). Analyses of Variance (ANOVAs) on participant and item means revealed that the effect of cue location was significant (both p's<.05).

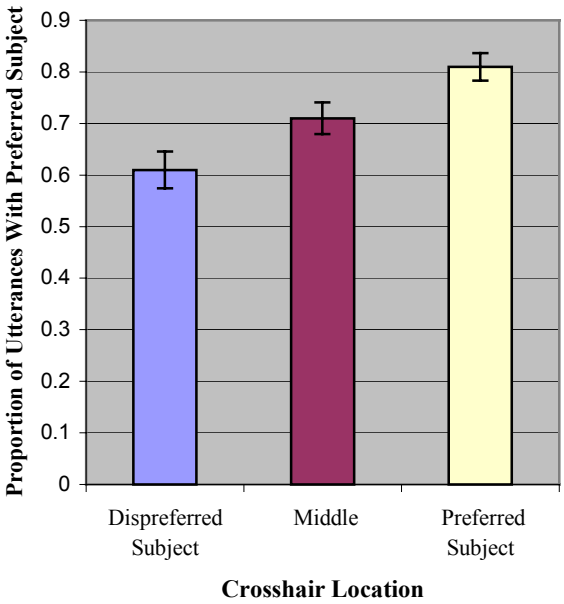


Figure 2: Proportion of utterances beginning with the preferred subject and verb, by condition, in Experiment 1.

Experiment 2

Following the results of Experiment 1, a couple of questions arose. First were concerns regarding demand characteristics of the crosshair fixation point manipulation. Although subjects did not seem to sense the specific purpose of the experiment (namely, subject and verb selection), many noticed that the crosshair's position frequently corresponded to an object in the upcoming scene. We worried that this knowledge alone might have subtly influenced their linguistic choices. To this end, we developed an attention-capture cue (see Jonides & Yantis, 1988), as discussed below, which successfully directed subjects' attention to a particular region of the scenes, without being consciously perceptible.

Secondly, as discussed previously, much prior research on sentence production and linguistic choice has compared the role of many different factors, from animacy to size of entities, on differing constructions. Specifically, different variables seem to contribute differently to linguistic choice in simple conjoined noun phrases (e.g. *the dog and the man* vs. *the man and the dog*) than to linguistic choices involving thematic role assignment (e.g. *the dog chased the man* vs. *the man fled the dog*). In Experiment 2, we wanted not just to replicate our prior result, but also to compare the influence of our covert attention-capture manipulation on these sorts of different constructions. To this end, twelve additional items were added in Experiment 2, depicting events in which two scene participants were engaging in an activity together (see Figure 3, designed to elicit "The cat/dog and the dog/cat are growling at each other"). These Conjoined Noun Phrase (CNP) items were aimed at eliciting descriptions containing a conjoined noun phrase in the sentential subject position (e.g. A dog and a cat are growling), so as to investigate the effect of our covert manipulation on word order in a simple conjoined noun phrase.

Figure 3: Sample Conjoined Noun Phrase item from Experiment 2.

Methods

Norming and Stimuli CNP pictures were first normed for baseline preferences. Twenty-one monolingual English-speaking University of Pennsylvania Intro Psychology students described 64 scenes (52 fillers, and the 12 CNP items), absent any manipulations or cues. In an effort to avoid utterances beginning with uninformative sentential subjects (e.g. "Two people are…"), CNP stimuli consisted of scenes with animal, rather than human, participants (e.g. a dog and a cat growling, see Figure 3).

Baseline rates of first-mentioned scene participants varied less than with the PV stimuli in Experiment 1 (See Table 2). Most items were relatively unbiased, but scene participants with even a slight advantage were dubbed *Preferred First-Mentioned*, and referred to as such from this point onward, for the sake of simplification. Overall, preferred first-mentioned participants were mentioned first 56% of the time, dispreferred first-mentioned only 44%.

Table 2: Norming study baseline rates of first-mentioned participants in CNP stimuli in Experiment 2. Percentage of total usage across all utterances in parentheses.

Item	Preferred First-mentioned	Dispreferred First-mentioned
Biking	Turtle (61.9)	Dog (38.1)
Dancing	Fish (57.1)	Bear (42.9)
Eating	Koala (52.4)	Panda (47.6)
Growling	Cat (52.4)	Dog (47.6)
Juggling	Elephant (52.4)	Seal (47.6)
Jumping	Frog (57.1)	Cat (42.9)
Playing cards	Pig (57.1)	Dog (42.9)
Playing horns	Rhino (52.4)	Snail (47.6)
Rowing	Bear (52.4)	Snowman (47.6)
Skating	Monkey (57.1)	Rabbit (42.9)
Swinging	Elephant (61.9)	Monkey (38.1)
Waiting	Penguin (52.4)	Deer (47.6)

An additional consideration that arises when adding the CNP stimuli is orientation. As previously mentioned, one factor driving word order in conjoined noun phrases is the left-to-right bias, with leftmost participants more likely to be first mentioned. This prediction bore out in the current norming study as well, with leftmost participants mentioned first 78.2% of the time for CNP items (as compared to only 52.8% of the time for PV items in prior norming study).

Participants and Design Forty monolingual English-speaking Introductory Psychology students at the University of Pennsylvania participated in this study for course credit. Both the location of the attention-capture cue and the left-to-right orientation of the scene were systematically varied, creating a 2 X 2 design (cued participant X leftmost participant) and four stimulus lists. Manipulations were within-subjects, with each subject assigned randomly to one of these four lists.

Procedure Subjects in this experiment were presented with 64 scenes: the same 40 fillers and 12 PV scenes used in Experiment 1 and the 12 normed CNP scenes. Subjects were instructed to describe each picture using one simple sentence, and subjects' utterances and eye movements were recorded throughout the task.

Prior to stimulus presentation, subjects fixated a crosshair fixation point (equidistant from the two scene participants) for 500 msec. Subjects were misled to believe that position of the crosshair was randomized, to assist the experimenters in maintaining eyetracker calibration accuracy (no subject reported suspecting anything otherwise). The fixation point was then followed by a brief, covert attention-capture manipulation. This manipulation consisted of a small black target area (subtending an area of approximately 0.5X0.5 degrees of visual angle) against a white background, with a duration of 60-80 msec, followed immediately by the stimulus. Although no subject reported noticing the subliminal cue, it was highly effective in capturing attention. Subjects looked first to the cued location a median of 76% of the time.

Results

Table 3 shows rates of mentioning the Preferred First-Mentioned participant first for the CNP stimuli, and Table 4 shows rates of using the Preferred Subject for the PV stimuli for all four conditions in the 2X2 design. Collapsing across sentence types, significant effects of Left-Right Position and Attention-Capture were observed; leftmost and cued entities were more likely to be first-mentioned (p's<0.01). Further analyses showed that Left-Right orientation was significant only for word order in CNP stimuli ($p<0.01$), not for subject selection in PV items. Both sentence types, however, showed significant, stable effects of Priming, with primed characters more likely to appear first in CNPs ($p<0.05$) and to be the subject of a perspective verb ($p<0.01$).

Table 3: For all four conditions of Conjoined Noun Phrase stimuli, proportion of utterances in which subjects mentioned Preferred First-Mentioned participant first

	Preferred First-Mentioned Primed	**Dispreferred First-Mentioned Primed**	Average
Preferred First-Mentioned on Left	79.3%	63.8%	**71.6%**
Dispreferred First-Mentioned on Left	58.1%	41.4%	**49.7%**
Average	**68.7%**	**52.6%**	

Table 4: For all four conditions of Perspective Verb stimuli, proportion of utterances in which subjects mentioned used Preferred Subject

	Preferred Subject Primed	**Dispreferred Subject Primed**	Average
Preferred Subject on Left	87.4%	66.4%	**76.9%**
Dispreferred Subject on Left	77.3%	60.1%	**68.7%**
Average	**82.3%**	**63.2%**	

General Discussion

Language Production Overall, our results show a role for perceptual prominence in constituent ordering, and may be taken as support for a more incremental approach to sentence production.

It is important, however, to keep these results in the context of the current literature on the subject of speech production. Although Griffin and Bock (2000) found no correlation between first-fixated scene participants and first-mentioned participants, in an extensive investigation into the time course of message extraction from a visual scene, they *did* show tightly linked eye movement and speech patterns once an utterance was to begin; subjects looked reliably to an object less than a second before producing the corresponding word. This, and other research in this vein (Bock, Irwin, Davidson & Levelt, 2003), implies a system that begins with an initial, message-planning stage, followed by a more incremental process of retrieving the necessary lexical elements to construct an utterance (see Bock, in press, for discussion).

Our result is in no way inconsistent with this model of speech production. It is quite possible that subjects in our studies, rather than beginning to incrementally code their final utterance at the onset of the stimulus, begin with an information-extracting, message-planning stage, and that the perceptual priming effects we see take effect in this early stage. In the analogous ambiguous-figure literature, such attentional manipulations seem to affect the way subjects *perceive*, or interpret a stimulus. This may well be what's resulting from our similar attention-driving tools: a different perception, or interpretation of the stimulus. Ongoing research will investigate the effects of the same perceptual prime on both transitive verbs – where subjects must shift to an infrequent, passive structure to alter subject role assignment – and symmetrical predicates – where prominent information tends to appear in the object role/position (e.g. "I met Meryl Streep" vs. "Meryl Streep met me") (Gleitman et al., 1996). These explorations into the underlying nature of the perceptual prime should begin to determine where and how it is having its effect.

Language Acquisition These results have interesting implications for word learning studies as well. It has been noted that perspective verb pairs should be specifically very difficult for children acquiring a language to learn, as in many cases both members of these pairs necessarily co-occur under the same situational circumstances (Gleitman, 1990; Fisher, Hall, Rakowitz & Gleitman, 1994); for instance, a child is not apt to be presented with a situation that involves chasing but, at the same time, does *not* involve fleeing, and vice versa. How can the young learner figure out, then, whether the mother was saying "chase" or "run away?" These studies showed that syntactic information can inform the listener/learner as to the speaker's intended meaning. By varying the syntactic frame in which a novel verb appeared while referring to a perspective verb stimulus (e.g. "The man is glorping the dog" vs. "The dog is glorping the man," with regard to Figure 1) Fisher et al. showed that young listeners are quite adept at using this syntactic input, or "zoom lens," to arrive at the same interpretation intended by the speaker.

Another "zoom lens" that is more closely related to the present studies, is joint visual attention of speaker and listener. Infants as young as 2-months-old engage in such gaze-following activities (Bruner, 1998), looking where an adult is looking, during conversation. Moreover, by 12 to 18 months of age, the infant can successfully use this gaze-

direction information as a cues for how to label new objects (Baldwin, 1993). Contributions of attentional cues to word learning have not been as broadly or rigorously investigated for the case of verb learning. Given our current result on the relationship between attention-direction and variation between subject and verb choice, we suggest that similar attentional cues are available to the young language learner in successfully parsing and interpreting speech as well, even in the especially difficult case of perspective verbs.

Conclusion

Taken together, the results of these two experiments as they interface with relevant prior investigations clearly demonstrate a relationship between attention and language production. Further investigation will be necessary to delve into the detailed nature of this relationship, and explore the way it fits into a model of language production. These results, though, and the implications they have for attentionally-aware young language learners trying to interpret the speech stream, open exciting new investigative doors in both language production and acquisition.

References

Baldwin, D. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*, 395-418.

Bock, J.K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 575-586.

Bock, J.K., Irwin, D., Davidson, D. & Levelt, W.J.M (2003). Minding the clock. *Journal of Memory and Languag*e, *48*, 653-685.

Bock, J.K., Irwin, D.E. & Davidson, D.J. (in press) Putting First Things First. In J. M. Henderson & F. Ferreira (Eds.), *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Bruner, Jerome S. (1998). Routes to reference. *Pragmatics & Cognition*, *6,* 209-227.

Chomsky, N. (1957). *Syntactic Structures.* Oxford, England: Mouton.

Fisher, C., Hall, D.G., Rakowitz, S. & Gleitman, L. (1994). When It Is Better to Receive Than to Give: Syntactic and Conceptual Constraints on Vocabulary Growth. *Lingua, 92*, 333-375.

Forrest, L. B. (1996). Discourse goals and attentional processes in sentence production: The dynamic construal of events. In A. E. Goldberg (Ed.), *Conceptual structure, discourse and language*. Stanford, CA: CSLI Publications.

Georgiades, M. & Harris, J.P. (1997). Biasing effects in ambiguous figures: Removal or fixation of critical features can affect perception. *Visual Cognition, 4,* 383-408

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3-55.

Gleitman, L., Gleitman, H., Miller, C. & Ostrin, R. (1996). Similar and similar concepts. Cognition, 58, 321-376.

Griffin, Z.M. & Bock, J.K. (2000). What the eyes say about speaking. *Psychological Science, 11*, 274-279.

Jonides, J. & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics, 43,* Apr 1988, pp. 346-354

MacDonald, J.L., Bock, J.K. & Kelly, M.H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology, 25,* 188-230.

Osgood, C.E. & Bock, J.K. (1977). Salience and Sentencing: Some Production Principles. In Sheldon Rosenberg (Ed). *Sentence Production: Developments in Research and Theory*. Hillsdale, N.J.: Erlbaum.

Tomlin, R.S. (1997). Mapping conceptual representations into linguistic representations: The role of attention in grammar. In Nuyts, Jan & Pederson, Eric (Eds). *Language and conceptualization. Language, culture and cognition.* New York: Cambridge University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352

Discovering and Supporting Temporal Cognition in Complex Environments

Christopher Nemeth (cnemeth@uchicago.edu)
Cognitive Technologies Laboratory, The University of Chicago MC4028,
5841 S. Maryland Avenue, Chicago, IL 60637 USA

Richard Cook (ri-cook@uchicago.edu)
Cognitive Technologies Laboratory, The University of Chicago MC4028,
5841 S. Maryland Avenue, Chicago, IL 60637 USA

Abstract

Building new information tools to support cognitive work requires research at a level that, within the constraints of time and resources, will reveal higher-order cognition among practitioners. Practitioners develop cognitive artifacts in order to perform technical work. As densely encoded representations of work domains, their artifacts embody the most meaningful information in the task setting. The study of cognitive artifact development and use makes it possible to study individual and team cognition. This approach reveals what information is important, and how practitioners capture and use it. As replicas of physical artifacts, digital cognitive artifacts often amount to only meager representations of what matters in the work environment. This clumsy automation imposes a burden on practitioners by forcing them to cope with its shortcomings. In this setting, user-centered automation must support reasoning through time. The study of physical artifacts indicates ways that digital artifacts might better support temporal reasoning.

Use of Cognitive Artifacts to Understand Technical Work

The coordination of anesthesia assignments at a major urban teaching hospital spans 50 to 80 cases a day and requires the orchestration of multiple departments including anesthesia, surgery, nuclear medicine, obstetrics and gynecology, gastrointestinal endoscopy, diagnostic and interventional radiology, and psychiatry. This activity involves a distributed cognition (Hutchins, 1995), that is comprised of the shared awareness of goals, plans, and details that no single individual grasps. Through socially distributed cognition (Perry, 1999), individuals cultivate the mutual awareness and understanding that is needed to collectively accomplish shared goals.

Surgeons, anesthesiologists and the others at the hospital work to a Standard of Medical Expertise (SME). Resources among care settings, patient populations and system are constrained and must be allocated prudently in order to meet a Standard of Resource Use (SRU). (Sharpe and Faden, 1998). A few of the senior anesthesiologists serve in the role of daily coordinator, assigning staff to perform a full schedule of anesthesia, sedation or pain management procedures each weekday. To do this, the coordinator must evaluate the number and types of procedures, determine the number and types of staff available, assign staff to perform procedures, and evaluate the balance between the two. The coordinator typically manages the execution of that schedule on the following day. Management of this process involves the synchronization of complex, changing activities through time. This requires an accurate grasp of the number and nature of available staff as well as an accurate, up-to-the-minute account of procedures that have been performed so far, are underway, and have yet to be performed within work setting constraints.

Research into cognitive activity in this setting is challenging for a number of reasons. Healthcare practitioners may have little insight into how their work is organized. Information and interaction at the sharp (operator) end is dense, complex, varies widely, and changes rapidly. (Cook and Woods, 1994)

In order to understand cognition in this environment, the researcher needs to employ a number of methods. Woods and Roth's (1988) cognitive engineering approach studies behavior in actual environments in order to change behavior and to improve performance. Klein's (2000) naturalistic decision making (NDM) approach accounts for the performance of decision makers in actual settings. Hutchins' (1995) ethnomethodology describes how distributed cognition includes artifacts that make it possible for a group to accomplish shared goals.

The development and use of cognitive artifacts makes it possible to perform the otherwise impossible process of assignment coordination. Cognitive artifacts are an efficient representation of what matters here because they represent only the information that is critical in this work domain. Previous work (Nemeth 2002, 2003a) describes the use of observational studies to discover how the acute care team uses cognitive artifacts to make the plan for the day's work. It also explains how controlled study of artifact creation reveals the strategies that coordinators employ in order to create a feasible future for the next day of procedures. Two artifacts are essential to the coordinator while developing a plan. The Daily Availabilities sheet is used to account for the status of each of the members in the department who are available for assignment. The preliminary copy of the Master Schedule lists all procedures that are scheduled to be

performed the following day that will require anesthesia, sedation or pain management. These two artifacts are the tools that are used to create the Master Schedule.

An example shows how one coordinator uses the Daily Availabilities sheet and the preliminary copy of the Master Schedule to build a final version of the Master Schedule. Figure 1, from Nemeth (2003b), represents the schedule development process in three ways. The left column shows the verbatim transcript of how the coordinator describes his deliberations using Verbal Protocol Analysis (VPA). The Daily Availabilities and Preliminary Copy that he refers to are shown at center, along with indications of where he is paying attention. A diagram and comments at right show the analysis of his cognitive activity as he assigns available attending and resident anesthesiologists to cover procedures in eight outpatient clinic rooms. After scoping the supply of staff resources and evaluating the type and number of procedures, the coordinator assigns staff to particular procedures and then assesses the assignments. In eleven minutes, he has assigned attending and resident anesthesia staff to perform a day's

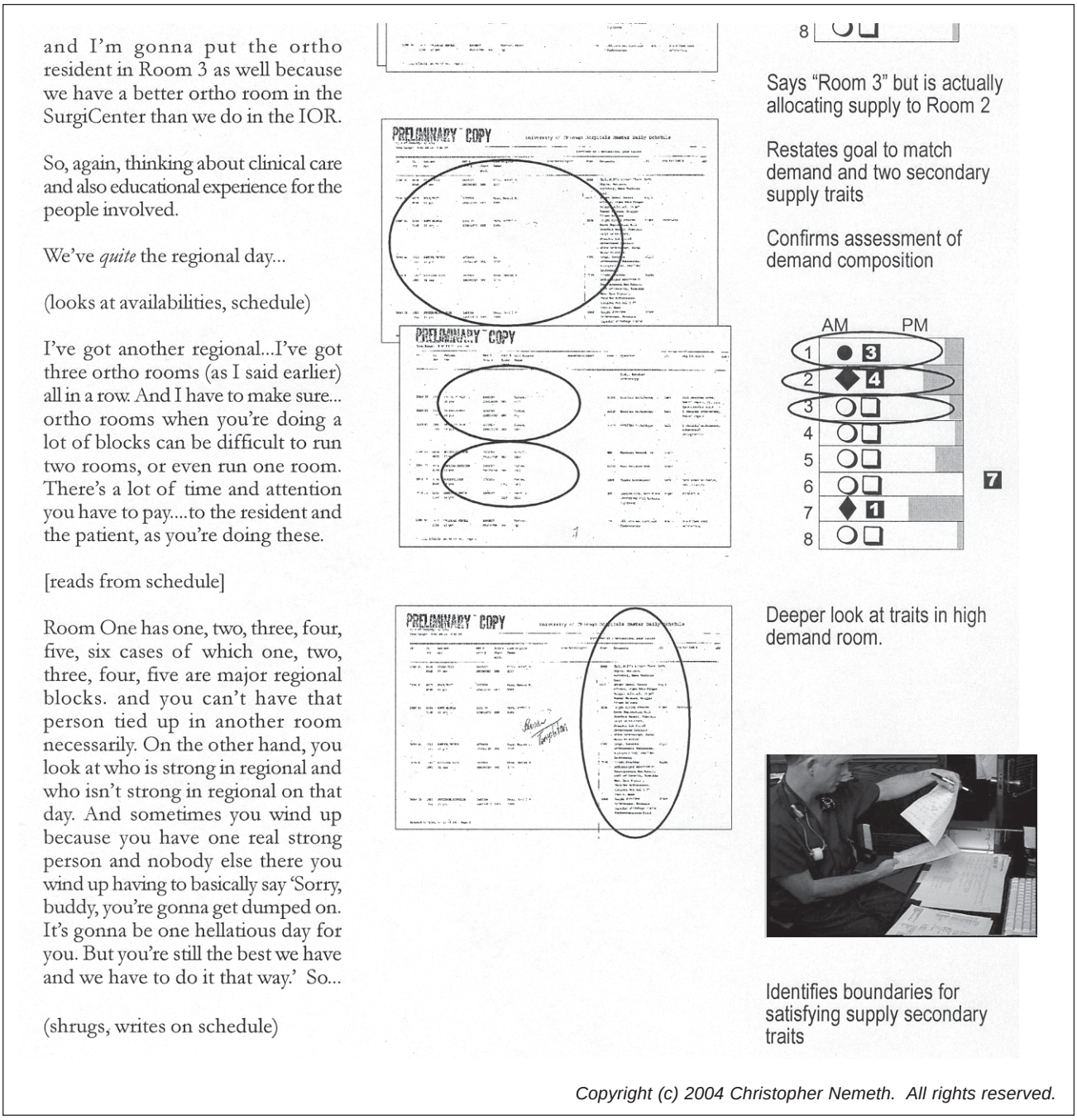
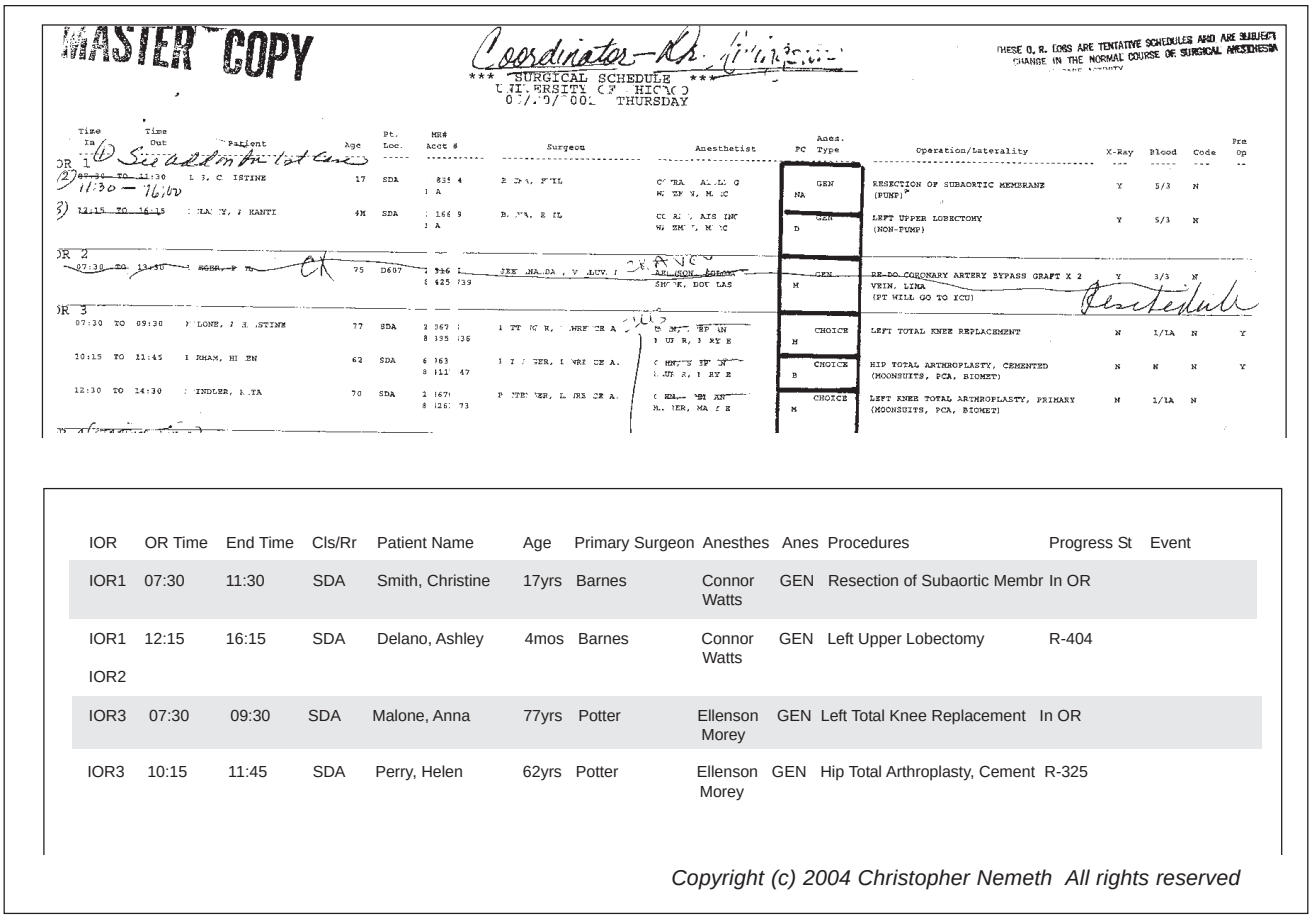


Figure 1:Anesthesia Coordinator Schemata Analysis *(selected portion)*



IOR	OR Time	End Time	Cls/Rr	Patient Name	Age	Primary Surgeon	Anesthes	Anes	Procedures	Progress St	Event
IOR1	07:30	11:30	SDA	Smith, Christine	17yrs	Barnes	Connor Watts	GEN	Resection of Subaortic Membr	In OR	
IOR1	12:15	16:15	SDA	Delano, Ashley	4mos	Barnes	Connor Watts	GEN	Left Upper Lobectomy	R-404	
IOR2											
IOR3	07:30	09:30	SDA	Malone, Anna	77yrs	Potter	Ellenson Morey	GEN	Left Total Knee Replacement	In OR	
IOR3	10:15	11:45	SDA	Perry, Helen	62yrs	Potter	Ellenson Morey	GEN	Hip Total Arthroplasty, Cement	R-325	

Figure 2: Master Schedule Physical Artifact (upper) and OpAssign format (lower)
All names are fictitious

outpatient procedures. This planning process requires deep domain knowledge and deft diplomacy. It also requires the ability to exploit opportunities, to create trial solutions and to ȧssess their possible consequences. No two coordinators approach the process in the same way. Until recently, the coordinator would use only hard copies of the artifacts to develop the Master Schedule. During the day the coordinator would track and update case status by making marks on the Master Schedule hard copy that was posted at the Inpatient Operating Rooms (IOR) coordinator station. Team members also used the OR Board, a white marker board with magnetic plaques, as a platform to discuss assignments, negotiate trade-off decisions, plan and re-plan assignments, speculate about how to re-balance changes in demand and staff.

Digital Artifact Concerns

Hard copies of the Master Schedule have recently been replaced by a computer-based system, OpAssign. The Daily Availabilities remains a hard copy report that the coordinator refers to while composing the Master Schedule.

The OpAssign display is an alphanumeric table that is shown on a flat screen monitor. As a mimic of the physical Master Schedule, OpAssign shows surgical procedures that are organized by room and by time of day within the room. All procedures are represented by alphanumeric characters and each procedure occupies the same amount of space in the layout. Colored bars are intended to indicate case status such as called for, arrived on unit, in OR, in-progress, delayed, and concluded. Procedures appear identical on the display even though they differ as markedly in criticality and duration as a circumcision and a coronary artery bypass graft (CABG).

A number of changes have occurred as a result of the transition from physical to digital cognitive artifact.

▪ The physical artifact had previously made it possible for the coordinator to control the accuracy of information that was used to make decisions. Only the coordinator would make marks on the one original hard copy that was posted at the coordinator station. The physical artifact also allowed for the coordinator to make margin notes to keep track of unofficial, yet important information such as the name and extension of a staff member who had called with information

that pertained to a case. Now that the Master Schedule is no longer a physical artifact, neither the coordinator nor the acute care team can annotate it.

- Data field limitations force certain compromises in the information that can be shown. Many elements of information are truncated on the OpAssign display. Details can only be found by drilling down through multiple levels of the interface.

- Information on case status was traditionally written onto the Master Schedule hard copy by the coordinator as a patient was wheeled past the coordinator station toward an IOR room. Now, case status must be entered via laptop from a very busy IOR room. This means that information on the Master Schedule display can lag actual events by 30 minutes or more. The lag causes the coordinator to second guess the display and to do additional cognitive work to check on case status. It also erodes the coordinator's ability to manage decisively.

- Case location on the original hard copy remained the same. Team members could use the fixed location for each case to find and refer to it. As case status changes through the day on OpAssign, their location on the screens also changes and team members have to search to find them.

These and other shortcomings have caused team members to do additional cognitive work to cope with limitations of the digital display, impeding team performance. Figure 3 shows OpAssign in use at the coordinator station.

Figure 3: OpAssign Display at Coordinator Workstation

Opportunities to Improve Displays

Interestingly, both the hard copy of the Master Schedule and the OpAssign display list scheduled start and end times. However, neither of these two reflect the time-related demandsand complexity that are the primary drivers in this environment. Figure 4 illustrates a conceptual prototype for a digital version of the Master Schedule information that draws on the findings from research into coordinator schedule development and team schedule use. Six of the IOR rooms are shown in the figure. Information on each case is shown in a horizontal bar that is aligned next to the label of the operating room to which it is assigned. A shaded segment follows each procedure to indicate the 45 minute period that is required to clean-up and restock the room. The arrow at top of the display indicates that the time is 0800 on the day of procedures that are being conducted in the Inpatient Operating Room (IOR) unit.

IOR1 shows that a half hour is open after the first procedure and clean-up have been completed. The procedure that is scheduled for 0730 to 1330 in IOR2 has just been cancelled and the display indicates that the room is scheduled and prepared for use. IOR3 shows that cases are scheduled efficiently. IOR4 is available for any general surgery to be added on after 1215. Somehow, the second procedure scheduled for IOR6 has been slated to start before the technicians would be able to finish clean-up. The solid bar can be used to display more information on the cases by choosing it with an input device such as a mouse or touching the screen.

Certain information is crucial in order to optimize assignments. This includes knowing when procedures are likely to finish, which procedures can be moved into another room, and which opportunities (such as Medicare payment) might be exploited. Such information can be made available by polling the database of scheduled cases to see what opportunities may exist.

The example in Figure 4 is based on research into the work domain in which it would be used. Because of this, it avoids many of the shortcomings that the OpAssign display encountered. It may also improve on the OpAssign design in a number of respects.

- The visual organization of the display remains the same as it evolves. By using a graphic representation of time, the team can understand and evaluate relationships among events through time.

- Relevant variables such as age are shown within each case window, which saves the need to locate and assemble information that is related but is displayed separately.

- Cases that were performed remain on the display in sequence, making it possible to review the entire day's activities while they are still underway.

- Aspects of schedule management that were previously hidden are made evident. These include requirements that are the objects of coordinator cognitive work such as showing conflicts and gaps in timing, and constraints on schedule management such as room clean-up and restocking.

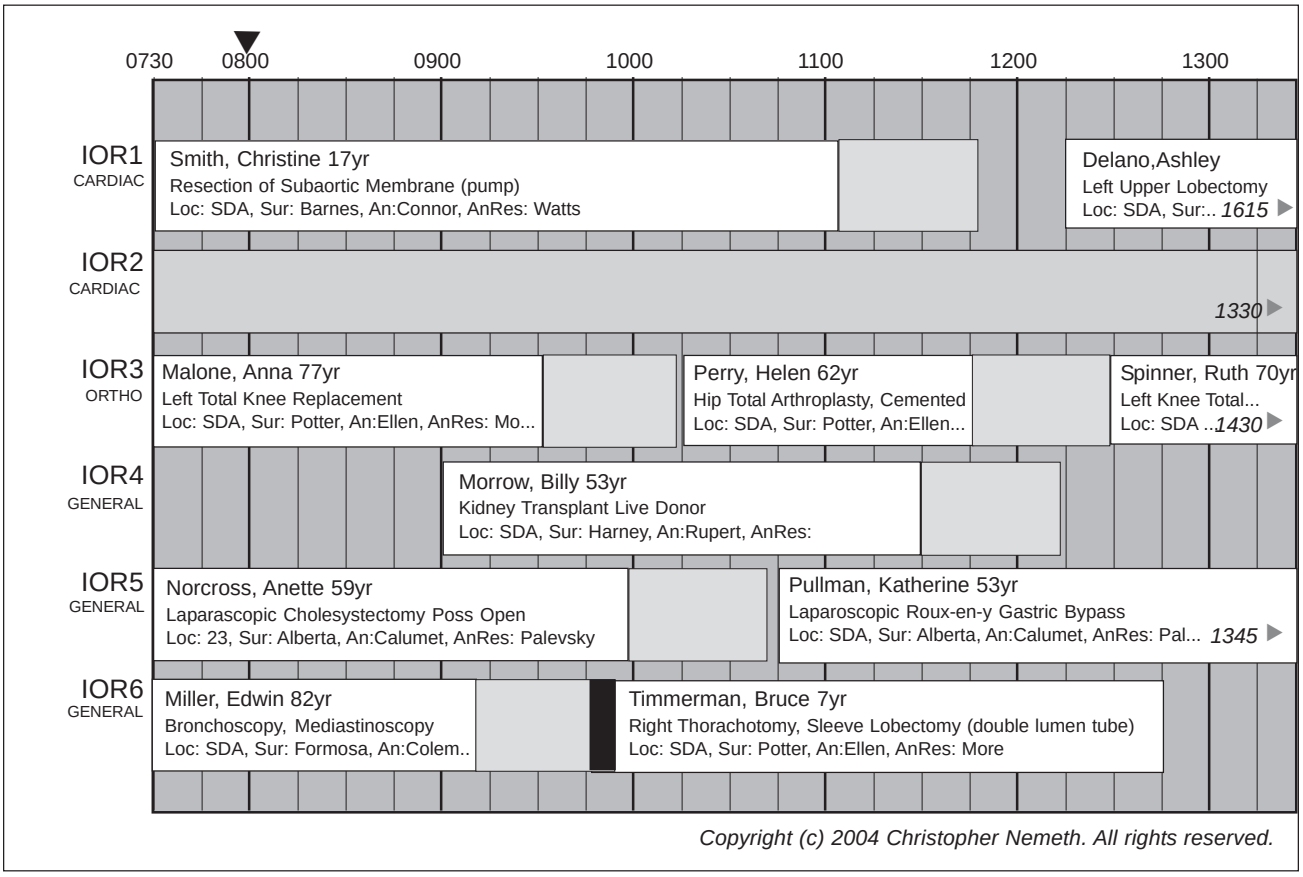


Figure 4: Prototype Digital Display of the Master Schedule
All names are fictitious

Adding Value to Cognitive Tools

Any tools that are created to assist these complex and highly sensitive interactions need to reflect the underlying complexity of the work that is to be performed. A digital version of the Master Schedule might improve team performance by supporting work in ways that the research that was described earlier in this paper demonstrated.

The flexibility that digital representation offers is powerful and can be used to support cognitive work. However, this does not happen automatically. The digital artifact's design must represent constraints and opportunities that are relevant in this domain. Because time is the key aspect here, organizing display design according to time allows users to easily track changes, to anticipate future events, and to respond to emerging situations.

Further features such displays might provide include:

Prompting—Digital artifacts might survey information in the distributed cognition for gaps and inconsistencies that go unnoticed and unaccounted for. Nominating the item(s) for consideration would enrich and improve the cognition.

Speculation—Digital artifacts can enable coordinators to speculate about and choose among possible courses of action. For example, speculation about plans for the afternoon staff is currently limited to the OR Board. Developing potential courses of action would make it possible to evaluate how desirable they might be.

Consequences—Applying evaluation criteria to potential courses of action could make it possible to display the consequences of choices. One example is to show how billing might be increased, or costs might be minimized, by opening one operating room or closing another.

Value-based decisions—Digital artifacts can be used to develop templates of schedule planning strategies. Coordinators could review and use the template that best matches their values and preferences. Such templates can capture scheduling expertise and make it available beyond a single individual. Study of template use through time might open the way to insights about coordinator training and the development of further schedule models that might ease coordinator workloads.

Conclusion

This paper has described a detailed study of the operational aspects in a complex, high hazard work setting, assessed the role and effect of physical and digital cognitive artifacts on cognitive work, and presented a display concept that embodies the task demands that workers confront. Support for the cognitive work of those who labor in this setting is the hallmark of user-centered automation. (Billings, 1997)

As a readily available source of information, cognitive artifacts make it possible to study cognition in complex environments. Because those who work in the environment have created them, artifacts are highly encoded representations of what matters most in complex settings. The creation and use of cognitive artifacts also provide the researcher with a means to understand deeper structure of behavior in the work domain.

Findings from such research can be used to identify the functions of computer-supported displays that are needed to not only support but to improve performance. Validation of those findings and related display designs will come from operator acceptance in actual use. Improving work efficiency and reliability can make it possible for work teams to be more effective, thereby improving medical safety.

Acknowledgements

This work was supported by a grant from the National Library of Medicine.

References

Billings, C. (1997). *Aviation Automation: The Search for a Human Centered Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cook, R.I. and Woods, D. (1994). Operating at the Sharp End: The Complexity of Human Error. In Bogner, M. S. (Ed.). *Human Error in Medicine*. (pp.255-310). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: The MIT Press.

Klein, G. (2000). *Sources of Power*. Cambridge, MA: The MIT Press.

Nemeth, C. (2003a). How Cognitive Artifacts Support Acute Care Distributed Cognition, In Cook, R. and Woods, D., Insights From Technical Work Studies in Healthcare, Symposium at *Human Factors and Ergonomics Society National Conference*, Denver.

Nemeth, C. (2003b). *The Master Schedule: How Cognitive Artifacts Affect Distributed Cognition in Acute Care.* Dissertation Abstracts International 64/08, 3990, (UMI No. AAT 3101124).

Nemeth, C., Cook, R.I., O'Connor, M., and Klock, P.A. (2003, October) Using Cognitive Artifacts to Understand Distributed Cognition. In Xiao, Y., Special Session on Distributed Planning. *IEEE International Conference on Systems, Man & Cybernetics*, Washington, D.C.

Nemeth, C., Klock, P.A., Daves, S., and Cook. R.I. (2002, October). A Study of How Cognitive Artifacts Affect Distributed Cognition in Operating Room Management. *Proceedings of the American Society of Anesthesiologists Annual Meeting.* Orlando.

Perry, M. (1999). The Application of Individually and Socially Distributed Cognition in Workplace Studies: Two Peas in a Pod? *Proceedings of European Conference on Cognitive Science*, pp.87-92, Sienna, Italy.

Sharpe, V. and Faden, A. (1998) *Medical Harm*. New York: Cambridge University Press.

Weiner, E. (1985). Beyond the Sterile Cockpit. *Human Factors*, 27: 1,75-90.

Woods D., and Roth E. (1988). Cognitive Systems Engineering. In Helander, M. (Ed.) *Handbook of Human-Computer Interaction.* pp. 3-43. Amsterdam: North-Holland.

Semantic Effects in Speech Production

Adrian Nestor (adriannestor@students.nbu.bg)
Elena Andonova (eandonova@cogs.nbu.bg)
Department of Cognitive Science
New Bulgarian University, 21, Montevideo Street
Sofia 1618, Bulgaria

Abstract

The paper reports empirical and computational research on semantic facilitation and inhibition in the picture-word interference paradigm for different values of stimulus onset asynchrony (SOA). The main claim it makes is that purely semantic facilitation effects are conditioned by the degree of picture-word semantic similarity in contrast to inhibition effects for which categorical relatedness is typically sufficient. The experimental results support this claim and a simulation with an attractor neural network attempts to provide a unified account of semantic facilitation and inhibition in this paradigm as a function of SOA.

Introduction

Semantic facilitation from context words in a picture naming task, just like semantic priming in visual word recognition, has been a somewhat controversial issue. However, in picture naming, unlike word recognition, the issue does not seem to have been settled in favor of facilitation. It is naturally to ask then why purely semantic facilitation should show up when making a lexical decision or reading words but not when naming pictures. Both methodological and theoretical grounds are examined as potentially responsible for the failure to establish such a result with a picture naming task. The claim we attempt to support is that the categorical view on semantic similarity, as opposed to a graded featural one, is mainly responsible here. However, if facilitation is shown to be possible, one has to face the challenge of accommodating it with other experimental results, notably picture-word interference, in the framework of a general theory of speech production.

Originating with the study of Rosinski, Golinkoff & Kukish (1975), the picture-word interference paradigm established a classical result in the psycholinguistic literature. To put it succinctly, it takes more time to name the picture of an object in the presence of a semantically related word than in the presence of an unrelated one. For instance, naming the picture of a DOG in the presence of the word 'cat' takes longer than in the presence of 'tree'.

The classical study of Glaser and Düngelhoff (1984) opened the way to time-course analyses of the interference effect in the picture-word paradigm by systematic manipulation of the stimulus onset asynchrony (SOA) values, i.e. the temporal difference between the onsets of the two stimuli on a trial, the target picture and the context word. Examining different SOA values ranging from –400 to 400 ms, where negative SOA values correspond to word onset preceding picture onset, the two researchers found semantic inhibition in a small time window around synchrony (SOA=0) with a maximum inhibitory effect at 100 SOA. A facilitation effect at – 400 ms SOA, on the other hand, was not found to be significant.

Semantic interference around synchrony has been replicated many times and extended to cover various modifications of stimulus properties such as modality of the prime, category of named objects, stimulus duration, SOA etc. (Alario, Segui & Ferrand, 2000; Damian & Martin, 1999; La Heij, 1988; Roelofs, 1992; Schriefers, Meyer & Levelt, 1990; Starreveld & La Heij, 1996).

Facilitation at early SOAs, on the other hand, was more of a debated possibility. While the early studies of Sperber, McCauley, Ragain and Weil (1979) and Carr, McCauley, Sperber and Parmelee (1982) reported facilitation, later research failed to obtain it (Alario, Segui & Ferrand, 2000; Glaser & Glaser, 1989; La Heij, Dirx & Kramer, 1990). The reason invoked for this difference was twofold. First, early studies used long or very long SOAs – Sperber et al. (1979), for instance, reported facilitation for an interstimulus interval of as much as one second – which opened the gate to strategies on the part of subjects. Second, they failed to separate semantic from associative relatedness. 'Dog' and 'cat', for example, are good candidates not only for cohyponims, i.e. members of the same semantic category, but also for close associates as recorded by free association norms. The nature of the facilitation could have been then associative rather than semantic as suggested by robust purely associative facilitation results at negative SOA values (Alario, Segui & Ferrand, 2000; La Heij, Dirx & Kramer, 1990; Lupker, 1988). As a result, currently the issue seems to be settled pretty much against the possibility of purely semantic facilitation at negative SOAs.

The debate on facilitation in the picture-word paradigm can be paralleled quite interestingly with the debate on automatic semantic priming in lexical decision and word reading. Does a word prime speed up, for instance, the lexical decision for a semantically related word compared to a semantically unrelated one? The examination of many studies investigating this issue cannot give us a straight answer (see Neely, 1991 for a review). Similarly to the line taken in the picture-word paradigm, it was suggested that automatic priming in such a task reflects only an associative relationship (Shelton & Martin, 1992).

McRae and Boisvert (1998) suggested that the divergence of results and the failure to elicit purely semantic priming in

a number of studies could be due to an inadequate conception of semantic relatedness. In many of the previous studies two concepts have been taken as semantically related if they belonged to the same intuitive semantic category as judged by the researcher. However, if categorical relationships do not reflect accurately the organization of our semantic memory and only approximate quite roughly the graded nature of semantic similarity, e.g. featural similarity (McRae, de Sa & Seidenberg, 1997), then our choice of semantically related prime-target pairs in the experiment should be performed accordingly. McRae and Boisvert tested their hypothesis in a series of experiments with a lexical decision task controlling the semantic similarity of stimuli. The results supported their hypothesis. Highly similar pairs, e.g. *turkey – goose*, did elicit semantic priming at –250 ms while less similar pairs, e.g. *turkey – robin*, did not. Purely semantic priming was also obtained with synonyms and near-synonyms by Perea and Gotor (1996) both in lexical decision and word reading tasks, a result which adds further support to the hypothesis mentioned above.

In picture-word experiments researchers also tend to adhere to a categorical understanding of semantic relationship. Two concepts are judged as semantically related if they relate as cohyponyms or as category – exemplar pairs. This is enough to guarantee a reliable interference effect, so facilitation is searched for by the same standards. What if facilitation is a smaller effect sensitive to the degree of similarity between picture target and context word? An average or even low degree of similarity might be enough for a large and robust interference effect around synchrony but facilitation at earlier SOAs might remain undetected.

If one could show, as we attempt below, that there is semantic facilitation and, additionally, that it is conditioned by the degree of semantic similarity, that would be a successful extension of the results from one experimental paradigm to another and additional support for a graded view of semantic relatedness. However, it would amount to more than simply importing an experimental result from one paradigm to another. In picture-word experiments, in contrast to primed lexical decision or word reading, semantic interference is the rule. Putting together semantic facilitation and inhibition as a function of SOA is the challenge one would have to face next. A simulation with an attractor neural network is our candidate for accommodating these opposite effects.

The proper way of understanding and deploying semantic relatedness, on the other hand, might not be the only problem in detecting a small facilitation effect in picture-word experiments. One particular aspect of the procedure made use of in such experiments can also be a source of worry. In most of the studies in the paradigm, subjects are allowed to familiarize themselves with the target pictures and their proposed names in a pre-experimental session. Presenting the pictures and their desired names in advance certainly offers a number of benefits like keeping low the rate of misnaming errors. Familiarization with the target pictures is also supposed, as argued by Glaser and Düngelhoff (1984), to homogenize subjects with respect to visual processing of the pictorial stimuli. All of this, however, comes at a price. Subjects are primed before the experiment with the visual stimuli. Asking them to name the pictures in a pre-experimental session extends the scope of priming to all levels of processing in speech production. Moreover, having subjects learn the names of the pictures in advance could be a way of turning a lexical task into more of a memory task. Instead of accessing semantic memory for the concept corresponding to the picture, subjects may try to recall the name used for it before the experiment. The impact of these factors on the detection of a semantic effect, especially a small one, is hard to predict. The alternative we offer is a careful control of the pictorial stimuli, in particular their name and image agreement. Optimizing the stimulus material rather than preparing subjects for the test could be a reasonable alternative.

Experiment

The experiment addresses mainly the possibility of a semantic facilitation effect at an early SOA in the picture-word interference paradigm. The hypotheses we explore ascribe the failure to find such an effect in previous studies to an inadequate view on semantic similarity and to procedural specificity.

First, the experiment aims at exploring the role of semantic similarity in this paradigm. The question we ask is whether a high degree of semantic similarity, as opposed to simple categorical relatedness, could induce such an effect at an early SOA.

Second, we attempt to improve on the experimental procedure by eliminating subjects' familiarization with the stimulus material before the experiment and replace it with careful control of the stimuli. If this familiarization is the main culprit for hiding facilitation, we would expect to detect such an effect independent of the degree of picture-context word semantic similarity as long as they qualify for categorical relatedness.

Finally, as we deploy a somewhat nonstandard procedure and as it is the first time an experiment in this paradigm is run with Bulgarian stimuli, it is desirable to seek the replication of the classical result in the paradigm, i.e. semantic inhibitory effects around synchrony.

Participants

We tested 45 undergraduate students at the New Bulgarian University in Sofia, Bulgaria. All were native Bulgarian speakers and reported normal or corrected-to-normal vision. None of them participated in the norming studies performed in advance on the stimulus material used in the experiment.

Materials and Design

Thirty-six black-and-white drawings of common objects were selected from the pool of items for which norms are available in Bulgarian as part of the crosslinguistic study of

Bates et al. (2003). The choice of items was guided mainly by the attempt to maximize their nameability or consensus, i.e. percent of subjects naming the picture with the same name (dominant name), and image agreement as rated by subjects on a 1-to-7 scale. The main properties of the materials are summarized in Table 1.

Table 1: Properties of the target picture stimuli.

	Mean	Range
Consensus (%)	92	78-100
Image agreement (1-7)	5.7	4.6-6.6
Subjective frequency (1-7)	4.7	3.5-6.2
Concreteness (1-7)	5.9	4.5-6.7

Next, 36 picture-word pairs were constructed such that each picture and its corresponding word denote concepts from the same intuitive semantic category, e.g. *spider* and *ant* as insects. Care was taken that the pairs would not be close semantic associates by the examination of free association norms for Bulgarian (Gerganov, Ivancheva, Kurlova, Nikolov & Nikolova, 1984).

The total set of pairs was split in two halves as they scored higher (M=5.03, SE=0.15) or lower (M=3.33, SE=0.1) on a semantic similarity test in which 20 subjects were asked to rate the degree of similarity of the objects denoted by the pair items on a 1-to-7 scale (1 – very dissimilar, 7 – very similar).

Each picture was additionally assigned a nonword and an unrelated word by random choice of a word from a different semantic category. Thus, our first factor was type of context: semantically related, unrelated and nonword. Context type was crossed with the SOA. Three values were chosen to cover the time-course of semantic effects in picture naming: -350, 0 and 100 ms. Negative SOA values indicate that the context item was presented before the target picture. Both factors were within-subject.

Nine different lists were constructed containing all picture-word pairs, an equal number of pairs per condition. Half of the semantically related pairs in each list fell in the higher semantic similarity group while the other half in the lower semantic similarity one.

Procedure

Subjects were randomly assigned to lists and tested individually using the PsyScope Experimental Control Shell on a Macintosh computer with a 14-in. monitor.

The pictures were scaled and centered in an imaginary square with dimensions 4.5° X 4.5° visual angle. The (non)words, displayed below the pictures, were selected for each target so as to maximize the length match with their corresponding pictures. Both stimuli on a trial, the picture and the distractor, were thus displayed within an imaginary rectangle with dimensions 4.5° X 6° centered in the middle of the screen.

The subjects were instructed to name the pictures that would appear on the screen as quickly and accurately as possible with the best name they could think of and to avoid false starts, hesitations or extraneous material.

Each experimental trial had the following structure. A fixation crosshatch appeared in the center of the imaginary rectangle containing the picture and the (non)word for 200 ms followed by a 50 ms blank interval. Depending on the SOA, next appeared the picture, the (non)word or both. The (non)word stimulus duration was set to 200 ms. The target picture remained on the screen for a maximum of 2 seconds. The picture disappeared from the screen as soon as a vocal response was registered by a voice key.

The experimental session was preceded by a 6 trial warm-up set to let subjects familiarize with the task. There was no previous presentation of the stimulus material in the experiment and once presented stimuli were not repeated later with the same subject.

Results

Responses such as productions of names different from the dominant name, verbal disfluencies or no responses were scored as errors and excluded from the analysis. Three subjects and two items were excluded from the analysis because of the large percentage of errors they produced (more than 25 %). The error percentage for the remaining trials reached a level of 6.4%. Latencies greater than 3 SDs above the mean per condition were also removed. The resulting mean naming latencies per condition averaged by subjects are presented in Table 2. Positive values for semantic effects indicate facilitation while negative values indicate inhibition.

Table 2: Mean naming latency and semantic effect per condition[1].

Context type	SOA		
	- 350	0	100
Semantically related	892	973	958
Semantically unrelated	911	941	934
Nonword	898	910	910
Semantic effect	19	-32*	-24*

An overall analysis of variance (ANOVA) was conducted on the RT data, with context type and SOA as within-subject factors. Both analyses, by subjects and by items, revealed significant main effects for distractor type[2] ($F_1(2,82)=12.99$; $F_2(2,66)=8.51$), for SOA ($F_1(2,82)=23.96$; $F_2(2,66)=13.48$) and for their interaction ($F_1(4,164)=6.52$; $F_2(4,132)=2.97$). The lexical effect between the unrelated word condition and the nonword one was also found significant in a separate analysis ($F_1(1,41)=9.11$; $F_2(1,33)=66$). Naming latencies were significantly smaller at the negative SOA compared to either of the other two conditions. More importantly, we obtained semantic facilitation at –350 ms SOA and inhibition at 0 and 100 ms. Planned comparisons run for each of the SOA values established significant inhibitory effects both at synchrony

[1] Effects significant at $p<.05$ are marked with an asterisk *.

[2] In all analyses reported, $p<.05$ unless otherwise stated.

($F_1(1,41)=6$; $F_2(1,33)=11.27$) and at 100 ms ($F_1(1,41)=4.8$; $F_2(1,33)=4.3$). However, at the negative SOA there was only a trend to semantic facilitation both by subjects and by items ($F_1(1,41)=3.48$, $p<.07$; $F_2(1,33)=3.42$, $p<.08$).

Next, planned comparisons by items between the related and unrelated conditions were run at each SOA separately for the high and low similarity sets. They both replicated the inhibition pattern found at 0 and 100 ms with all items. The only difference showed up for the negative SOA where semantic facilitation reached significance for the high similarity set ($F(1,16)=4.57$) but not for the low similarity set ($F(1,16)=0.5$, $p<.56$). An additional power analysis showed the power to detect with all items an effect of the same size as the one obtained with the high similarity set was as high as 0.84.

The error analysis did not lead to any significant results.

Discussion

The experiment tested the possibility of detecting semantic facilitation at a negative SOA along with inhibition around synchrony in the picture-word interference paradigm.

One hypothesis ascribing the failure to detect facilitation in previous experiments to potential procedural drawbacks, namely pre-experimental familiarization of subjects with the stimulus material, was not supported by the results. Without familiarization we only noticed a trend to facilitation at −350 ms SOA (Glaser & Düngelhoff, 1984). However, we were able to replicate the inhibition results they and many others reported for small positive SOAs.

A second hypothesis linked this failure to too rough a conception of semantic similarity, namely categorical relationship. A move from a categorical view to a featural similarity view on semantic similarity proved successful in obtaining semantic facilitation in the word recognition paradigm (McRae & Boisvert, 1998). Making the same move in the picture-word interference paradigm proved to be a successful strategy, too. Categorical relatedness alone was not enough to deliver a significant facilitation effect. But high semantic similarity did provide us with such an effect. Additionally, the power analysis indicated that categorical relatedness, that is, the analysis including all semantically related pairs, was very likely to detect a significant effect of the size of the one we observed with the high semantic similarity set, if such an effect existed. The absence of a graded inhibitory effect at positive SOAs, on the other hand, was not a surprising result. As categorical relatedness is able to ensure large inhibitory results, a ceiling effect seems a likely explanation for this absence.

However, one worry we still need to address is the possibility of strategy use. Facilitation alone is not at stake. One has to ensure it is automatic. We could argue against expectancy effects at least on two grounds. First, 350 ms is still a small interval. It does not depart so much from the 250 ms limit proposed by Neely and Keefe (1989) for ensuring automaticity of effects and is still far from the large SOA values used in the early studies. Second, a lower relatedness proportion is believed to diminish or eliminate expectancy effects for long SOAs, e.g. de Groot (1984). Related pairs made up only a third of the stimuli subjects were presented and they could have deployed expectancy strategies meaningfully only at –350 ms SOA, that is, on one ninth of the total number of trials. Taken together, the size of the interval and the proportion of the occasions for potential strategy use could offer reasonable support for the claim that the semantic effect we observed was automatic.

Finally, our results and interpretation need to be qualified by the remark that they were obtained in Bulgarian, a Slavic language. To what degree effects like the one we investigate are sensitive to language characteristics is an experimental question. The study of Bates et al. (2003) raised such an issue with respect to timed picture naming. One peculiarity noticed about Bulgarian in this study was that naming latencies were longer than in other languages. In our experiment we also obtained naming latencies a little longer than the ones reported in similar experiments carried out in other languages. However, we were successful in replicating semantic interference effects for this language as they have been observed in many others. That only gives further ground for the expectation that semantic effects are robust enough across language dissimilarities.

Simulation

Once experimental evidence is available, the next step would be to offer an explanation for the effect in the general framework of a speech production theory. A unified account of semantic inhibition and facilitation in a picture naming task as a function of SOA would be the main challenge here.

If we conceive of the speech production system as a series of layers (stages) one on top of the other, we might get a very general idea of the way most theories in the field go. The sequence semantic representations (semantic processing) – lemmas (syntactic processing) – lexemes (morpho-phonological processing) – phonetics is a good candidate here (Dell, 1986; Levelt, Roelofs & Meyer, 1999). The interference effect is typically explained in this framework by competition / inhibition of the lexical items at the level built on top of semantics, e.g. lemmas (Roelofs, 1992) or lexemes (Starreveld & La Heij, 1996). A semantically related word compared to an unrelated one activates more its corresponding lexical unit at the next level as it receives additional support from the semantic representation of the target to be named. Thus it makes a more powerful competitor than the lexical representation of an unrelated word and interference shows up as a result.

In order to explain the pair of facilitation - inhibition effects in single word production we set up a small 3 layer neural network simulating lexical item activation based on semantic input. The input layer hosts semantic representations distributed over a set of feature units. The output layer stands for localist representations of lexical items at some level in the speech production system, possibly lexemes. The hidden layer thus compresses the intermediate levels of processing, notably lemma (syntactic) processing. Therefore, we may view the network as a

simplified model of speech processing in three stages: concept activation, lemma retrieval and lexeme selection.

A 20 X 30 X 20 feedforward architecture was enriched with a full set of interconnections at the hidden layer (without self-connections). Continuous spread of activation and a continuous version of backpropagation through time (Pearlmutter, 1989) as a training algorithm were made use of in order to be able to simulate temporal properties and relations of the stimuli like SOA and stimulus duration as well as naming latencies (see Plaut, 1995).

The network was trained for a fixed amount of time on 20 input-output pairs. The inputs were random distributed patterns on the semantic layer while the outputs were localist lexeme representations. That is, each input pattern was supposed to activate only its corresponding lexeme node on the output layer. Semantic relatedness was simulated by featural overlap. We had 15 related input pairs sharing on average around 5 features and 3 for a minimum. Another 15 unrelated pairs had at most 2 features in common. The first item in each pair stands for the concept denoted by the context word while the second stands for the meaning evoked by the picture.

In the simulation context meanings were presented to the network by clamping the state of the semantic layer to their assigned representations for a limited amount of time. Target meanings, in contrast, were presented to the network until the network settled, that is, until no lexeme unit changed its state by more than an output tolerance value. The time needed for the network to settle provided us with an estimation of naming latencies.

The three SOAs in the experiments were simulated by presenting the context meaning before the meaning of the target, at the same time or after it. Figure 1 presents the results. Both the mean latency differences and the size of the three SOAs are given in network time units. (The absolute time scale of the network is arbitrary.)

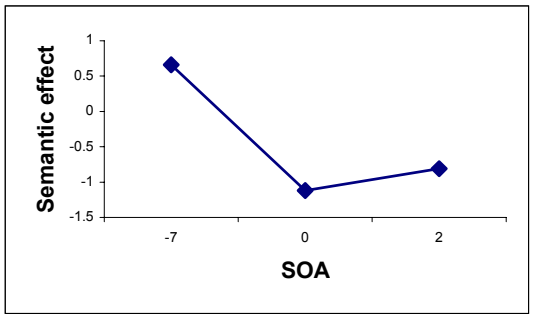


Figure 1: Mean latency differences between the related and unrelated conditions for three SOA values

The simulation replicated the experimental pattern of results: facilitation at an early SOA and inhibition for positive SOA values. Tracing the patterns of activation at both the hidden and the output layer suggested to us the following explanation of the effects. Related input patterns tend to have similar hidden layer representations or, at least, more similar than unrelated input patterns. Thus, it takes less time for a negative SOA to change the overall hidden layer configuration moving from a related context to the target than from an unrelated one. Therefore, the locus of the facilitation effect seems to be the hidden layer, that is, in our interpretation, the lemma level. Next, inhibition shows up at synchrony and at a short positive SOA as the lexeme node for a related context tends to be more active than the lexeme of an unrelated context making it a stronger competitor for the target node. On the other hand, processing of related and unrelated contexts did not seem to make much difference at the hidden layer. So, we could conclude that the locus of the inhibition effect is the lexeme level rather than the lemma level (see also Starreveld & La Heij, 1996). To conclude, the simulation offers a tentative explanation of both effects suggesting, however, that they reside at different levels and stages of processing.

The theoretical relevance of the simulation and the architecture we proposed intends to reach, however, farther than the simulation of the two opposite semantic effects we discussed. The key feature, we could say, is a distributed conception of lemma representation and the view that lemma retrieval corresponds to slipping into an attractor rather than activating above threshold a distinct lemma node. While the theoretical advantages of such a proposal are certainly in need of evaluation, we believe it could answer a major challenge put forward by Caramazza (1997): why should the speech production system be in need of lemma entities? The answer we propose is that lemmas play the role of a hidden layer in a connectionist network, namely they re-represent the semantic input (see also Dell, Schwartz, Martin, Saffran & Gagnon, 1997 for a similar suggestion). However, if lemmas continue to be conceived of as separate individual nodes as current models describe them (Dell, 1986; Levelt, Roelofs & Meyer, 1999), it is hard to see how such a re-representation could take place. Distributing lemmas across a set of features, on the other hand, can make them serve that purpose.

Conclusions

A graded featural view of semantic relatedness, as opposed to simple categorical relatedness, proves to be essential for ensuring purely semantic facilitation effects not only in word recognition but also in picture naming. Besides giving support to a featural account of semantic similarity and semantic memory organization, this also serves as a warning that categorical relations might be too rough a conception of semantic relatedness for the detection of potential semantic effects in various tasks.

Different loci for semantic facilitation and inhibition in picture-word experiments as a function of SOA are identified based on a simulation with an attractor neural network – the lemma level and the lexeme level in our interpretation. A hypothesis concerning the distributed nature of lemma representation is advanced with the more general goal of motivating the presence of lemmas in the speech production system.

Acknowledgments

The paper summarizes the main results reported in the master thesis of the first author under the supervision of the second. We thank Maurice Grinberg for his help in designing the simulation. We also thank Armina Janyan and Velina Balkanska for their help in designing and running the experiment. Finally, we are much indebted to Keith Stenning for his helpful comments and his review of the original material this paper is based on.

References

Alario, F.-X., Segui, J., Ferrand, L. (2000). Semantic and associative priming in picture naming. *The Quarterly Journal of Experimental Psychology*, 53A, 741-764.

Bates, E., D'Amico, S., Jacobsen, T., Szekely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C., Pechmann, T., Pleh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin and Review*, 10 (2), 344-380.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*, 177–208.

Carr, T.H., McCauley, C., Sperber, R.D. & Parmelee, C.M. (1982). Words, pictures and priming: On semantic activation, conscious identification, and the automaticity of information processing. *Journal of Experimental Psychology: HPP,* 8, 757-777.

Damian, M.F. & Martin, R.C. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Language, Memory and Cognition,* 25, 345-361.

de Groot, A.M.B. (1984). Primed lexical decision: Combined effects of the proportion of similar prime-target pairs and the stimulus asynchrony of prime and target. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology,* 36(A), 253-280.

Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review,* 93, 283-321.

Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M. & Gagnon, D.A. (1997) Lexical access in normal and aphasic speech. *Psychological Review*, 104, 801-838.

Gerganov, E., Ivancheva, L., Kurlova, R., Nikolov, V., Nikolova, T. (1984). *Bulgarski normi na slovesni asotsiatsii*. Sofia: Izdatelstvo Nauka i Izkustvo.

Glaser, W.R. & Düngelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: HPP,* 10, 640-654.

Glaser, W.R. & Glaser, M.O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General,* 118, 13-42.

La Heij, W. (1988). Components of Stroop-like interference in picture naming. *Memory and Cognition,* 16, 400-410.

La Heij, W., Dirkx, J. & Kramer, P. (1990). Categorical interference and associative priming in picture naming. *British Journal of Psychology,* 81, 511-525.

Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A Theory of Lexical Access in Speech Production. *Behavioral and Brain Sciences*, 22, 1-75.

Lupker, S.J. (1988). Picture naming: An investigation of the nature of categorical priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14, 444-455.

McRae, K. & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Language, Memory and Cognition,* 24, 558-572.

McRae, K., de Sa, V., Seidenberg, M. (1997). On the nature and scope of featural representation of word meaning. *Journal of Experimental Psychology: General,* 126, 99-130.

Neely, J.H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In Besner, D. & Humphreys, G.W. (eds.) *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.

Neely, J.H., & Keefe, D.E. (1989). Semantic context effects on visual word processing: A hybrid prospective / retrospective processing theory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory,* vol. 24. New York: Academic Press.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*, 263–269.

Perea, M. & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition,* 62, 223-240.

Plaut, D.C. (1995). Semantic and associative priming in a distributed attractor network. *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp 37-42). Hillsdale, NJ: Lawrence Erlbaum Associates.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition,* 42, 107-142.

Rosinski, R.R., Golinkoff, R.M. & Kukish, K.S. (1975). Automatic semantic processing in a picture-word interference task. *Child Development,* 46, 247-253.

Schriefers, H., Meyer, A.S. & Levelt, W.J.M. (1990). Exploring the time course of lexical access in speech production: Picture-word interference studies. *Journal of Memory and Language,* 29, 86-102.

Shelton, J.R. & Martin, R.C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology*: *Language, Memory and Cognition,* 18, 1191 – 1210.

Sperber, R.D.,McCauley, C., Ragain,R.D., &Weil, C.M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition*, 7, 339-345.

Starreveld, P.A. & La Heij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Language, Memory and Cognition,* 22, 252-255.

You Can't Play Straight TRACS and Win: Memory Updates in a Dynamic Task Environment

Hansjörg Neth (nethh@rpi.edu) **Chris R. Sims** (simsc@rpi.edu) **Vladislav D. Veksler** (vekslv@rpi.edu) **Wayne D. Gray** (grayw@rpi.edu)
Cognitive Science Department
Rensselaer Polytechnic Institute

Abstract

To investigate people's ability to update memory in a dynamic task environment we use the experimental card game TRACS™ (Burns, 2001). In many card games card counting is a component of optimal performance. However, for TRACS, Burns (2002a) reported that players exhibited a baseline bias: rather than basing their choices on the actual number of cards remaining in the deck, they chose cards based on the initial composition of the deck. Both a task analysis and computer simulation show that a perfectly executed memory update strategy has minimal value in the original game, suggesting that a baseline strategy is a rational adaptation to the demands of the original game. We then redesign the game to maximize the difference in performance between baseline and update strategies. An empirical study with the new game shows that players perform much better than could be achieved by a baseline strategy. Hence, we conclude that people will adopt a memory update strategy when the benefits outweigh the costs.

Introduction

Optimal performance in dynamic environments requires that we base our decisions on the current state of the world, not on past states. Radar operators must act on the basis of continuously changing variables such as plane altitude and heading. Drivers constantly need to monitor the current speed limit, posted road signs and the traffic behind and in front of them. Failure to mentally update these types of information can lead to dangerous decisions and catastrophic behavior. Even our chances to win at card games like Blackjack or Bridge are closely tied to our ability to count cards and update memory.

Previous research suggests that human ability to monitor and adjust to change is limited and dependent on various factors. Yntema (1963) found that people are better at tracking a small number of variables with a large range of values each, than a large number of variables with a small set of possible values each. In addition, reducing the frequency of update can improve performance. Other manipulations, such as increased predictability of a sequence, provide little or no advantage in remembering the current state of the environment. Venturino (1997) distinguished the memory capacity for static information from that for dynamically changing information and showed that the latter is highly limited, particularly when the to-be-remembered attributes are similar. Hess, Detweiler and Ellis (1999) added that update performance is improved when spatial invariants constrain where different data values are presented on a visual display.

In general, human rational behavior is constrained by the structure of task environments and the computational capabilities of the actor (Simon, 1990). To capture functional relationships of complex tasks while abstracting away from domain specific details we advocate the use of synthetic task environments, or microworlds (Gray, 2002). If the properties of the synthetic task environments are known and manipulable, the scope and limits of human rationality can be assessed. Moreover, the effects of environmental changes are tractable.

Table 1: Baseline distribution of cards in the deck. The back of every card shows only its shape, whereas the front shows both its shape and color.

Shape:	▲	●	■	▲	●	■
Color:	red	red	red	blue	blue	blue
Initial deck:	6	4	2	2	4	6

Straight TRACS

TRACS™ is a 'Tool for Research on Adaptive Cognitive Strategies,' designed and developed by Kevin Burns (2001, 2004). Being both entertaining card game and experimental research tool, TRACS provides a microworld which promises to bridge the gap between mathematical rigor and real-world relevance. We will limit our discussion to *Straight* TRACS, which is the simplest version of an entire family of games.[1]

TRACS is played with a deck of 24 cards. The back of each card shows one of three shapes—circle, triangle, or square—filled in with black. The front of each card shows both its shape, and one of two colors (red or blue). Table 1 shows the initial deck distribution for each of the six possible card types. This baseline information is always available to the player. As hands are played the number of cards remaining in the deck decreases, and the odds for each shape change accordingly.

At the start of a game, three cards are dealt in a row. The middle card is dealt face up (showing both its shape and color), while the left and right cards are dealt face down, showing their shape not their color. The task for the player is to choose the card, either left or right, most likely to match the *color* of the middle card. The chosen card is then turned over, revealing its color. If the chosen card matches the color of the middle card, a *hit* is credited to the player's score. A mismatch is scored as a *miss*. The two face up cards (the middle and the chosen card) are then removed from the game. On

[1] Online versions are available at www.tracsgame.com.

the next turn, the unchosen card is flipped over and becomes the new middle card, and two new cards are dealt face down to the left and right. A game lasts 11 turns, at which point there are not enough cards in the deck to deal another hand. A player's objective in TRACS is to maximize the number of hits.

As a probe of the player's assessment of odds at each turn, Burns (2002a, 2002b) added a confidence meter to the task. On each turn, players were presented with a red to blue color gradient for each of the two face-down cards. Prior to choosing a card, the participants used the gradient to indicate the likelihood of each candidate card to be red or blue. In another condition, Burns used a scale of nine buttons rather than a continuous spectrum. For consistency reasons all gradient estimates were rounded to the nearest button, corresponding to the nearest 12.5%.

Burns (2002a) characterized players' likelihood estimates as exhibiting a *baseline bias*; i.e., their judgments of odds deviate systematically from the actual odds in the direction of the initial card distribution. There are six types of color–shape combinations. Burns (2002b) reports that players could only monitor 2–4 types of cards with reasonable reliability. He concludes that the dual tasks of concurrently counting and normalizing numbers 'are naturally hard' and that continuously updating odds exceeded the cognitive capacity of the 'unaided mind' (Burns, 2002a, p. 159).

In the following sections, we will challenge this claim both theoretically and empirically. To preview our conclusions, we find that subtle constraints in the task environment can have profound effects on the strategy adopted by participants. The reported baseline bias is revealed as both rational and adaptive when considered in light of a cost-benefit analysis of the environment. We then demonstrate that players will adopt a more effortful memory strategy if the cost-benefit structure of the environment rewards this.

Tracking TRACS

Given the original finding that players find it challenging to succeed at TRACS, a natural starting point for our investigation is a task analysis. What specifically makes this game so difficult to play?

Task Analysis

In describing TRACS as a game of 'confidence and consequence' Burns (2001, 2002b) distinguishes two subtasks of diagnosis and decision. On each turn, a player first provides an odds judgment for each face-down card and then chooses one on the basis of these estimates.

Extending Burns' analysis, we suggest that each turn involves a minimum of *three* distinct cognitive tasks: a memory retrieval task, an odds conversion task, and a decision task (see Figure 1). The first subtask on each turn consists in remembering how many cards of each candidate shape and color remain in the deck. As the initial card distribution is provided in terms of frequencies and players encounter card instances through a process of natural sampling, we assume that this retrieval is framed in terms of natural frequencies. Secondly, the retrieved frequencies need to be converted into odds, which is a non-trivial process involving Bayes' rule for natural frequencies (Gigerenzer, 2000). For example, to determine the likelihood of a red triangle, a player has to divide the number of red triangles currently left in the deck by the sum of red and blue triangles left in the deck. As people are notoriously bad at dealing with probability information (see Gigerenzer, 2000, and Koehler, 1996, for reviews) it is conceivable that this translation process incurs a loss of accuracy. If so, merely asking for likelihood estimates confounds memory updates with probability judgments and may underestimate players' true memory capacity. As a third subtask, a player needs to integrate all estimates and decide which candidate card is more likely to score a hit on the current trial.

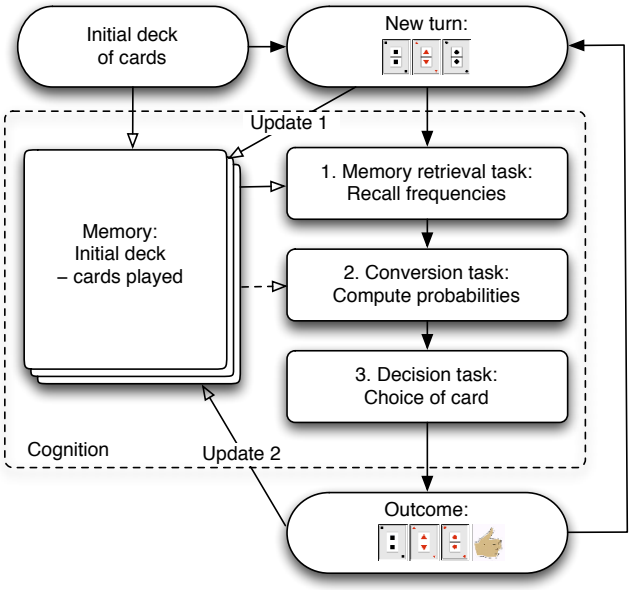


Figure 1: Subtasks and memory updates required on each turn of Straight TRACS.

In addition to these three subtasks, each turn requires two distinct updates of memory. The first update is necessary as soon as the middle card is revealed. If the middle card happens to be a red triangle, the player needs to realize that there now is one less red triangle left in the deck. The second update ought to occur at the end of a turn when the chosen card is revealed. This second update is critical, as at this point in the game, players may be distracted by focusing exclusively on the correctness of their choice and ignoring the additional information revealed.

This task analysis reveals both the complexity and simplicity of TRACS. On one hand, multiple subtasks and memory update requirements make the game quite challenging. Even if frequency information on card types was readily available, the conversions into probabilities, comparisons between odds, and selection of cards introduce potential sources of error. On the other hand, remembering and updating a list of six numbers (representing the current frequency of each card type) does not in itself seem beyond the capacity of human memory.

The Impact of Memory

At first glance, it seems that TRACS is a 'memory game' (Burns, 2001, 2002a) in which players can succeed only by remembering which cards have left the deck. However, our

experience playing TRACS casts some doubts on the importance of memory. Due to the random card selection process a typical game contains many knowledge-indeterminate turns. For example, whenever both face-down cards show the same shape, a player has no choice but to guess. Likewise, both face-down cards frequently have the same color, so that the player scores a hit or miss regardless of knowledge or choice. Even when the cards differ in shape, color, and odds, it is possible that selecting the card with higher actual odds results in a miss, whereas choosing the 'wrong' card scores a hit.

These concerns raise questions about whether memory really matters. To what extent can poor performance be blamed on failures of memory? Would better memory improve performance? The non-deterministic nature of the game makes it hard to answer these questions analytically; thus, we implemented the game as a computer simulation.

Simulation As Allen Newell and Herbert Simon famously stated, "Just as a scissors cannot cut paper without two blades, a theory of thinking and problem solving cannot predict behavior unless it encompasses both an analysis of the structure of task environments and an analysis of the limits of rational adaptation to task requirements." (1972, p. 55). In this spirit, we created a simulation in MATLAB™ in which 'pure' cognitive strategies could be formalized and implemented. By running these artificial agents for thousands of trials, we were able to determine precise performance levels, despite the dynamic and nondeterministic aspects of the game.

We compared four cognitive agents that differed in their memory resources and strategies, but did not make any errors in odds translation or judgment. A *baseline* agent has perfect knowledge of the initial deck distribution, but is amnesic with regards to the cards played during a game. In contrast, the *update* agent enjoys perfect memory of every hand played, and bases all choices on the actual odds at any given moment.

Two additional agents bracket the performance of baseline and update agents: *random* agent has neither memory nor knowledge of the initial distribution, and hence is forced to blindly guess at every turn. On the other end of the scale, *omniscient* agent effectively enjoys X-ray vision and can observe the colors of both candidate cards, allowing for optimal card selections without the need for memory or odds estimates.

The mean score for the random agent across 10,000 simulated games was 5.24 (out of 11 possible) hits per game. To our surprise, baseline and update agents performed about the same, scoring 6.57 and 6.79, respectively. Thus, the average performance difference between the baseline and update agents was roughly two tenths of one point per game. Further, both strategies achieved only marginally better scores than the random strategy.

Figure 2 shows the mean percentage of hits per turn for each agent. It is obvious that the performance of baseline and update agents are very similar, except for an increasing benefit of update strategy late in a game. The entire range between random and omniscient performance scores is only 25%, which is essentially due to 25% of all turns not allowing for a hit.

While an optimal update agent acts to maximize performance regardless of the effort involved, humans have limited cognitive resources and are required to negotiate cost–benefit tradeoffs (Anderson, 1990; Simon, 1990, 1992). Given these

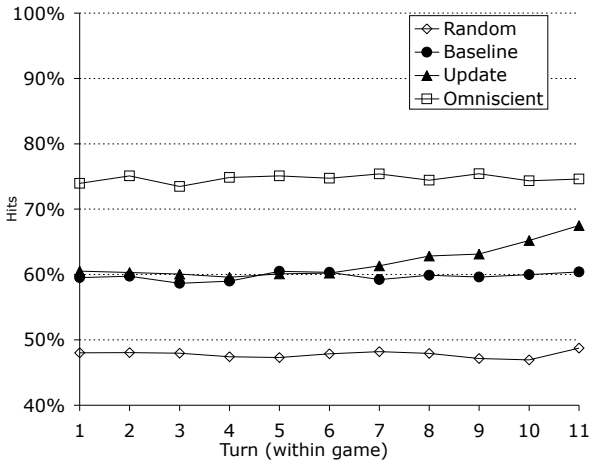


Figure 2: Simulation results for four artificial agents playing 10,000 games of original TRACS.

constraints and the minimal benefits of an update strategy, participants might well have adopted a baseline strategy for good reasons. Thus, our analysis suggests a re-interpretation of Burns' original findings: In Straight TRACS, memory update yields no performance benefit over adopting a much easier baseline strategy. Hence, adopting the baseline strategy is both adaptive and *rational*.

TRACS*

The simulation results suggest that—by not offering an incentive to a memory update strategy—Straight TRACS is inadequate for investigating people's willingness and capacity to monitor and update changing environmental circumstances. In this section we introduce TRACS*, which provides a clear benefit for adopting an update strategy, as well as introduces additional probes of memory performance.

In designing TRACS*, we sought to create a variant of the game for which a memory update strategy clearly benefits performance. We achieved this by carefully controlling the cards dealt to the players. While cards were selected randomly, they were selected from a card space constrained by two rules. First, only pairs of face-down cards that would not have equal odds of matching the target color would be dealt. By eliminating ties, this rule eliminates the need to guess. Second, pairs were not selected if the card with the lower odds resulted in a hit, or if the card with the higher odds did not. This rule aimed to reduce the influence of luck by eliminating win-win and lose-lose situations, thus driving a wedge between the baseline and update strategies.

Figure 3 illustrates the effects of these changes. The mean score for the random agent in TRACS* remained stable, at 5.49 (out of 11) hits per game across 10,000 games. However, baseline and update scores rose to 8.22 and 10.83, respectively. Hence, our game modifications were successful in introducing a substantial benefit of the update strategy over the random and baseline strategies. Given that baseline and update strategies now yield unique performance signatures, it should be possible to determine which strategy our participants actually adopt in the game.

Our second alteration in TRACS* was procedural. In addi-

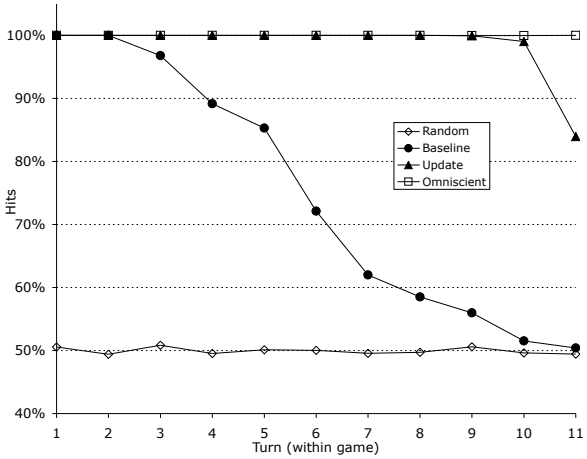


Figure 3: Simulation results for four artificial agents playing 10,000 games of modified TRACS*.

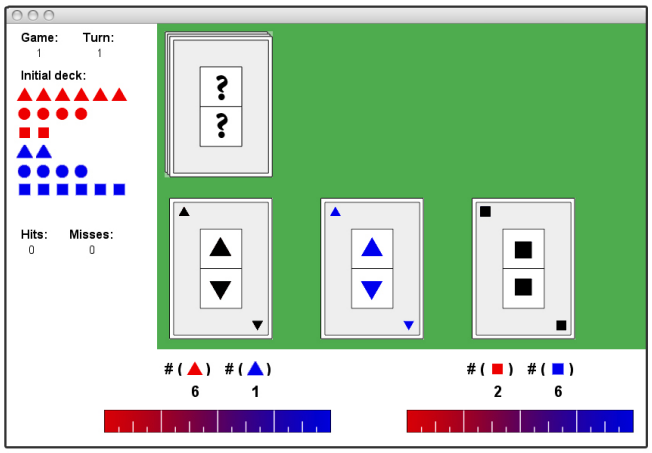


Figure 4: Screenshot of the TRACS* interface requesting odds estimates (after the completion of the recall task).

tion to using continuous color gradients to assess our participants' odds calculations, we introduced memory recall boxes to judge the accuracy of their memory. In this way we hoped to elucidate whether Burns' findings indicated an actual baseline bias, or merely just difficulty in converting accurately recalled frequencies into points along a likelihood gradient.

Experiment

Method

Twenty-five undergraduates from Rensselaer Polytechnic Institute participated in partial fulfillment of a course requirement. They ranged in age from 18 to 22 years, with an average of 19.6 years. Participants were tested individually.

The experimenter spent about ten minutes instructing each participant on the rules of original TRACS. Each participant played a total of 10 games of 11 turns each. On every turn, players had to complete the recall task, provide odds estimates, and choose a card.

On the newly added recall task participants were asked, for each face-down card, to report the number of red and blue cards of that shape which remained in the deck. Answers were typed into text boxes immediately below each face-down card. Players then estimated the odds of each face-down card being red or blue by placing a marker on a continuous color gradient. Gradients were red on the left and blue on the right, and 300 pixels wide (≈10 cm), allowing for a precision below one percent (see Figure 4 for a screenshot). Finally, participants chose a card by clicking on it. Feedback on correctness was then provided by a thumbs-up/thumbs-down image and the next turn was initiated by clicking on the feedback image.

The game was implemented in Macintosh Common Lisp 5.0 running on OS 10.2 with a 17" flat panel display set to a 1024×768 screen resolution. The initial card distribution and a hit/miss counter were shown to the left of the game window.

Results

We will assess participants' performance before turning to more detailed analyses of various error types.

Performance TRACS* allows for a straightforward correspondence between a player's awareness of the current game state and his or her outcome score. Thus, scores reliably exceeding the expected values of a simulated baseline agent would signal a memory update strategy.

On average, participants scored 9.3 hits per game with 22 out of 25 players (88%) exceeding the theoretic baseline score of 8.2 hits. This strongly suggests that memory updates contributed to task performance.

To allow for a statistical assessment of these differences, we let our simulated baseline and update agents both play the same number of games as human participants. A comparison of mean scores over the sequence of ten games per player showed that human players scored significantly more points than baseline agents [9.3>8.2, $t(26)=2.1$, $p<.001$], and significantly fewer hits than update agents [9.3<10.8, $t(25)=2.1$, $p<.001$]. Figure 5 contrasts the performance of human participants with that of simulated agents on a within-game resolution. It is obvious that human players did not perform on

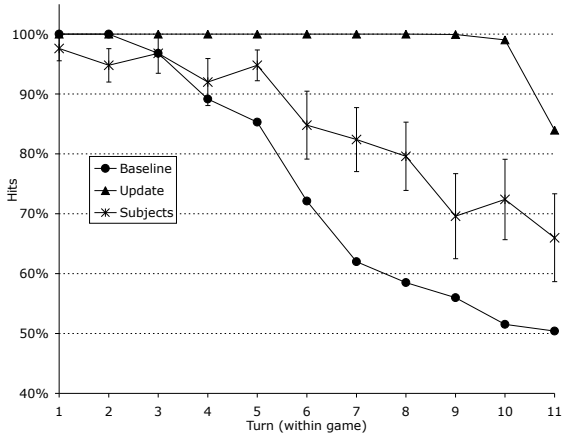


Figure 5: Participants' mean percentage of hits by turn compared to those of simulated baseline and update agents. (Error bars indicate 95% confidence intervals.)

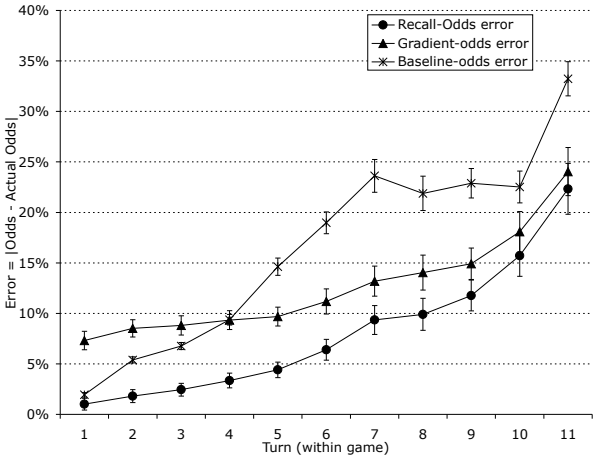


Figure 6: Average errors of odds by turn. (Error bars indicate 95% confidence intervals.)

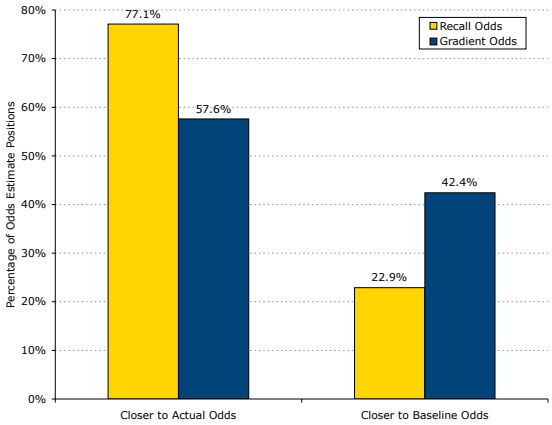


Figure 7: Percentage of odds selections closer to the baseline vs. closer to the actual value (based on n=4404 estimates).

the level of an ideal update agent, but did reliably better than a baseline agent.

To assess possible effects of learning we conducted an ANOVA with game number as a within-subjects factor. A significant main effect [$F(9,216)=3.0$, $p<.01$] indicated that players improved their scores reliably from an average of 8.8 hits in earlier to about 9.7 hits in later games. Subsequent comparisons showed that human participants outperformed a pure baseline agent in all but the initial two games.

Errors Even though human participants performed better than a baseline agent, their performance was worse than that of an ideal update agent. In this section, we examine this discrepancy by first considering erroneous frequency and likelihood estimates before assessing errors of internal consistency.

As participants estimated card frequencies as well as likelihoods we were provided with two distinct indices of memory. To allow for direct comparisons of both indices on a single scale, we converted reported frequencies into 'recall odds'. For both recall odds and likelihood estimates (as indicated on the gradient scales) we then calculated and summed up the absolute difference from the actual odds.

Figure 6 illustrates that both recall-odds and gradient-odds errors increase over the course of a game, but errors in frequency recall (with a mean of 8.0%) are significantly lower than the errors in likelihood estimates provided on gradient scales (12.6%). The third line in Figure 6 shows the mean size of the 'baseline-odds' error (16.5%) which would result if participants had adopted a baseline strategy on the given trial. Even though the mean gradient-odds error exceeded the baseline-odds error on the first three trials, the general trend indicates that participants' actual errors on both scales were lower than suggested by a baseline bias.

Taking into account the direction of deviations rather than just error magnitudes, we can also ask whether empirical recall and gradient odds are closer to the baseline or to the actual odds. Whenever the actual odds value deviates from the baseline value there are two possible attractors: Participants might specify odds closer to the baseline odds, or they might select odds closer to the actual odds. A *bias* is defined by a systematic preference. If participants—due to update failure or memory decay—were more likely to choose odds closer to the baseline than to the update value this would constitute a baseline bias. Likewise, an "update bias" could be diagnosed if participants were more likely to select odds in the vicinity of the actual value. Figure 7 shows that, in TRACS*, the evidence for an update bias clearly outweighs the evidence for a baseline bias. Participants' preference for actual values seems particularly pronounced when odds are based on recall frequencies (77.1% vs. 22.9%). In contrast, the same preference is weaker when odds estimates are measured by probability gradients (57.6% vs. 42.4%). As the baseline attractor seems to exert less gravitational pull when providing frequency estimates than when responding on a gradient scale, examining only the latter (e.g., Burns, 2002a, 2002b) might overestimate the size of a baseline bias.

All errors reported so far were deviations of empirical estimates from either true or baseline values. Our finding that participants' frequency estimates are closer to the actual values than to the initial baselines makes it implausible that participants' frequency estimates are governed by a baseline bias. At the same time, it raises questions about alternative breakdowns in performance. On the basis of our initial task analysis, the complexity of TRACS allows for a variety of non-memory related errors. In the following and final sections we consider conversion errors and errors of choice as examples of errors of internal consistency.

Due to our sequential procedure of first requiring frequency information and then asking for probability estimates, participants' responses on the likelihood gradients ought to be a direct function of recall performance. Nonetheless, people's notorious problems with probabilities can cause *conversion errors* when transforming recalled frequencies into odds on continuous scales. To assess the occurrence of such errors, we compared subjective recall odds (based on the card frequency entries of each participant and turn) with the likelihood estimates provided on the same turn. An average deviation of 6.6% indicates that this translation process was indeed non-trivial and error-prone. The magnitude of this error is striking not only as it is almost as large as the average error in fre-

quency recall (8.0%, see Figure 6), but also when considering that players reported their subjective frequencies immediately before indicating their judgment of odds and had all relevant frequencies displayed directly above the gradient scales (see Figure 4). Thus, we conclude that a large proportion of participants' error-prone responses on likelihood scales were due to errors in odds conversion.

Two curious errors of internal consistency address the relation between odds estimates and card selections. *Recall-choice errors* can be defined as instances in which the card with lower recall odds (based on the subjective card frequency estimates) is selected by the participant. Similarly, *gradient-choice errors* occur whenever the card with lower likelihood odds (based on probability estimates) is chosen.

There were 4.3% (119 out of 2750 choices) recall-choice errors, but 8.3% (229) gradient-choice errors. Given that any conflict between judgment and choice is relatively bizarre, both errors are more frequent than we would have expected. As the gradients are evaluated immediately before a choice is made, we interpret the relative size of both errors as evidence that players were more likely to base their choices on perceived frequencies than on perceived odds.

Discussion

Our first result is of a methodological nature: When creating artificial task environments to assess the scope of human rationality, the cost–benefit structure of the task must provide an incentive to display the behavior in question. Our simulation of Straight TRACS revealed that the original game provides only minimal benefits for adopting an effortful memory update strategy. This led us to re-interpret Burns' (2002a, 2002b) original finding of a 'baseline bias' as an adaptive and rational response to the properties of the task environment.

Our critique, however, does not imply that TRACS is not an interesting game and valuable research paradigm—quite to the contrary! We now believe that TRACS is both more complex and more interesting than it at first appeared. Our task analysis has suggested the need to distinguish three cognitive components: retrieving numbers of cards from memory, converting frequencies into probabilities, and mapping frequency or probability estimates to choices.

We are particularly intrigued by the errors our players made when converting natural frequency information to likelihood estimates. Players who had to provide the same information in two different formats within seconds and saw the frequencies displayed in front of them while computing probabilities still made substantial errors when coming up with simple likelihood estimates. Interestingly, our analysis of choice errors revealed that players seemed less likely to act on their inaccurate probability estimates than on their perceived frequencies even though the former just preceded their choice.

A potential caveat of our study is that by altering the cost-benefit structure of the task and assessing players' memory for card frequencies we introduced *two* changes to the original game. It is conceivable that the mere query for frequencies made the necessity to count cards more explicit, whereas it remains rather implicit in the original game. The extent to which each of our modifications contributed to the improved performance and to which a procedural task demand may have inadvertently prompted different memory strategies is an empirical question to be addressed in future studies.

Finally, the performance results of our modified version TRACS* provide a more optimistic view of the human capacity for concurrent memory updates than do previous studies. As our players were able to reliably exceed baseline performance, we conclude that the previously reported 'baseline bias' may be an artifact of the original game.

Despite our criticisms, our results agree with those of Burns (2002a, 2002b) that people are able to take baserate information into account. However, we additionally demonstrate that—when memory matters—people are also able to dynamically update their memory while being engaged in a highly demanding task.

Acknowledgments

We are grateful to Kevin Burns for allowing us to use TRACS and providing many helpful comments. In addition, we thank Christopher Myers, Bram van Heuveln and Jamie Sowder for many valuable contributions. The work reported was supported by grants from the Air Force Office of Scientific Research (AFOSR #F49620-03-1-0143), as well as the Office of Naval Research (ONR #N000140310046).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Burns, K. (2001). TRACS: A tool for research on adaptive cognitive strategies: The Game of Confidence and Consequence. At www.tracsgame.com (May 2004).

Burns, K. (2002a). On Straight TRACS: A baseline bias from mental models. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 154-159. Hillsdale, NJ: Lawrence Erlbaum.

Burns, K. (2002b). Dealing with TRACS: The game of confidence and consequence. *Proceedings of the American Association for Artificial Intelligence*, Symposium on Chance Discovery.

Burns, K. (2004). Making TRACS: The diagrammatic design of a double-sided deck. *Proceedings of the 3rd International Conference on the Theory and Application of Diagrams.*

Gigerenzer, G. (2000). *Adaptive thinking. Rationality in the real world.* Oxford, UK: Oxford University Press.

Gray, W.D. (2002). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. *Cognitive Science Quarterly 2(2)*, 205–227.

Hess, S.M., Detweiler, M.C. and Ellis, R.D. (1999). The utility of display space in keeping track of rapidly changling information. *Human Factors 41(2)*, 257–281.

Koehler, J.J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19*, 1–53.

Newell, A., and Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Simon, H.A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–19.

Venturino, M. (1997). Interference and information organization in keeping track of continually changing information. *Human Factors, 39(4)*, 532–539.

Yntema, D.B. (1963). Keeping track of several things at once. *Human Factors 5*, 7–17.

Defining New Words in Corpus Data: Productivity of English Suffixes in the British National Corpus

Eiji Nishimoto (enishimoto@gc.cuny.edu)
Ph.D. Program in Linguistics, The Graduate Center
The City University of New York
365 Fifth Avenue, New York, NY 10016 USA

Abstract

The present study introduces a method of identifying potentially new words in a large corpus of texts, and assesses the morphological productivity of 12 English suffixes, based on some 78 million words of the written component (books and periodicals) of the British National Corpus (BNC). The method compares two corpus segments (created by randomly sampling at the level of documents within the BNC), and defines new words as those that are not shared across segments (segments being interpreted as randomly sampled speaker groups). The approach taken differs from others in the literature in that new words are identified irrespective of how many times a given word is used by the same speaker (author). A productivity ranking of the 12 English suffixes is obtained, and the results are shown to be intuitively satisfying and stable over different sample sizes. With a psycholinguistic interpretation of the data, implications for the nature of intuitions about productivity are considered.

Introduction

Morphological productivity is central to the study of word formation, but it continues to defy a solid, uniform description (see e.g., Aronoff, 1976; Bauer, 2001; Plag, 1999). The coinage of a "new" word is abundant in our daily use of language; for example, a person who is being gossiped about may be referred to as a *gossipee*, or a used book may be *cleanish*. Affixation in English (as in *gossip* + *-ee* → *gossipee*; *clean* + *-ish* → *cleanish*) is a productive word formation process, and there is plenty of evidence that affixes differ in their *degree of productivity* (e.g., Aronoff, 1976; Bauer, 2001); for example, words can in general be formed more easily with *-ness* than with *-ity* (and thus we may accept *cleanness* but not *cleanity*). The majority of researchers investigating the issue of productivity are interested in accounting for varying degrees of productivity, and several productivity measures have been proposed in the literature (e.g., Aronoff, 1976; Baayen, 1992, 2001; Bauer, 2001; Plag, 1999). Assessing the degree of productivity, however, has proven to be a complex task (Bauer, 2001): while the consensus seems to be that capturing the coinage of new words is essential in assessing productivity, there is an inherent difficulty in defining what a "new" word is.

Most notable among previous studies is a corpus-based approach proposed by Baayen (1992, 2001). Based on word frequency in a large corpus of texts, his productivity measure is formulated as $P = n_1/N$, where given a particular affix, n_1 is the number of word types with that affix that occur only once (the so-called *hapax legomena*, hereafter *hapaxes*), N is the sum of word tokens with that affix, and P is the productivity index.[1] P is interpreted as expressing the probability of encountering a word type with a given affix that has not been seen in the sampled corpus. Thus, new words are defined under this measure as "unseen" words in a corpus. An important characteristic of P is that it is based on token frequency—N directly refers to a count over tokens, and a word is included in the n_1 count only if it occurs just once. The measure P, with its focus on hapaxes as estimators of unseen words, is motivated by the probability estimation method of Good (1953)—or the *Good-Turing* estimation method (Church & Gale, 1991).[2]

While a dictionary provides another source of data for quantifying morphological productivity, a corpus-based approach has many advantages. A large corpus of texts contains productively formed words that are typically not listed in a dictionary (e.g., *gossipee*), and corpus data reflect how words are actually used (Baayen & Lieber, 1991; Baayen & Renouf, 1996).

The present study pursues and extends the corpus-based approach by introducing a new method of identifying new words and assessing productivity.

Type Frequency and Deleted Estimation

It has been suggested that the type frequency for an affix (the number of word types with an affix) in a corpus, represented by V, is inadequate in expressing its degree of productivity. Baayen and Lieber (1991: 804) point out that in their reference corpus of 18 million words, the type frequencies for *-ness* (497) and *-ity* (405) do not adequately express the fact that *-ness* is intuitively felt to be much more productive than *-ity*. They find that the P indices for *-ness* (0.0044) and *-ity* (0.0007) are more in line with linguists' intuitive estimates for these suffixes. There are, however, some aspects of the measure P that can be quite counter-intuitive. In Baayen and Lieber (1991), for example, the P index for verbal suffix *-ize* (0.00007) is substantially lower

[1] As is usually the case in a corpus study, the term *token* refers to each occurrence of a word, and the term *type* refers to each distinct word. For instance, if we have {*awareness*, *fairness*, *fairness*, *sharpness*, *sharpness*}, the *token frequency* for *-ness* is 5 (the sum of all occurrences of *-ness*), whereas the *type frequency* for *-ness* is 3 (the number of distinct words with *-ness*).

[2] For more detail, see Baayen (2001).

than that for *-ity* (0.0007). To correctly interpret these data, we need to take into account the fact that *P* is dependent on token frequency, and that verbs and nouns generally differ in their overall frequency in a corpus (Baayen & Lieber, 1991). Consequently, an across-the-board comparison of affixes across lexical categories is ruled out.

The view that type frequency in a corpus is problematic for assessments of degree of productivity holds only if type frequency alone is examined, for an entire corpus. A use of type frequency is suggested by Nishimoto (2003) in a productivity measure that adopts the mechanism of *deleted estimation* (Jelinek & Mercer, 1985; see also Manning & Schütze, 1999: 210–211), a probability estimation method used in Language Technology. The basic concept underlying the proposed productivity measure is the cross-comparison of corpus segments to identify word types that are not shared. The P_{DE} measure, a productivity measure based on the deleted estimation method, is formulated as:

$$(1) \quad P_{DE} = \frac{V_0^{AB} + V_0^{BA}}{V^A + V^B} = \frac{(V_0^{AB} + V_0^{BA})/2}{(V^A + V^B)/2} = \frac{V_N}{V}$$

Given a particular affix and two corpus segments *A* and *B* (both of some size *m*), V^A is the number of word types with that affix that are present in segment *A*, and V_0^{AB} is the number of word types with that affix that are present in segment *A* but are absent (unseen) in segment *B*. V^B and V_0^{BA} are defined similarly. Averaging the elements of the denominator and of the numerator separately, we obtain *V*, the total number of word types with that affix in a corpus segment of size *m*, and V_N (V-New), the number of *otherwise-unseen* word types with that affix (unseen being dependent on the relationship between segments). P_{DE} expresses the degree of productivity of an affix as the likelihood that a given word type with an affix will be unseen, hence potentially new. In addition to *V* and V_N, we also define V_{NN} (V-Non-New) as the number of word types with the relevant affix that are seen in both segments, hence non-new. What is essentially achieved by the P_{DE} measure is the division of sampled word types (*V*) into new word types (V_N) and non-new word types (V_{NN}). The relationship $V = V_N + V_{NN}$ holds in each application of the measure.

What are the grounds for associating new words with words that are not shared by two corpus segments? In the *British National Corpus* (BNC), data from unique sources are sampled in single documents, and thus each document could be considered to represent a set of words used either by one speaker (author) or by a few speakers (co-authors). Randomly distributing these documents into two corpus segments therefore gives us two groups of randomly sampled speakers (and the words that they used). Words that are not used in common by the two speaker groups are more likely to be new than words used by both groups. Cross-comparing two corpus segments offers a crude yet computationally simple method of separating words into potentially new words and potentially non-new words.[3]

Simulation Environment

We will examine the performance of the P_{DE} measure, based on the written component (books and periodicals) of the BNC, which offers some 78 million words sampled in 2,688 documents. Each sampled document is randomly assigned to one of two corpus segments, until each segment has a specified number of words in total (say, 30 million words). Documents are sampled without replacement, so no document is shared by two corpus segments. One simulation run (i.e., one application of the P_{DE} measure) consists of creating two corpus segments (as above) and obtaining values for *V*, V_N, and P_{DE}, based on formula (1).

Table 1 lists 12 English suffixes and 1 non-suffix control selected for the current study.[4] At least one suffix is included for each major lexical category.

Table 1: 12 English suffixes and 1 non-suffix control.

Suffix	Category	Prediction
-ness, *-ity*	Nominal	*-ness* > *-ity*
-er, *-ee*	Nominal	*-er* > *-ee*
-ion, *-ment*	Nominal	*-ion* > *-ment*
-th	Nominal	Unproductive
-ish, *-ous*	Adjectival	*-ish* > *-ous*
-ize, *-ify*	Verbal	*-ize* > *-ify*
-ly	Adverbial	Productive
ch#	Noun ending	Unproductive

The predicted differences in productivity in the last column of Table 1 are largely based on views expressed in the literature. We also examine *ch#*, the word ending of a noun (as in *church*), as a presumably unproductive non-suffix control that provides a baseline for determining whether suffixes are productive (or unproductive). Different semantic patterns among words formed with a suffix are ignored: for example, *amputee*, *absentee*, and *employee* exhibit different semantic patterns, but they are collectively treated as *-ee* words. We do not distinguish words with a suffix by the class of bases that the suffix attaches to: for example, *-er* includes *employer* (verb base) and *islander* (noun base). Ordinal numbers are excluded from *-th*.

A database of 17,347 word types representing the 12 suffixes and the non-suffix control was compiled, based on 100 million words occurring in the entire BNC. The database crucially relies on decisions about what constitutes a word type with a suffix. Most problematic are prefixation and compounding, which could dramatically increase the number of word types with a suffix. Removing all prefixes

[3] Nishimoto (2004) offers more detailed exploration of the mechanism of the P_{DE} measure, by increasing the number of cross-compared corpus segments (speaker groups) to 6.

[4] These are suffixes whose productivity is often discussed in the morphology literature. We focus on suffixes only, as they play a more prominent role than do prefixes in English word formation.

has some negative consequences, such as *encouragement* → **couragement* or *disagreement* → *agreement*.[5] On the other hand, allowing all prefixes does not seem plausible, since words such as *anti-institution* that appear to be cases of prefixation would count as distinct word types with a suffix. Compounding poses a similar problem, and the issue is further complicated by the variable hyphenation of words. In solving this familiar problem, we make use of entries in the *Oxford English Dictionary* (OED) and *Webster's Third New International Dictionary* (WD). All prefixed forms and compounds are checked against the OED/WD, and any preceding part of a word that cannot be spelled without a hyphen in both the OED and WD is removed. As a result, for example, *anti-institution* will be treated as *institution*, but *disagreement* will remain as *disagreement*. With the assumption that the OED and WD are conservative in accepting novel word forms, the current treatment effectively prevents novel cases of prefixation and compounding from inflating the count of word types with a suffix. Each word type in the database was inspected to exclude errors (e.g., misspelled words, words with a pseudo-suffix).[6] See Nishimoto (2004) for further detail.

Evaluation of Data

Productivity Indices

Table 2 presents mean values for V, V_N, and P_{DE}, averaged over 100 simulation runs, with 30 million words in each of the two corpus segments required by the P_{DE} measure (i.e., the total of 60 million words were sampled in each run). The suffixes in Table 2 are sorted by their P_{DE} value, to achieve a productivity ranking.

Suffixes *-ish* and *-ness* meet our expectations by being found at the more productive end of the ranking (although we might have expected *-ness* to be more productive than *-ish*), and *-th* and *ch#* fall at the less productive end. We consider the P_{DE} index for *ch#* to arise from processes other than affixation, such as the coinage of simplex words, compounding, or some sources of noise including the occurrence of rare or obsolete words.

Taking *ch#* as a baseline for determining productivity in affixation, we find that *-th* is unproductive. The finding that *-ment* is effectively non-productive matches Bauer's (2001: 8–9) observation that the productivity of *-ment* has been in decline so that new *-ment* words are synchronically rare.

Table 2: Mean values of the P_{DE} measure.

	V	V_N	P_{DE}
-ish	261.3	90.6	0.347
-ness	1354.9	431.2	0.318
-ee	88.6	26.1	0.295
-ize	437.6	114.5	0.262
-ity	1008.5	234.4	0.232
-er	2517.8	558.6	0.222
-ly	3585.0	754.3	0.210
-ify	105.8	21.1	0.199
-ous	639.1	107.1	0.168
-ion	2152.9	348.7	0.162
-ment	424.2	61.6	0.145
ch#	213.6	29.7	0.139
-th	40.9	3.5	0.085

The high productivity of *-ee* is somewhat unexpected, on its face. Based on the measure *P*, Baayen and Lieber (1991) also find *-ee* (0.0016) to be more productive than *-er* (0.0007), and they attribute the high productivity of *-ee* to the "vogue" nature of this suffix, as suggested by Marchand (1969).

In contrast to *-ee*, the productivity of *-er* is lower than we might have expected. Also lower than expected is the P_{DE} index for *-ly*. The result for *-ly* seems unsatisfactory considering the high regularity in *-ly* word formation—the suffix attaches to almost any adjective to form an adverb, with few restrictions (Aronoff, 1976: 37 fn 4; Baayen & Renouf, 1996: 82–83). The low P_{DE} indices for *-er* and *-ly* might be thought to arise from large values of V for these suffixes; however, Spearman's test shows no significant correlation between V and P_{DE}, $r_s = .203$, $p > .10$. We will return later to the data for *-er* and *-ly*.

Overall, we find that the P_{DE} measure yields results that largely accord with the productivity expected for the suffixes examined.

Sample Size Dependency

A question that naturally arises in evaluating the P_{DE} measure is to what extent the measure is dependent on sample size. Could it be the case, for example, that the productivity ranking of suffixes would differ markedly if the two corpus segments were smaller? Table 3 presents P_{DE} as a function of corpus segment size.[7] Again, each P_{DE} value is a mean over 100 simulation runs.

We find that P_{DE} values are remarkably similar across three series with different corpus-segment sizes. Friedman's test finds no significant difference in P_{DE} among the three, $\chi^2(2,13) = 2.627$, $p > .10$. Spearman's test shows that the P_{DE} indices are highly positively correlated: for 10 vs. 20 million words, $r_s = .990$, $p < .01$; for 10 vs. 30 million words, $r_s = .971$, $p < .01$; and for 20 vs. 30 million words, $r_s = .984$, $p < .01$. Thus, the P_{DE} measure offers a consistent

[5] Removing *dis-* from *disagreement* appears to be undesirable if we view *disagreement* as a nominalization of *disagree*.

[6] There are 6,797 rules defined for these corrections (mostly generated automatically, but some inevitably defined manually for cases such as "dona-a-a-ation" → *donation*). The number of rules is large, but it must be noted that some are needed to obtain correct forms for irrelevant words (so that they can be deemed irrelevant), and that a given word can be misspelled in a number of ways. Errors in a corpus cannot be overlooked. Evert and Lüdeling (2001) point out, for example, that each error in a corpus typically occurs only once and could greatly distort the number of hapaxes.

[7] As is clear from the formulation of the measure, a change in corpus segment size applies simultaneously to both corpus segments.

characterization of the productivity of these suffixes over different sample sizes.

Table 3: P_{DE} as a function of corpus segment size.

	P_{DE}		
	10 million	20 million	30 million
-ish	0.322	0.332	0.347
-ness	0.371	0.336	0.318
-ee	0.313	0.301	0.295
-ize	0.262	0.260	0.262
-ity	0.238	0.235	0.232
-er	0.260	0.236	0.222
-ly	0.226	0.215	0.210
-ify	0.179	0.191	0.199
-ous	0.164	0.165	0.168
-ion	0.169	0.163	0.162
-ment	0.153	0.148	0.145
ch#	0.164	0.153	0.139
-th	0.078	0.081	0.085

Token Frequency of New Words

One advantage of the P_{DE} measure is that new words in a corpus are identified in a way that is not solely dependent on token frequency. To ensure that advantage, it is crucial to implement the measure by creating corpus segments via random sampling at the level of documents (hereafter *RD*), rather than random sampling at the level of words (hereafter *RW*). The data presented in the preceding sections arise in implementations using RD.

If we were to follow RW, which words become identified as new would be dependent on their token frequency in the whole corpus.[8] Under the P_{DE} measure, a word is identified as new if all its tokens are distributed into only one corpus segment. If we were to randomly distribute words into two corpus segments (i.e., RW), the probability P that word *w* with token frequency *r* (in the whole corpus) will be identified as new is given by: $P(w\text{: new}) = 2(0.5^r)$. Figure 1 shows how P(*w*: new) changes as a function of *r*. We find that words that occur more than a few times in the whole corpus are highly unlikely to be identified as new. Hapaxes are exceptional in that they are guaranteed to be new, regardless of whether RD or RW is adopted. What is of interest regarding the difference between RD and RW is how many *non-hapaxes* are found to be new.

[8] The *whole corpus* here refers to the set of data used to create two corpus segments.

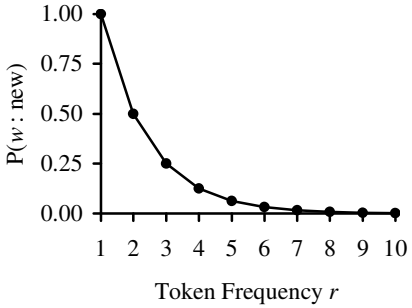


Figure 1: Probability of word *w* being identified as new.

We compare the outcome of RD and RW as follows. V_N values obtained under RD are listed in Table 2. For each of the 100 simulation runs generating these values, we sum the two corpus segments to obtain the whole corpus, and then, based on token frequencies in this whole, calculate a V_N value expected under RW, $E(V_N)$, based on the following formula:

$$(2) \quad E(V_N) = \sum_{r=1} N_r(0.5^r)$$

Here, *r* is the token frequency of a word, and N_r is the number of word types that occur *r* times. Table 4 contrasts V_N under RD and $E(V_N)$ under RW.

Table 4: Mean values for V_N (RD) and $E(V_N)$ (RW).

	V_N	$E(V_N)$
-ish	90.6	85.9
-ness	431.2	401.4
-ee	26.1	18.3
-ize	114.5	101.4
-ity	234.4	186.5
-er	558.6	463.9
-ly	754.3	707.0
-ify	21.1	18.5
-ous	107.1	89.1
-ion	348.7	270.6
-ment	61.6	49.4
ch#	29.7	23.0
-th	3.5	2.8

We find that each value of $E(V_N)$ is consistently an underestimation of the V_N. That is, more new words are captured by RD than by RW.

What kind of words are responsible for the discrepancy between RD and RW that is exhibited in Table 4? Consider *causee* (undoubtedly new to the majority of English speakers, except perhaps those who are syntacticians), which occurs 9 times in 1 document of the written component of the BNC. Under RW, the probability that all 9 tokens of *causee* will be distributed into only a single corpus segment is as low as about 0.002—in effect, *causee* is virtually guaranteed to be identified as non-new under RW.

Under RD, on the other hand, all 9 tokens of *causee* (occurring in just 1 document) will inevitably be distributed into only one corpus segment, and *causee* will thus be identified as new. The advantage of the P_{DE} measure (when implemented with RD) is that it captures as new words those words such as *causee* that are repeatedly used by the same speaker.

Implications for Linguistic Intuitions

Intuition-Based Interpretation

Although intuitions about productivity may not be reliable, they play an informal yet important role in evaluating results for a productivity measure—such results are often said to be intuitive or counter-intuitive. However, to the extent that little is known about the nature of such intuitions, determining the validity of a productivity measure on this basis may not be viable. Nevertheless, we may still ask what kind of information could be available to speakers (linguists) when they offer intuitive judgments about productivity.

We found in Table 2 that P_{DE} indices for *-er* and *-ly* are unexpectedly low, but one possibility is that the data on which P_{DE} is built are simply not in the form to be accessible to intuition. Speakers presumably cannot tell with any precision, for example, how many *-er* and *-ee* word types exist in the BNC, and thus, exact values of V, V_N, and P_{DE} may have little relevance to speakers' intuitions. On the other hand, speakers may be able to predict that "more" *-er* words than *-ee* words will occur.

What will be attempted here is a transformation of the data underlying the P_{DE} measure into a form that could be relevant to speakers' intuitions. There are two points to consider. The first is the possibility that whatever type frequency information speakers may have access to may be better represented on a logarithmic scale. Word frequency effects have been well studied in psycholinguistics (since Howes & Solomon, 1951; see Monsell, 1991, for an overview), where it has been noted that reaction time for a word in a lexical decision task is inversely proportional to the log frequency of that word. Although word frequency effects are normally discussed with respect to token frequency, the possibility that we entertain is that a similar logarithmic scaling may be also applicable to type frequency information.

The second point to consider is Baayen's (1993: 204) view that speakers' intuitive judgments on productivity are ordinal rather than interval in nature. Speakers presumably cannot tell to what extent *-ness* is more productive than *-ity*, but they may reject a productivity ranking in which *-ness* is ranked lower than *-ity*. Baayen also suggests that intuitions about productivity may simply be unavailable for some affixes.

Transforming V and V_N into $\log_{10}V$ and $\log_{10}V_N$ and taking their ratio (by analogy to V_N/V) is too simplistic a solution, and suffers a problem in that the complementary relationship between V_N and V_{NN} will be broken: the order of suffixes defined by V_N/V should be the reverse of the order defined by V_{NN}/V, but that relationship will no longer hold when values are log-transformed values. A solution to this dilemma is to shift our point of view, and to think of $\log_{10}V_N$ as the *extent* to which words are new and of $\log_{10}V_{NN}$ as the *extent* to which words are non-new. These are two conflicting factors that may simultaneously affect speakers' "impression" about a given word formation process. When $\log_{10}V_N$ (the extent to which words with a suffix are new) approaches $\log_{10}V_{NN}$ (the extent to which words with that suffix are non-new), the word formation process for that suffix may be felt to be productive, with a degree that can be calculated by the ratio of $\log_{10}V_N$ to $\log_{10}V_{NN}$.[9] Table 5 presents a productivity ranking of suffixes calculated in just this way.

Table 5: Intuition-oriented productivity ranking of suffixes.

	$\log_{10}V_N$	$\log_{10}V_{NN}$	Ratio
-ness	2.63	2.97	0.886
-ish	1.96	2.23	0.879
-er	2.75	3.29	0.836
-ly	2.88	3.45	0.835
-ize	2.06	2.51	0.821
-ity	2.37	2.89	0.820
-ee	1.42	1.80	0.789
-ion	2.54	3.26	0.779
-ous	2.03	2.73	0.744
-ment	1.79	2.56	0.699
-ify	1.32	1.93	0.684
ch#	1.47	2.26	0.650
-th	0.54	1.57	0.344

Following the view that intuitive judgments on productivity have an ordinal character, we concentrate only on the ranking of suffixes shown in Table 5. Interestingly, we seem to have gained many improvements as compared to Table 2. We particularly note the following: (a) *-ness* now counts as the most productive suffix; (b) *-er* and *-ly* move up in the ranking to be close to *-ness* and *-ish*; (c) *-ee* moves down in the ranking but is still close to *-ize*; and (d) *-ify* is now much lower in the ranking. Perhaps one unsatisfactory result is that *-ly* still does not emerge as the most productive suffix.

Although the exploration offered in this final section is based on speculation about what information could be available to speakers, the fact that the productivity ranking of suffixes in Table 5 is intuitively satisfying, by and large, suggests that the approach merits further investigation in future research.

Conclusion

The analysis of the data for the P_{DE} measure demonstrates that the deleted estimation method offers an effective means of capturing new words in corpus data and of assessing the

[9] The complementary relationship between V_N and V_{NN} is of course maintained by the ratio of $\log_{10}V_{NN}$ to $\log_{10}V_N$.

productivity of affixes. An interesting characteristic of the P_{DE} measure is that its identification of new words is not dependent on token frequency, and this may be construed as an advantage, given potential burstiness in the use of new coinages. The current measure identifies a word as new regardless of whether it is used repeatedly by the same speaker. The measure is also shown to be stable over different sample sizes.

Some findings appeared to deviate slightly from our intuitive expectations, but we proposed, (appealing to a psycholinguistic interpretation of the data), that it may be necessary to draw a distinction between raw corpus statistics and information that could be accessible to intuitions about productivity. Corpus statistics, scaled in psychologically plausible ways, may offer insights into the kind of information available to speakers when they make intuitive judgments on productivity.

A description of productivity obtained with a corpus-based productivity measure will be useful in many forms of linguistic research, not necessarily limited to the study of word formation. The success of the present study provides another indication that the corpus-based approach to the study of productivity advocated by Baayen (1992, 2001) is worthy of many future extensions.

Acknowledgments

The author wishes to thank Dianne Bradley, Harald Baayen, and the anonymous reviewers for their insightful comments on this paper. Any errors remain the responsibility of the author.

References

Aronoff, M. (1976). *Word Formation in Generative Grammar.* Cambridge, MA: MIT Press.

Baayen, R. H. (1992). Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991* (pp. 109–149). Dordrecht: Kluwer.

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1992* (pp. 181–208). Dordrecht: Kluwer.

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer.

Baayen, R. H., & Lieber, R. (1991). Productivity and English word-formation: A corpus-based study. *Linguistics, 29,* 801–843.

Baayen, R. H., & Renouf, A. (1996). Chronicling the Times: productive lexical innovations in an English newspaper. *Language, 72,* 69–96.

Bauer, L. (2001). *Morphological Productivity.* Cambridge, UK: Cambridge University Press.

British National Corpus (World Edition) [CD-ROM]. (2000). Oxford, UK: Oxford University Computing Services.

Church, K. W., & Gale, W. A. (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language, 5,* 19–54.

Evert, S., & Lüdeling, A. (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? *Proceedings of the Corpus Linguistics 2001 Conference* (pp. 167–175). Lancaster, UK.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40,* 237–264.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41,* 401–410.

Jelinek, F., & Mercer, R. (1985). Probability distribution estimation for sparse data. *IBM Technical Disclosure Bulletin, 28,* 2591–2594.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

Marchand, H. (1969). *Categories and Types of Present-Day English Word-Formation.* München: Beck.

Monsell, S. (1991). The nature and locus of word frequency effects in reading. In D. Besner & G. W. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition* (pp. 148–197). Hillsdale, NJ: Lawrence Erlbaum Associates.

Webster's Third New International Dictionary, Unabridged (Version 2.5) [CD-ROM]. (2000). Springfield, MA: Merriam-Webster.

Nishimoto, E. (2003). Measuring and comparing the productivity of Mandarin Chinese suffixes. *Journal of Computational Linguistics and Chinese Language Processing, 8* (1), 49–76.

Nishimoto, E. (2004). *A Corpus-Based Delimitation of New Words: Cross-Segment Comparison and Morphological Productivity.* Doctoral dissertation, City University of New York.

Oxford English Dictionary (2nd ed., Version 3.0) [CD-ROM]. (2002). Oxford, UK: Oxford University Press.

Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation.* Berlin: Mouton de Gruyter.

Testing Three Theories of Knowledge Transfer

Timothy J. Nokes (tnokes@uic.edu)

Department of Psychology (M/C 285)
University of Illinois at Chicago
Chicago, IL 60607-7513 U.S.A.

Abstract

Three theories of knowledge transfer -- analogy, knowledge compilation, and constraint violation -- were tested across three transfer scenarios. Each theory was shown to predict human performance in distinct and identifiable ways on a variety of transfer tasks. Results support the hypothesis that there are *multiple mechanisms* of transfer and that a general theory of transfer must incorporate each mechanism in principled ways.

Introduction

In order to understand human thinking and problem solving in complex and novel situations we need to have a general theory for how people use and adapt their prior knowledge to solve new problems. Aspirations towards such a goal have traditionally been discussed in terms of *transfer*, or how knowledge acquired from one task or situation can be applied to a different situation (Bransford & Schwartz, 1999; Detterman & Sternberg, 1993; Salomon & Perkins, 1989).

Work in cognitive science over the past thirty years has progressed towards this goal by investigating separate strands of transfer phenomena that occur in particular learning and problem solving situations. Although this research strategy has proven successful in developing local, independent explanations of knowledge transfer for particular experimental scenarios (e.g., analogical transfer and transfer appropriate processing), it has done little to bring us closer to a *general theory of transfer*. It is time to begin to weave these separate strands of investigation into a more complete theory that incorporates each strand in principled ways.

It is this charge of theoretical synthesis that motivates the two hypotheses under investigation in the current study. First, it is proposed that there is no single knowledge transfer mechanism, but multiple ones. These mechanisms include (but are not limited to) analogy, knowledge compilation, and error correction. Second, the particular transfer mechanism used depends on both (a) the knowledge actually present and how it is represented, and (b) the processing demands of the transfer task.

Below I summarize some of the prior work on transfer that is relevant to the investigation of these two hypotheses.

Mechanisms of Knowledge Transfer

The first mechanism of interest is *analogical transfer* (Gentner, Holyoak, & Kokinov, 2001). Analogical transfer is composed of three subprocesses: retrieving a prior knowledge structure, creating a mapping between it and the current problem or situation, and then using that mapping to generate new knowledge structures relevant to the application context. The transferred knowledge is typically assumed to be a declarative representation, but it can also include procedural attachments (Chen, 2002).

The empirical evidence for analogical mapping is extensive (Catrambone & Holyoak, 1989; Gentner & Toupin, 1986). However, the evidence also shows that although people are capable of mapping deep relational structures, the retrieval of an analogue is heavily dependent upon matches between the surface features of the current problem and prior problem solving experiences (Catrambone, 2002; Ross & Kilbane, 1997). Therefore, analogy is perhaps a better explanation for near transfer than for far transfer.

The second transfer mechanism of interest is *knowledge compilation* proposed by John R. Anderson and co-workers (Anderson, 1983; Neves & Anderson, 1981). Knowledge compilation was specifically proposed to explain how declarative knowledge is brought to bear on problem solving in the context of the ACT-R theory. This computational mechanism operates through the deliberate and explicit, step-by-step interpretation of a declarative statement that generates new production rules as a side effect. Those rules are then optimized via rule composition and the result is a procedural representation of the content of the declarative knowledge given a specific goal.

The knowledge compilation mechanism can be viewed as a translation device that translates or interprets declarative knowledge (e.g., advice, instructions, and strategies) into a set of procedures and actions that can be used to solve problems. Since knowledge compilation operates on declarative knowledge it can be used in a wide variety of application contexts because the knowledge has yet to be proceduralized, or tied to the goals of a particular problem solving context. This mechanism embodies a tradeoff between applicability and efficiency in that it has wide applicability across many contexts but requires a complicated and lengthy application process to translate the declarative knowledge into a set of actions. There is some empirical support for knowledge compilation but the evidence is not extensive (Anderson, Greeno, Kline, & Neves, 1981; Neves & Anderson, 1981).

The third transfer mechanism of interest is Ohlsson's (1996) *error correction* mechanism. Ohlsson and co-workers (Ohlsson, 1996; Ohlsson & Rees, 1991) have proposed that

the role of declarative knowledge is primarily to help a learner identify and correct his or her own errors. The constraint violation theory has both declarative and procedural components that operate in parallel, and the function of declarative knowledge is to constrain possible problem solutions. When incomplete or faulty procedural knowledge generates undesirable outcomes, these are recognized as violations of those constraints and the responsible rules are revised accordingly.

The power of declarative knowledge is that it can help the learner pinpoint the cause of an error, and transfer is the process by which errors are identified and remedied. This mechanism has wide applicability in that the constraints can be applied to a variety of problems that may require *different* strategies or sequences of actions to produce the correct solution. The constraint violation theory has been shown to generate power law learning curves (Ohlsson, 1996) and to support the design of successful tutoring systems (Mitrovic & Ohlsson, 1999).

In addition to each transfer mechanism using different cognitive processes, each mechanism has also been hypothesized to operate on specific types of prior *knowledge structures*. Analogy uses exemplar knowledge that consists of a declarative representation that may also have procedural attachments (Gentner, 1983). Knowledge compilation uses declarative knowledge such as instructions, advice, or tactical knowledge (Anderson, 1983). Error correction uses declarative knowledge of the constraints for a particular problem domain (Ohlsson, 1996).

In summary, researchers have proposed multiple alternative transfer processes including analogy, knowledge compilation, and error correction. Each mechanism has been associated with a particular kind of transfer scenario that specifies the conditions necessary for transfer (i.e., type of prior knowledge and application context). The purpose of the current study is to test the predictions of each transfer theory, and ask whether we can predict what transfer mechanism will be triggered for a given set of transfer scenarios.

The Present Study

In order to test these theories I implemented a between-groups training study in which subjects were given one of three training scenarios (exemplar, tactics, or constraints) and then were tested on a common set of problem solving tasks.

Each training scenario was designed to facilitate the construction of one of the three of the aforementioned knowledge structures associated with each transfer mechanism (i.e., exemplars for analogy, tactics for knowledge compilation, and constraints for error correction). In the exemplar training condition participants solve problems similar to those used in the transfer phase. In the tactical training condition participants learn instructional tactics for solving the transfer problems. In the constraints training condition participants learn the constraints associated with the problem solving task domain.

The transfer task is Thurstone's *letter extrapolation task* (Thurstone & Thurstone, 1941). In this task subjects are given a sequence of letters containing a pattern and their task is to find the pattern and continue it. Here is a simple example, A B M C D M . . . the correct continuation is E F M G H M. An important aspect of these problems for the current purposes is that prior declarative and procedural knowledge can make them easier to solve.

Although letter extrapolation is an invented task, it has several elements in common with many real world tasks including: a prior knowledge base (e.g., the alphabet), conceptual content (e.g., the pattern), materials to study (e.g., tactics), and generativity (e.g., one has to generate a sequence of coordinated actions).

Three different extrapolation problems were used in the transfer phase. Each problem was constructed with different properties or affordances, to elicit quantitative (accuracy, solution time, self-corrected errors) and qualitative (solution type) differences in performance from each training group.

The first transfer problem was designed to have a similar surface and deep pattern structure as that used in the exemplar training problems. This problem can also be solved by applying either tactical or constraint knowledge. The second transfer problem is open-ended and depending on how the given sequence is interpreted, different solution types are expected. This problem shares the same deep structure as the exemplar problems. However, the surface similar characteristics are misaligned and suggest a different interpretation. If the given sequence is interpreted as similar to the *surface sequence* one solution is expected. If it is interpreted as a *deep analogy* a second solution is expected. Tactical knowledge can also be used to solve this problem and biases one towards the second solution. Constraint knowledge can be applied as well and does not provide an a-priori bias towards any one of the correct solutions. The third transfer problem has neither surface nor deep structure similarity to the exemplar problems. The tactics are also not directly applicable. However, the constraints can be applied to find a unique solution.

In addition to comparing task performance across training groups, each training group was compared to a *no-training* control group for a measure of transfer relative to baseline performance.

Predictions

Exemplar Training. If participants in this training condition use exemplar knowledge and analogy to solve the first transfer problem they are expected to show high accuracy and fast solution times with few error-correcting behaviors as compared to the no-training group. They should show fast solution times for this problem because there is both surface and deep similarity to the training exemplars (i.e., fast memory access). They should show few error-correcting behaviors because they can transfer both declarative and procedural knowledge from the exemplars. For transfer problem 2 participants are expected to show high accuracy with slower solution times and few self-corrected errors. In

addition, they should show a bias for the surface similar problem solution. For transfer problem 3 they should show similar performance to no-training participants.

Tactical Training. If participants in this training condition use tactical knowledge and knowledge compilation to solve the first two transfer problems they should show high accuracy but similar solution times and error-correcting behaviors to that of the no-training group. In addition, for transfer problem 2 they should show a bias for the tactics relevant solution. For transfer problem 3 they should show similar performance to that of the no-training participants.

Constraints Training. If participants use constraint knowledge and error correction to solve all three transfer problems they should show high accuracy, similar solution times, and many error-correcting behaviors compared to the no-training group. In addition, they should show more variability in solution types for transfer problem 2.

Methods

Participants

One hundred and twenty-five undergraduate students from the University of Illinois at Chicago's subject pool participated in return for partial course credit.

Materials

Training Materials. The training materials for the *exemplar group* consisted of four sequence extrapolation problem isomorphs. Each problem was presented on a separate sheet of paper. All four training problems had the same deep pattern structure as each other and the first two transfer problems, but each was instantiated with different surface features. Below are two examples:

Exemplar 1: L M Z M L Y M N X . . .
Exemplar 2: E F S F E R F G Q . . .

The training materials for the *tactics group* consisted of a general tutorial, a tactic summary sheet, and several blank recall sheets. The tutorial (10 pages) provided instruction on specific kinds of pattern relations including: *forward*, *mirror-flip*, *backward*, *repeat*, and *identity*. Each pattern relation was defined and multiple examples were given. The tactics summary sheet consisted of one pattern continuing tactic and four pattern finding tactics including: (1) look for mirror flips or periods to break apart the pattern, (2) repeated letters may signal a mirror-flip order of symbols, group repeat, or period marker, (3) letters that are far apart in the alphabet may signal a mirror-flip alphabet, (4) letters close together may signal backward or forward relations. The tactics could be used to solve the first two transfer problems.

The training materials for the *constraints group* consisted of a constraints tutorial, constraint summary sheet, blank recall sheets, and letter string violation worksheet. The tutorial (5 pages) provided instruction on four letter pattern constraints: (1) all *completed* letter strings must be divisible into six groups of letters, (2) the number of letters in each similar group must be the same, (3) each letter group must be derived from either the immediately preceding letter group or the letter group two back, (4) letter operations must be repeated. The string violation worksheet provided a series of completed letter strings in which the participants' task was to identify constraint violations.

Test Materials. The test tasks were three letter extrapolation problems. See Table 1 for each transfer problem and its solution(s). The first extrapolation problem had a periodicity of three letters. It was superficially similar to the exemplar training problems and shared the same deep pattern structure. This problem could also be solved by applying either tactics or constraint knowledge. Subjects were asked to continue the solution to six positions.

The second extrapolation problem also had a periodicity of three letters. However, the correct continuation was ambiguous and was dependent on how the subject interpreted or "parsed" the given sequence. There are two primary solutions depending on the interpretation of the given sequence. If the letters are parsed into cross period relations of *forward-1* and *backward-1* comparable to surface similar relations used in the exemplar problems, one solution type will be derived (see Table 1, solution 1). However, if the given string is instead parsed as cross period relations of *mirror-flip-alphabet* and *backward-1* relations as suggested by a deep analogy or pattern finding tactic 3, a different solution will be derived. Subjects were asked to continue the solution to nine positions. In addition, there were four other possible correct solutions.

The third problem had a periodicity of two letters and had neither surface nor deep structure similarity to the exemplar problems. In addition, there was no pattern finding tactic that directly applied to this problem. However, a unique solution could be derived by constraint application. The pattern consists of pairs of letters incrementally increasing through the alphabet, each pair skipping an additional letter as the pattern progresses.

Table 1. Transfer problems and their solutions.

Problem Type	Given letter sequence & the correct extrapolation
Transfer 1	
	Given: R S F S R E S T D T S C . . .
	Solution → T U B U T A
Transfer 2	
	Given: B C P X Y O C D N . . .
	Solution 1 → Y Z M D E L Z A K
	Solution 2 → W X M D E L V W K
Transfer 3	
	Given: B A C B E D H G . . .
	Solution → L K Q P

Transfer problems were presented on a Macintosh computer with a 17'' color monitor, standard keyboard and mouse. Problems were presented in black 30 pt font in the center of the screen. The transfer portion of the experiment was designed and presented using PsyScope software.

Design

A between-subjects design was used with subjects randomly assigned to one of four training conditions: *exemplar training* (n = 31), *tactic training* (n = 31), *constraint training* (n = 33), and *no-training* (n = 30). Participants were tested individually. The procedure consisted of a training phase and a transfer phase.

Procedure

Training procedure for the exemplar group. Participants were first given general instructions for solving extrapolation problems. Next they were given three minutes to solve the first training problem. After three minutes participants received feedback on each position of their solution. If they extrapolated any position of the solution incorrectly they were given another instance of the same problem and three minutes to solve it. This cycle continued until the problem was solved correctly or the participant made four attempts to solve that problem. After the first problem this same procedure was continued for the remaining three training problems.

Training procedure for the tactics group. Participants first read the general tutorial. Next they memorized a summary sheet of the tactics for three minutes. Then they were given a simple unrelated distractor task to solve (e.g., three arithmetic problems). Participants were then asked to recall and write down all of the tactics. The experimenter assessed memory performance for recall of each tactic. If the subject omitted or incorrectly recalled any of the tactics they were given the tactic summary sheet to study again for another two minutes. After the second memorization phase they were given another distractor task followed by recall. This cycle was continued until the subject recalled all five tactics. After correct recall the subjects were asked to explain each tactic to the experimenter. If the subject gave an incorrect explanation the experimenter provided the correct explanation.

Training procedure for the constraints group. Participants first read the constraints tutorial. Next they memorized a summary sheet of the four constraints for three minutes. They were then given an unrelated distractor task to solve. Participants were then asked to recall the constraints and were given feedback on their recall performance. If they omitted or incorrectly recalled any of the constraints they were given the constraint summary sheet to memorize for another two minutes. After the second memorization phase they were given another distractor task followed by a blank recall sheet. This cycle continued until participants recalled all four constraints. After correct recall of the constraints subjects were asked to explain each constraint to the experimenter. If the subject gave an incorrect explanation the experimenter provided the correct explanation. Participants were then given the string violation worksheet.

Training procedure for the no-training group. Participants in this condition did not receive any training and served as a comparison condition of baseline performance on the transfer tasks.

Test procedure for all training groups. Subjects were seated at the computer and were told that they were to solve three extrapolation test problems. They were instructed that the given string of each transfer problem would be presented on the left side of the computer screen and that there would be an empty box for each letter position they were to extrapolate and fill in. Subjects were informed that they could re-enter new letters in any given position as many times as they would like. Subjects were told to click the mouse on the "Finished" field after all solution positions were filled and they were finished solving the problem. After the initial instructions participants were presented with each problem one at a time and given six minutes to solve each one.

Results and Discussion

Training Performance

Subjects in all three training groups were trained to criterion. The criterion measure for the exemplar group was solving at least two of the training problems completely correct. The criterion measure for the tactical and constraints groups was complete recall and correct explanation of the tactics and constraints respectively. Three subjects in the constraints training condition and one subject in both the exemplar and tactical training conditions did not pass the criterion. These subjects were excluded from further analysis leaving thirty subjects (n = 30) in each training group.

The training criterion provides evidence that each subject learned the target knowledge during the training phase (i.e., subjects in the exemplar group could solve training problems and subjects in the tactical and constraints group could recall declarative knowledge from memory). Next, I examine whether these subjects could transfer this knowledge to the problem solving tasks.

Transfer Performance

The three measures of central interest for the transfer phase were participants' accuracy scores and behavioral profiles across the three transfer problems, as well as the type of solution used to solve problem 2.

Accuracy Performance. To assess overall transfer performance participants' accuracy scores were examined for each training group. The *accuracy score* was the proportion of solution positions correctly extrapolated for a given transfer problem. The mean accuracy scores and standard deviations for each training group on the transfer problems are presented in Table 2.

Table 2. Mean proportion of solution positions correctly extrapolated for each transfer problem.

Training	Transfer1	Transfer2	Transfer3
Exemplar	.93* (.19)	.80 (.26)	.21 (.43)
Tactics	.78* (.32)	.71 (.32)	.22 (.37)
Constraints	.70* (.34)	.74 (.33)	.23 (.41)
No-training	.40 (.36)	.69 (.38)	.29 (.43)

A 4 (training) X 3 (problem type) mixed-analysis of variance (ANOVA) revealed a significant interaction of training by problem type, F (6, 232) = 6.46, $p < .05$. Follow-up comparisons showed that the interaction was best explained by the large advantage of the training groups over the no-training group on transfer problem 1, F (6, 232) = 43.14, $p < .05$, but not on problems 2 and 3, F (6, 232) = .76, *ns* and F (6, 232) = 1.41, *ns* respectively.

As predicted, all three training groups showed high accuracy in solving the first two transfer problems. Problem 1 in particular shows that the knowledge generated from each training condition facilitated transfer resulting in significantly higher accuracy performance than the no-training group. Although the constraints training group showed high accuracy on the first two transfer problems they did not show high accuracy scores on the final problem. One potential explanation for this lack of predicted transfer is that solving the first two transfer problems provided participants with partial exemplar knowledge that interfered with constraint application (this issue is further discussed in the conclusion).

Behavioral Profile. In order to assess whether a given participant used a particular transfer mechanism an ideal behavioral performance profile was created for each transfer mechanism. The use of a particular transfer mechanism can be evidenced by a constellation of scores across a set of dependent variables, what I term the *behavioral signature*.

The dependent variables used in this assessment included the accuracy score, the solution time, the number of self-corrected errors, and the checking time. The *solution time* was the total time in seconds to solve the problem. The *self-corrected error score* was the total number of times a subject re-entered a new letter into a given solution position that changed a previous response. The *checking time* was the amount of time in seconds between a subject's last extrapolation response and clicking on the finished button. This was presumably an indirect measure of error-checking behavior.

Using this set of dependent measures an ideal behavioral signature was created for each transfer mechanism (see Table 3). The qualitative indices (e.g., fast vs. slow) for a given variable are in comparison to the average no-training baseline performance.

Table 3. Ideal behavioral signatures for each transfer mechanism.

Behavior	**Transfer Mechanism**		
	Analogy	Knowledge Compilation	Error-Correction
High Accuracy	√	√	√
Fast Solution	√		
Error Checking			√

The ideal behavioral signature for analogical transfer was a high overall accuracy, a fast solution time on problem 1, and few error correcting behaviors. The ideal behavioral signature for knowledge compilation was a high overall accuracy, similar solution times and error correcting behaviors. The behavioral signature for error-correction was high accuracy, similar solution times, and a high number of error correcting behaviors.

Each subject's performance was examined as to whether it fit with a particular behavioral signature. For a subject's accuracy performance to be classified as *high* he or she had to have an overall accuracy score higher than the average (collapsed across problem) of the no-training group. For a subject's solution time to be classified as *fast* it had to be at least 1 standard deviation faster than the average solution time of the no-training group. Subjects were classified as having high error checking behavior if their performance met one of two criteria. The participant must have either scored 1 standard deviation above the average no-training group on both of the error measures (i.e., many self-corrected errors and long checking time), or have scored 2 standard deviations above on a single error measure. The number of subjects classified under each behavioral signature is shown in Table 4.

Table 4. Number of subjects classified under each behavioral signature.

Training Condition	**Behavioral Signature**			
	Analogy	Knowledge Compilation	Error-Correction	Other
Exemplar	19*	9	0	2
Tactics	2	16*	6	6
Constraints	3	7	13*	7

Chi-square tests showed that the training groups differed in the number of subjects classified for a given behavioral signature, χ^2 (6, N = 90) = 43.10, $p < .05$. Follow-up tests showed that more subjects trained on exemplars used analogy than those trained on tactics or constraints, χ^2 (2, N = 30) = 31.02, $p < .05$, more subjects trained on tactics used knowledge compilation than those trained on exemplars or constraints, χ^2 (2, N = 30) = 6.49, $p < .05$, and more subjects trained constraints used error-correction than those trained on exemplars or tactics, χ^2 (2, N = 30) = 16.94, $p < .05$.

In summary, the majority of subjects in a particular transfer condition showed the expected pattern of behavioral results as predicted by the three theories of transfer. This provides evidence that these three mechanisms are triggered under particular learning and transfer task conditions.

Solution Type. In addition to accuracy performance and behavioral profiles, further support for transfer can be assessed via the types of solutions participants used on problem 2. The number of subjects to use a given solution type is provided in Table 5.

Table 5. The number of subjects from each training group to use a given solution type.

Training Condition	Correct Solution Type		
	Solution 1	Solution 2	Others
Exemplar	19*	1	1
Tactics	5	12*	4
Constraints	12	0	4
No-training	9	0	9

Chi-square tests showed that the training groups significantly differed in the number of subjects to use a particular solution type, χ^2 (6, N = 77) = 41.68, $p < .05$. Of particular interest is that more exemplar training subjects used solution 1 than those given other forms of training, χ^2 (3, N = 77) = 20.80, $p < .05$, and that more tactics training subjects used solution 2 than subjects from the other groups, χ^2 (3, N = 77) = 29.70, $p < .05$.

In sum, these results provide further evidence that participants used training knowledge to solve the transfer problems. Subjects given exemplar training showed a preference for the surface similar solution and the tactics group showed a preference for the tactics relevant solution.

Conclusion

The results from this study provide support for the hypothesis that there are multiple mechanisms of transfer that are distinct and identifiable. Subjects in three separate transfer scenarios exhibited behavioral patterns of performance consistent with those predicted by three theories of knowledge transfer.

Several review articles have pointed out that the transfer literature exhibits a mixture of both positive and negative results (Bransford & Shwartz, 1999; Salomon & Perkins, 1989). While some studies have failed to find large transfer effects where we intuitively expect them, others have found transfer effects under particular types of study and test conditions. The complexity of the empirical results suggests that transfer is a heterogeneous phenomenon. Greater clarity might result if we assume that *different transfer processes are triggered in different types of transfer scenarios*. Results from the current study suggest that to understand transfer one must take a multifaceted approach and examine several interrelated aspects of the transfer scenario, not just one or two variables from a single theoretical perspective. Progress towards a general theory of transfer requires the synthesis and integration across current lines of research.

Future work should examine the interaction of these transfer mechanisms and investigate whether people are capable of *adaptively shifting between mechanisms* depending on their prior knowledge and the processing demands of the transfer task.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., Greeno, J. G., Kline, P. J., & Neves, D. M. (1981). Acquisition of problem-solving skill. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 191-230). Hillsdale, NJ: Erlbaum.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61-100.

Catrambone, R. (2002). The effects of surface and structural feature matches on the access of story analogs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 318-334.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1147-1156.

Chen, Z. (2002). Analogical problem solving: A hierarchical analysis of procedural similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 81-98.

Detterman, D. K., & Sternberg, R. J., (Eds.), (1993). *Transfer on trial: intelligence, cognition, and instruction*. Norwood, NJ: Ablex.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy, *Cognitive Science, 7*, 155-170.

Gentner, D., Holyoak, K. J., & Kokinov, B. N., (Eds.), (2001). *The analogical mind*. Cambridge, MA: MIT Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 277-300.

Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a data-base language. *International Journal of Artificial Intelligence and Education, 10*, 238-256.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 57-84). Hillsdale, NJ: Erlbaum.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103*, 241-262.

Ohlsson, S., & Rees, E. (1991) The function of conceptual understanding in the learning of arithmetic procedures. *Cognition & Instruction, 8*, 103-179.

Ross, B. H., & Kilbane, M. C. (1997). Effects of principle explanation and superficial similarity on analogical mapping in problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 427-440.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomena. *Educational Psychologist, 24*, 113-142.

Thurstone, L. & Thurstone, T. (1941). *Factorial studies in intelligence*. Chicago: University of Chicago Press.

Individual differences and implicit language: personality, parts-of-speech and pervasiveness

Jon Oberlander (J.Oberlander@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Alastair J. Gill (A.Gill@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

Dewaele and Furnham predict that in oral language Extraverts prefer to produce what they term implicit language. They use: more pronouns, adverbs and verbs; and fewer nouns, adjectives and prepositions. However, communication in a computer-mediated environment, such as e-mail, might disrupt these preferences. Also, other personality dimensions, such as Neuroticism, may be related to implicitness. The study exploited an existing corpus of e-mail texts written by native English speakers of known personality. Stratified corpus comparison used n-gram-based techniques from statistical natural language processing, to compare relative frequencies of use of (sequences of) parts-of-speech. Implicitness effects were found, and Neuroticism appeared to have a clearer impact than Extraversion.

Personality and language

Individuals differ in the way they speak and write. Some of those differences are systematic, and can be attributed to apparently deeper differences, such as personality traits, like Extraversion and Neuroticism. Extraversion is a trait strongly related to interpersonal interaction and sociability, whereas, Neuroticism, or Emotional Stability, is related to internal emotional states, rather than interaction. In the past, it has been found that both these personality traits do significantly influence an individual's language production behaviour in a variety of contexts (Pennebaker and King, 1999; Dewaele and Furnham, 1999). Recent work has investigated e-mail text, and suggested that there are characteristic sequences of words and punctuation associated with each end of both dimensions (Extravert or Neurotic) (Gill and Oberlander, 2002, 2003).

However, Mehl and Pennebaker (2003) note that linguistic style is more consistently described by its syntactic component, than by content. So, it could be that the relative use of different parts-of-speech (POSs) is a more important indicator of personality than the relative use of words or strings of words.

The work by Dewaele and Furnham suggests that, at least for Extraversion, there are real effects to be found in spoken language, at the level of POSs. In their account, implicit language involves a preference for pronouns, adverbs and verbs, whereas explicit language involves a preference for nouns, adjectives and prepositions. Heylighen and Dewaele (2002) suggest that Extraversion leads to implicitness due to greater visual-spacial capacities, and this is part of an overall preference for informal language. However, this work leaves open whether or not implicitness effects will be found for Neuroticism. Gill and Oberlander's work suggests that formality may also be a factor in Neurotic language behaviour, because the reduced resources of high Neurotics do not enable detailed language planning. But that work did not investigate implicitness in patterns of POS use. It would therefore be interesting to know whether Dewaele and Furnham's 'Implicit-Extravert hypothesis' applies in the genre of e-mail text—a genre close to spoken language—and if so, how.

To address this question, the rest of this paper is structured as follows. First, we give some background to help frame implicitness hypotheses that gives POS predictions for both Extraversion and Neuroticism. We then present the stratified corpus comparison methods used in analysing POS use in the e-mail corpus. Results were somewhat unexpected, in that implicitness predictions appear to be confirmed for Neuroticism, but not for Extraversion. We discuss possible ways of resolving the issue.

Background

Two personality traits

Extraversion and Neuroticism are traits which are common to the two major trait theories of personality: Eysenck's three factor model (Eysenck and Eysenck, 1991); and the five factor model developed by Costa and McCrae (Costa and McCrae, 1992) and others.

They are described as follows: High Extraverts are said to be sociable, easy-going, and optimistic, and to take chances. Low Extraverts (or Introverts) are said to be quiet, and reserved, and to plan ahead, and dislike excitement. High Neurotics are said to be: anxious, worrying, over-emotional, and frequently depressed. Low Neurotics are said to be: calm, even-tempered, controlled, and unworried (Eysenck and Eysenck, 1991).

Dewaele and Furnham

Furnham (1990) has proposed the following features of Extravert and Introvert language. Extravert language: is less formal; has a more restricted (rather than elaborated) code; uses vocabulary more loosely, where this is defined in terms of how correctly words are used, and how unusual they are. And it uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions). This last tendency directly involves POSs. Using factor analysis of syntactic tokens produced by L2 speakers, Dewaele and Furnham (2000) describe implicit language as a preference for pronouns, adverbs and verbs, and they contrast it with explicit language, seen as a preference for nouns, modifiers and prepositions. So Extraverts prefer implicitness, and Introverts prefer explicitness. For the purposes of this paper, we shall term this the Implicit-Extravert Hypothesis. The hypothesis appears to hold in both informal and formal situations, and is consistent with previous analyses of the individual linguistic categories (Dewaele, 2001). Cope (1969) also notes a lower lexical diversity (measured as type-token ratio), for Extravert native French speakers, with this also the case for non-native speakers of English (Dewaele and Furnham, 2000).

However, although they have discussed varieties of anxiety and their effects on communication, Dewaele and Furnham have not attempted to predict which part-of-speech patterns might be characteristic of the related trait Neuroticism. What might we expect to find?

An extension: Implicit-Neuroticism

Previous work by Gill and Oberlander (2002, 2003) gathered a corpus of e-mail messages, and analysed it for characteristic words and sequences of words. The corpus comprised 210 texts produced by 105 University students or recent graduates (37 males, 68 females). Each participant composed two e-mails *to a good friend whom they hadn't seen for quite some time*, spending around 10 minutes on each message. The first e-mail concerned their activities in the past week, the second discussed their plans for the next week. The total corpus size is around 65,000 words.

Following analysis of occurrences of individual words, and sequences of words, it was reported that the corpus results on Extravert words were broadly consistent with previous findings, for instance using informal language, looser punctuation, vaguer quantification and more co-ordination. This therefore appears to fit the Implicit-Extravert hypothesis; however, no POS analysis was reported.

However, there were also results on Neurotic language use. Pennebaker and King (1999) previously argued that High Neuroticism was associated with a language factor for 'Immediacy'. Gill and Oberlander (2003) extended these results, suggesting that 'High Neurotics show a preference for forms occurring frequently in speech, for example, *I, and, that*, rather than less common words such as *abject, suspicion, tether*. This preference for common words contributes towards the very low lexical density found in highly Neurotic texts, demonstrated by the high level of repetition over ten-word sections of text.'

What is interesting about this is that it suggests that Dewaele and Furnham's ideas about formality and implicitness might be as relevant to the Neuroticism dimension as they are to the Extraversion dimension. If they are, then we would expect that—like High Extraverts—High Neurotics will use more verbs, adverbs and pronouns, while Low Neurotics will use more nouns, adjectives, and prepositions. We call this the Implicit-Neurotic Hypothesis (INH). It obviously raises the question of whether or not *both* dimensions are related to implicitness, and the relative strength of any connections.

To address this question, we here apply to the existing e-mail corpus a series of techniques to derive POS frequencies, and POS sequences.

Syntactic Analysis of the Corpus

Method

The personality corpus was acquired as described above. It was tagged using the Penn part-of-speech tagset, using the MXPOST tagger (Ratnaparkhi, 1996). Further processing removed the original words, leaving their associated POS tags. A subsequent stage of processing reduced the POS tags from the detailed Penn tagset to more general syntactic categories. The 45 Penn tags (see Marcus, Santorini, and Marcinkiewicz, 1994, for more details) were converted to 10 broader categories, as implemented in the electronic version of the Shorter Oxford English Dictionary which is incorporated into the MRC Psycholinguistic Database (Wilson, 1987). These are: Noun (NN), Adjective (ADJ), Verb (VBN), Adverb (ADV), Preposition (PRP), Conjunction (CONJ), Pronoun (PRN), Interjection (INT), Past Participle (VPP), and Other [syntactic categories] (O). In addition to these categories, we also make use of ⟨p⟩ indicating punctuation, and 'NA', which indicates that a feature does not belong to any of the above categories and generally represents the ⟨END⟩, end of text marker. Note that here we use a different set of labels to enhance intelligibility, and these do not co-incide exactly with those used in the MRC database: for instance, we use 'PRP' instead of 'R'.

The reduced-tag corpus—with the more general syntactic categories—was then divided into stratified sub-corpora. In stratifying, we isolate a 'reference corpus' of text from authors with a personality profile which is not extreme on any of the measured dimensions. We can then compare authors from each of the extreme personality groups with this 'neutral' (here termed 'mid') group. Thus, High

and Low personality group samples were created by splitting them at greater than 1 standard deviation above and below the EPQ-R score for each dimension. The additional requirement was made that authors had to be *within* 1 standard deviation on the dimensions other than the one for which they were extremely high or low. Additionally, all texts which were within 1 standard deviation across *all* personality dimensions were assigned to the personality 'neutral' Mid sub-corpus. Thus, on any dimension, we have three groups to compare (High, Mid, and Low).

The resulting sizes of the subcorpora are as follows: Around 6,000 words for the high Extraversion, and over 2,000 words for the low Extraversion groups (11 and 4 authors respectively); Over 3,000 words for the high Neurotic and around 6,000 words for the low Neurotic groups (6 and 9 authors). The Neutral group was around 10,000 words (23 authors).

To identify collocations in the tagged sub-corpora, we calculate 1–5 word n-grams, and do not use a rank or frequency cut-off during calculation, but only present features with a frequency $\geq$5. This enables an accurate log-likelihood statistic (G^2) of their occurrence between groups to be calculated (cf. Rayson, 2003). We use N-gram software (Banerjee and Pedersen, 2003) to compute G^2 for 2- and 3-grams. To identify those robust collocations which distinguish one group from another, we need to make a three-way comparison of the linguistic features across the high-mid-low corpora for each group. We calculate the relationships between the three groups, and for each feature in each corpus we identify its frequency and relative frequency, and then where relevant its relative-frequency ratio and log-likelihood between High-Low, High-Mid and Low-Mid groups. This allows us to compare the relative usage and statistical significance of the difference in the use of features between groups.

Results

We first report the results of the unigram analysis for Extraversion and Neuroticism dimensions, we then report the findings of the overall n-gram analyses (1–5 item sequences). Following this, the results for Extraversion and Neuroticism are outlined.

Unigram Syntactic Analysis

Results of the unigram analysis for the reduced set of syntactic tags can be found in Tables 1 and 2. We display the results for all tags present in our data; however G^2 values which achieve significance of $p \leq 0.05$ or $p \leq 0.01$ are noted by $*$ or $**$ respectively.

In this presentation of the results, we draw attention to features which are characteristic of the High or Low groups, compared with the usage of the feature more generally. In the tables, we distinguish whether a feature is under- or over-used by one of the three groups (High, Mid or Low), relative to the two other groups; this information is given in the final three columns of each table, with over-use indicated by + and under-use by −. However, a more concise view of the results can be gained in the following way. At least two kind of features can be associated with (say) High Neuroticism: unigrams which are over-used by High Neurotics; and unigrams which are under-used by Low Neurotics. Thus, Figure 1 lists, for each dimension and each sub-group, the features which are associated with that group *either* via their over-use of the feature, *or* an opposite group's underuse.

High Extraverts [CONJ]
Mid Extraverts –
Low Extraverts [VPP]

High Neurotics [CONJ] [PRN]
Mid Neurotics –
Low Neurotics [ADJ] [NN]

Figure 1: Summary of unigram POS analysis

For Extraversion, conjunction (CONJ) is characteristic of High Extraverts, and past participle verbs (VPP) of Low Extraverts. The Mid Extravert group shows no significant under- or over-use of the general tags. For Neuroticism, conjunction (CONJ) and pronouns (PRN) are characteristic of High Neurotics, and adjectives (ADJ) and nouns (NN) of Low Neurotics. The Mid Neurotic group shows no significant under- or over-use of the general tags.

For these results, we note the generally modest levels of significant differences we found between personality groups. We may take this to indicate that these groups generally use relatively similar proportions of the relevant parts of speech. However, the POSs may also occur in different contexts or sequences, thus indicating differences in they way they are used. We therefore turn to the results of the n-gram analysis of the syntactic tag data.

N-gram Syntactic Analysis

There is insufficient space to display the full results. A concise view is therefore given in Figure 2. Notice that for the Mid groups, we have to distinguish features labelled specifically as under-use, since this is of course relative to both the High and Low groups.

The features here reach much higher levels of significance than the unigrams, so here we only discuss those which reach the critical value of 10.83 (i.e., $p \leq 0.001$). 32 n-gram features reach this value for Neuroticism, and 25 for Extraversion. Of these, the majority in each case reach the 15.13 critical value ($p \leq 0.0001$): 23 and 17, respectively. The features reaching this higher value are predominantly bigrams, exceptions being the longer n-grams for

Feature	Rank	High Freq.	High R.Freq	Mid Freq.	Mid R.Freq	Low Freq.	Low R.Freq	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
VPP	1	118	0.0173	202	0.0185	66	0.0260	0.34	5.43*	6.73**			+
CONJ	2	258	0.0378	338	0.0310	88	0.0347	5.80*	0.88	0.50	+		
ADV	3	562	0.0824	963	0.0882	238	0.0938	1.67	0.71	2.76			
PRP	4	679	0.0995	1100	0.1008	231	0.0910	0.06	2.02	1.40			
O	5	1071	0.1570	1714	0.1570	369	0.1454	0.00	1.82	1.64			
VBN	6	1156	0.1695	1804	0.1652	449	0.1769	0.44	1.65	0.60			
⟨p⟩	7	667	0.0978	1048	0.0960	228	0.0898	0.14	0.84	1.23			
ADJ	8	404	0.0592	617	0.0565	136	0.0536	0.53	0.32	1.03			
NA	9	23	0.0034	47	0.0043	9	0.0035	0.95	0.30	0.02			
PRN	10	696	0.1020	1118	0.1024	277	0.1091	0.01	0.89	0.89			
NN	11	1177	0.1725	1945	0.1782	442	0.1742	0.76	0.19	0.03			
INT	12	11	0.0016	21	0.0019	5	0.0020	0.23	0.00	0.13			

Table 1: Reduced syntactic tag unigram analysis, Extraversion.
Note. $*p < .05$, $**p < .01$, $df = 1$.

Feature	Rank	High Freq.	High R.Freq	Mid Freq.	Mid R.Freq	Low Freq.	Low R.Freq	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
ADJ	1	193	0.0501	617	0.0565	447	0.0660	2.15	6.15*	10.50**			+
CONJ	2	155	0.0403	338	0.0310	210	0.0310	7.09**	0.00	6.01*	+		
NN	3	625	0.1624	1945	0.1782	1230	0.1815	4.13*	0.27	5.22*	−		
PRN	4	424	0.1102	1118	0.1024	648	0.0956	1.62	1.93	5.06*	+		
INT	5	9	0.0023	21	0.0019	6	0.0009	0.23	3.19	3.48			
VPP	6	63	0.0164	202	0.0185	146	0.0215	0.74	1.95	3.44			
VBN	7	688	0.1787	1804	0.1652	1132	0.1671	3.04	0.09	1.94			
NA	8	13	0.0034	47	0.0043	19	0.0028	0.63	2.63	0.26			
PRP	9	352	0.0915	1100	0.1008	650	0.0959	2.55	0.99	0.53			
O	10	627	0.1629	1714	0.1570	1035	0.1528	0.62	0.48	1.60			
ADV	11	318	0.0826	963	0.0882	595	0.0878	1.04	0.01	0.78			
⟨p⟩	12	382	0.0992	1048	0.0960	657	0.0970	0.31	0.04	0.13			

Table 2: Reduced syntactic tag unigram analysis, Neuroticism.
Note. $*p < .05$, $**p < .01$, $df = 1$.

punctuation found for Neuroticism. In interpreting this data, we seek distinctive POS collocations. Table 3 shows, for each sub-group, how many distinctive collocations involving each POS were found.

Extraversion From the unigram analysis, we are particularly interested in collocations involving conjunctions (for the High E group) and past participle verbs (for the Low E group). As far as conjunctions are concerned, High Extraverts are associated with the use of [CONJ VBN] and [CONJ ADV], while Low Extraverts are associated with the use of [CONJ VBN PRN]. The latter offers a particularly distinctive collocation, since the pronoun switches the preference from High to Low E. Turning to past participles, we find that High E prefer [VPP PRP], but there are no preferred collocations for Low Extraverts.

Given Table 3, the remaining discrepancies between the High and Low E groups are as follows. Allowing that there are substantially more distinctive collocations for the High E group overall, we find that the High E group has notably more collocations involving: punctuation, adjectives, nouns, and POSs in the Other category. The Low E group has notably more collocations involving verbs and pronouns.

Neuroticism Here, we are most interested in collocations involving pronouns and conjunctions (for the High N group) and adjectives and nouns (for the Low N group). Taking pronouns first, we find a High Neurotic preference for [ADJ PRN VBN], [ADJ PRN] and [VBN PRN O]. Turning to conjunctions, they also show a preference for [VBN ADJ CONJ]. Three of these collocations also involve adjectives, which are used overall more by Low Neurotics. However, the rest of High N preferences for collocations involving pronouns instead involve adverbs: [VBN PRN O ADV VBN], [VBN PRN O ADV], [PRN VBN PRN O ADV] and [ADV PRN VBN PRN]. While Low Neurotics have only one pronoun collocation involving an adjective—[PRN ADJ]—the other three of their preferred pronoun or conjunction collocations also involve adverbs: [PRN ADV], [ADV PRN] and [CONJ ADV].

Given Table 3, and allowing that there are rather more distinctive collocations for the High Neurotic group overall, we find that the High Ns have notably more collocations involving verbs, and POSs in the Other category. The Low Ns have notably more collocations involving: past participle verbs and adverbs.

Discussion

Dewaele and Furnham's original Implicit-Extravert Hypothesis predicted that in spontaneous speech High Extraverts will use more verbs, adverbs and pronouns, and that Low Extraverts will use more nouns, adjectives, and prepositions (see Heylighen and Dewaele, 2002, for a discussion as to why certain POSs are preferred by Extraverts). The unigram analysis did not support these predictions. It indicated that High E use more conjunctions, and that Low E use more past participle verbs. No other overall differences were found, although it is perhaps

High Extraverts [CONJ VBN] [NN NN] [ADV ⟨p⟩] [PRN NN] [⟨p⟩ O] [ADV O] [ADJ ⟨p⟩] [NN ADV] [CONJ ADV] [VPP PRP] [ADJ O] [⟨p⟩ ADJ] [PRN O ADV] [VBN O NN ⟨p⟩] [PRN O ADV VBN] [⟨p⟩ O VBN ADJ ⟨p⟩] [⟨p⟩⟨p⟩⟨p⟩]

Mid Extraverts Underuse: [⟨p⟩ ADV] [⟨p⟩ NN]

Low Extraverts [ADV PRP] [PRN ADV] [VBN PRN O] [VBN PRN ADV] [CONJ VBN PRN] [VBN ⟨p⟩ PRN]

High Neurotics [VBN PRP] [⟨p⟩ O] [⟨p⟩⟨p⟩⟨p⟩⟨p⟩⟨p⟩] [⟨p⟩⟨p⟩⟨p⟩⟨p⟩] [⟨p⟩⟨p⟩] [⟨p⟩⟨p⟩⟨p⟩] [VBN PRN O] [ADJ PRN VBN] [PRP ADJ] [VBN O VBN ADV] [PRN VBN PRN O ADV] [VBN ADJ CONJ] [ADJ PRN] [VBN PRN O ADV VBN] [VBN PRN O ADV] [ADV PRN VBN PRN]

Mid Neurotics Underuse: [PRN ⟨p⟩ ADV] [NN VBN O ADJ] [NN VBN O ADJ NN] [PRN O VBN ⟨p⟩]

Low Neurotics [⟨p⟩ ADV] [PRN ADV] [ADV ADV] [ADJ ⟨p⟩] [ADV O] [VPP ADV] [O ADV] [ADV PRN] [CONJ ADV] [ADV VPP] [PRN ADJ] [VPP PRP]

Figure 2: Summary of n-gram POS analysis

worth noting that since we have both past participles and general verbs, our categories are slightly more fine-grained, which may affect the result.

The new Implicit-Neurotic Hypothesis predicted that High Neurotics will use more verbs, adverbs and pronouns, and that Low Neurotics will use more nouns, adjectives, and prepositions. The unigram analysis partially supported these predictions. It found that High N use more pronouns (and conjunctions), and that Low N use more nouns and adjectives. However, no overall differences were found for verbs, adverbs or prepositions.

At first glance, then, it appears that the Neuroticism dimension is more closely related to implicitness than the Extraversion dimension, in this corpus

POS	Extraversion			Neuroticism			Total
	High	Mid	Low	High	Mid	Low	
⟨p⟩	7	2	1	5	2	2	19
ADJ	4	0	0	4	2	2	12
ADV	6	1	3	5	1	9	25
CONJ	2	0	1	1	0	1	5
NN	4	1	0	0	2	0	7
PRN	3	0	5	7	2	3	20
PRP	1	0	1	2	0	1	5
VBN	4	0	4	9	3	0	20
VPP	1	0	0	0	0	3	4
O	7	0	1	6	3	2	19
NA	0	0	0	0	0	0	0
Total	39	4	16	39	15	23	136

Table 3: Distinctive collocations involving a given POS.

of e-mail text. Two potential explanations emerge to explain the difference between this and Dewaele and Furnham's results: Firstly, they were studying spoken, rather than written, language; and secondly, that they were largely dealing with L2 speakers. Perhaps implicitness is more closely related to Neuroticism in written language, and for Extraversion in spoken language; likewise it may have different effects for native and non-native language users. However, before following this line of reasoning, we should also consider the results of the n-gram analysis. At least two gross patterns are interesting.

First, where a High and Low group do not differ overall in the relative frequency of use of a POS, one group may have rather more types of distinctive collocation involving that POS than the other group. If overall use does not differ, it means that one group is using the POS in many different contexts; the other may be using it in a narrower, or perhaps more stereotypical, range of contexts. Let us call the greater-range case 'pervasive' use. Secondly, where a High and Low group do differ in relative frequency of use of a POS, it is interesting to note whether higher frequency is associated with a greater set of collocations involving that POS, or a smaller set. Intuitions here are not firm; but we might expect that greater relative frequency is associated with a greater range of use—and hence, with perhaps fewer stereotypical collocations. If so, frequency may track pervasiveness.

So, consider again the original Implicit-Extravert Hypothesis: High Extraverts will use more verbs, adverbs and pronouns, and Low Extraverts will use more nouns, adjectives, and prepositions. We find that High E prefer conjunctions overall, but that it is the Low E who tend towards POS-collocations involving verbs and pronouns. So High E use of verbs and pronouns may not be not greater overall, but it is pervasive. Equally, Low E prefer past participle verbs overall, but it is the High E who tend towards POS-collocations involving nouns, adjectives, punctuation, and the Other category. Perhaps Low E use of adjectives and nouns is pervasive. And since Low Extraverts actually use proportionately more VPP, their complete lack of distinctive robust collocations suggests that they use VPP pervasively.

Now, let us turn to the new Implicit-Neurotic Hypothesis. High Neurotics will use more verbs, adverbs and pronouns, and Low Neurotics will use more nouns, adjectives, and prepositions. We find that High N prefer pronouns and conjunctions overall, but that it is the Low N who tend towards POS-collocations involving past participle verbs and adverbs. So perhaps High N use of past participle verbs and adverbs is pervasive. Equally, Low N prefer adjectives and nouns overall, but it is the High N who tend towards POS-collocations involving verbs and the Other category. And again, perhaps Low N use of verbs and Other is pervasive.

This pattern is not quite so simple as the Extravert case, and this may in part be because we have split the verb category in two, distinguishing past participle verbs from verbs in general. Putting this to one side, however, we do find High N use of adverbs to be pervasive; and this at least fits the picture of pervasiveness that seemed to be emerging with Extraversion.

Conclusion

This paper set out to establish whether Dewaele and Furnham's Implicit-Extravert Hypothesis for oral language applies in the genre of written e-mail text produced by native English speakers.

At the simple unigram level, it appears that Neuroticism rather than Extraversion fits the implicitness predictions concerning frequency of use of parts-of-speech. However, we can drill down to the collocations level, and we may assume that the pervasive use of a POS tends to reduce the likelihood of finding stereotypical collocations involving it. If we do, then Extraversion does involve implicitness after all. On this interpretation, a POS can be characteristic of some personality group not because they use it more frequently than other groups; rather, it is characteristic because they use it more pervasively.

Applications of this work include affective text categorisation, and therefore could contribute towards the rapidly expanding field of sentiment classification. In taking this work further, we need to give the idea of pervasiveness a more solid basis. But this is only worth pursuing if the idea is really needed to explain the data. And we will only know this once we have tested the hypotheses against larger corpora in other domains. The corpora could be brand new; but it would certainly be possible to apply the analytic techniques presented here to other previously gathered personality corpora.

Acknowledgements

Our thanks to Jean-Marc Dewaele for his comments and suggestions about this paper. The second author gratefully acknowledges studentship support from the UK Economic and Social Research Council and the School of Informatics.

References

Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counselling Psychology*, **16**, 1–19.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Dewaele, J.-M. (2001). Interpreting the maxim of quantity: interindividual and situational variation in discourse styles of non-native speakers. In E. Nèmeth, editor, *Cognition in Language Use: Selected Papers from the 7th International Pragmatics Conference*, volume 1, pages 85–99. International Pragmatics Association, Antwerp.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**, 509–544.

Dewaele, J.-M. and Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, **28**, 355–365.

Eysenck, H. and Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. and Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; Neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461.

Heylighen, F. and Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science*, **7**, 293–340.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**, 313–330.

Mehl, M. and Pennebaker, J. (2003). The sounds of social life: A psychometric analysis of student's daily social interactions. *Journal of Personality and Social Psychology*, **84**, 857–870.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Wilson, M. (1987). MRC Psycholinguistic Database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford.

Identifying the Perceptual Dimensions of Visual Complexity of Scenes

Aude Oliva (oliva@mit.edu)
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139

Michael L. Mack (mackmic1@msu.edu)
Department of Computer Science, Michigan State University, East Lansing, MI 48824 USA

Mochan Shrestha
Department of Mathematics, Michigan State University, East Lansing, MI 48824 USA

Angela Peeper
Department of Psychology, Michigan State University, East Lansing, MI 48824 USA

Abstract

Scenes are composed of numerous objects, textures and colors which are arranged in a variety of spatial layouts. This presents the question of how visual complexity is represented by a cognitive system. In this paper, we aim to study the representation of visual complexity for real-world scene images. Is visual complexity a perceptual property simple enough so that it can be compressed along a unique perceptual dimension? Or is visual complexity better represented by a multi-dimensional space? Thirty-four participants performed a hierarchical grouping task in which they divided scenes into successive groups of decreasing complexity, describing the criteria they used at each stage. Half of the participants were told that complexity was related to the structure of the image whereas the instructions in the other half were unspecified. Results are consistent with a multi-dimensional representation of visual complexity (quantity of objects, clutter, openness, symmetry, organization, variety of colors) with task constraints modulating the shape of the complexity space (e.g. the weight of a specific dimension).

Introduction

Real-world scenes are composed of numerous objects, textures and colored regions, which are arranged in a variety of spatial layouts. Although natural images are visually complex, we are able to form a coherent percept amid numerous regions, and identify a complex scene at a glance (Potter, 1976), even in the face of visually degraded conditions (Schyns & Oliva, 1994). This presents the question of how a cognitive system may represent the level of complexity of a scene. Specifically, the following question motivated the experiment presented in this paper: can visual complexity be conceptualized along a single dimension? Or is visual complexity better represented as a multi-dimensional space where the axes might correspond to meaningful perceptual dimensions?

Visual complexity

The perception of visual complexity has been studied with natural texture images (e.g. Heaps & Handel, 1999; Rao & Lohse, 1993) and simple patterns (see Palmer, 1999 for a review). Heaps and Handel had participants rank texture images along several perceptual dimensions including complexity, connectedness, depth, orientation, repetitiveness, and structure. The authors defined complexity as "the degree of difficulty in providing a verbal description of an image". They observed that the complexity of a texture could be estimated along a one dimensional axis representing the degree of perceivable structure: textures with repetitive and uniform oriented patterns were judged less complex than disorganized patterns. This finding correlates with results in the domain of perceptual grouping by acknowledging that the presence of regularities (e.g., symmetry, repetition, similarity) simplifies a visual pattern (Feldman, 1997; Palmer, 1999; Van der Helm, 2000).

How can we represent the complexity of a stimulus like a scene, which has a high variability of parts and spatial layout organization? According to Heylighen (1997), the perception of complexity is correlated with the *variety* in the visual stimulus. Figure 1 illustrates two instances of *variety.* First, the perceived visual complexity can increase as a function of the quantity and range of objects. Second, the perceived visual complexity can increase as a function of the variety of materials and surface styles while the number of objects and surfaces remain constant. The representation of a real-world scene is likely to combine both levels of varieties (parts and surface styles). Intuitively, complex scenes should contain a larger variety of parts and surfaces styles, as well as more relationships between these regions than do simpler scenes.

A visual pattern is also seen complex if its parts are difficult to identify and separate from each other. Yet, paradoxically, when the parts are separated or conceptualized as a whole,

the valence of the complexity changes and the pattern becomes simpler (Heylighen, 1997). This suggests that the perceived complexity of an image also depends on the amount of perceptual grouping, a characteristic independent of the quantity of parts, an observer perceives in the scene. Additionally, the perception of visual complexity is likely to be dependent on the scale of observation (e.g. looking at a bookshelf or the books level), preexisting schemas and familiarity with the scene.

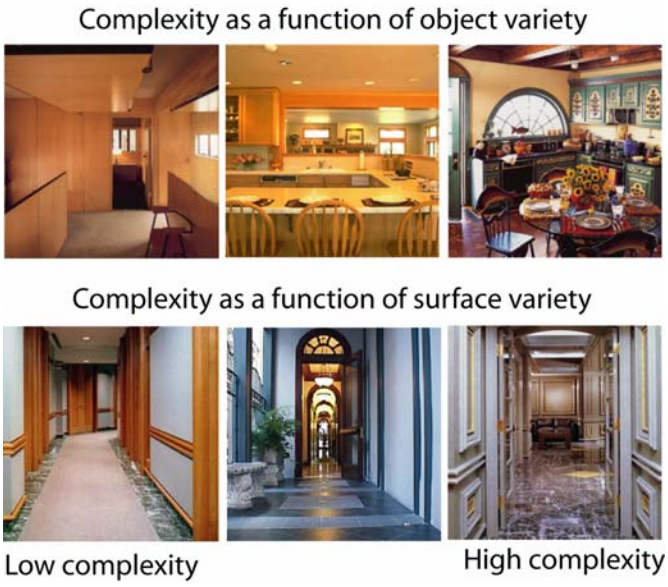


Figure 1: Illustration of how visual complexity evolves as a function of object variety (top) and surface variety (bottom).

If the perception of visual complexity is an interaction between the information in the image and task constraints, can we still identify a set of perceptual properties that participants consistently use to characterize visual complexity of real world scenes? The shape of the visual complexity representation could take three forms:
(1) Unique Perceptual Dimension: the properties of complexity are combined into one principal dimension, robust to subjectivity and task constraints. This is the case of the *naturalness* dimension in real world scenes (e.g. judging if a scene image is a natural or a man-made environment, Oliva & Torralba, 2001).
(2) Multi-dimensional Space Representation: most of visual complexity variability is explained by an identifiable number of perceptual dimensions. The weight of each dimension may vary with task constraints, but the principal dimensional vocabulary remains the same (Gardenfors, 2000). This seems to be the case of the representation of basic-level scene categories (e.g., beach, street, Oliva & Torralba, 2001).
(3) Flexible Space Representation: the properties that human observers use to represent the visual complexity of a particular scene vary with image characteristics (e.g., structure, clusters), tasks constraints, and attentional mechanisms. There is no specific vocabulary that is used for representing visual complexity.
These three levels of representation are not incompatible: for a particular task, the visual complexity space could be skewed towards a line (e.g. one perceptual property is dominant), but for a different task, the space of visual complexity might take into account multiple dimensions. The experiment presented below evaluates the format and content of the representation of visual complexity with the aim to tease apart the three levels of representation suggested above.

Experiment

The goal of the experiment is to study the representation of visual complexity while two groups of participants are told different definitions of visual complexity. Both groups performed a hierarchical grouping task with images of various levels of visual complexity. A hierarchical grouping task allows for identifying the explicit criteria participants used to perform a grouping task (see Oliva & Torralba, 2001) and helps to give a psychological interpretation of the axes provided by a multi-dimensional scaling algorithm (see Results section).

Method

Subjects Thirty-four students from an introduction to psychology course at Michigan State University participated in the study for course credits. Half were in the *control* group and the other half in the *structure* group.

Materials The present study used 100 pictures of indoor scenes. This subset was selected at random from a database of 1000 scenes previously ranked on their subjective visual complexity. The subset had the constraints to represent all levels of complexity along a scale from 1 to 100. The general scene database was originally composed from sources such as the web, magazines and various image databases. Since the volume of the space that a scene image represents is correlated with a given range of clutter (Torralba & Oliva, 2002), only scenes of a small volume range (indoors) were kept for this present study. Moreover, indoor scenes contain a greater variety of colors and objects in a variety of layouts compared to larger scaled environments (e.g. natural space, Oliva & Schyns, 2000).

Procedure The hierarchical grouping task was performed as follows (see Figure 2): starting with 100 pictures shown in a grid on a 23" Apple monitor, participants were asked to separate images into two groups on the screen, corresponding respectively to the most complex vs. the simplest scenes. In a second step, they were asked to split each group into two more subdivisions, and in a third step, split the four groups into two groups each, leading to a total of eight groups. For each subdivision, they were asked to follow a criterion corresponding to visual complexity

(simplicity) and give a verbal description of it. Participants could move each picture across boundaries at any stage, and see an enlarged version of the image by double clicking on it. Similarly to Heaps and Handel (1999), our *Control* group was told the following instruction: "Visual simplicity is related to how easy it will be to remember the image after seeing it for a short time. Visual complexity is related to how difficult it will be to give a verbal description of the image and how difficult it will be to remember the scene after seeing it for a short time." For the *Structure* group, the following instructions were given in addition to the control instructions: "Visual complexity is related to the structure of the scene and therefore, is not merely related to color or brightness. Simplicity is related to how you see that objects and regions are going well together. Complexity is related to how difficult it is to make sense of the structure of the scene". Both groups were forbidden to use a criterion related to the semantic class of the scene (e.g. kitchen) or the presence of a specific object or color.

Results

Table 1 summarizes a taxonomy corresponding to the most common criteria from the descriptions given by participants at the primary and secondary divisions. Each verbal description was recoded as a class of concepts. Some descriptions were a composition of concepts (e.g. pictures on the left seemed more *cluttered* whereas the ones on the right seemed more *open in space*), others were unique (e.g. *quantity* of objects). The percentage in Table 1 should be seen as an indicator of the strength of a perceptual property (most of the time used, often used or almost never used) and not as a fixed value, as variability among individual descriptions was high.

Table 1: Criteria of visual complexity used for the primary and secondary divisions and their % for both groups.

Criteria	Group:Structure	Group:control
Quantity of:		
object	*19*	*32*
detail	*8*	*8*
color	*2*	*19*
Quantity total	29	**59**
Clutter	**18**	5
Symmetry	**15**	2.5
Open Space	**18**	10
Organization	**13**	7
Contrast	<1	8

For the control group, where complexity was defined as a difficulty of verbal and visual recording, the criteria corresponding to *variety* and *quantity* of objects and color dominated the representation of complexity. In the second group where complexity was defined as relating to the structure of the scene, participants evenly used a set of criteria that the control group mentioned less frequently. The primary criterion of the structure group still concerns the quantity and variety of parts, participants referring either to the quantity of objects per se (19%), or the relationship between quantity of objects and spatial arrangement (18%, *clutter*). The other criteria were mostly concerned with spatial layout (symmetry, open space and organization {e.g. grid, centralized, cluster}).

For each condition, we investigated the consistency of the complexity ratings for the 100 images across subjects by computing a Spearman's rank-order correlation for each possible pairing of subjects (images within each subgroup were given the same complexity value, from 1 to 8). If participants were consistent, correlations among participants' rankings should be high. In both groups, Spearman's correlations were all statistically significant ($p < .01$) and were moderate to large in magnitude. Mean correlations of all the pair-wise comparisons were the same in the control and structure group, respectively, $r = 0.62$ and $r=0.61$; (stdev = 0.15 and 0.14).

Next, we applied a nonlinear dimensional reduction method (*Isomap*, Tenenbaum, de Silva, & Langford, 2000) onto a dissimilarity matrix constructed from participants' grouping for each condition (control and structure). To do so, a symmetric 100 x 100 matrix was constructed for each participant. Pairs of images placed in the same group versus in a different group were given respectively a score of 0 or a score of 1. Dissimilarity matrices from all participants from each condition were summed to create two pooled dissimilarity matrices. The Isomap analysis uses the dissimilarities of judgments given by human observers and provides a low dimensional visual representation of the mapping of proximities (i.e., distances) existing between images of various levels of complexity.

Figure 3 shows a two dimensional projection of the 100 images given by Isomap for the *Structure* group. The representation corresponds to the number of independent ways in which visual scenes can be perceived to resemble or differ in visual complexity. Although the dimensions per se are difficult to interpret and further experiments will be needed to assess more accurately the underlying dimensions of the space shown in Figure 3, it shows indeed a first principal direction corresponding to increasing "clutter" and quantity of objects. The second axis, illustrated in Figure 4, suggests an ordering along mirror symmetry and layout organization.

Albeit the correlation between the two first axes given by the Isomap representation for the structure and control group is nearly identical (0.98), the correlation between the ranks of images along the two second axes drops to 0.33 (see Figure 4), suggesting that participants used a different combination of criteria beside quantity while ranking the visual complexity of scenes. In the control group, participants were told that complexity was related to the difficulty of verbally describing an image. Consequently,

they estimated complexity almost exclusively based on the quantity and variety of objects and colors. In the structure group, participants were sensitive to spatial layout criteria, such as symmetry and open space.

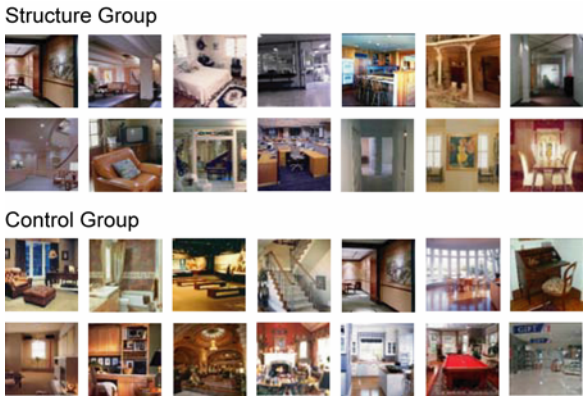


Figure 4: Sample of images projected onto the second principal dimension of *Isomap* for the structure group (top) and the control group (bottom). For the structure group, the images are organized from the top-left to the bottom-right following a property that resembles mirror symmetry. For the control group, the images are organized following a different combination of properties. These projections illustrate the differences in the criteria used between the two groups.

Discussion

The high correlations across participants for both groups (average of 0.61) suggest that participants used a same (or similar) set of holistic perceptual dimensions to represent complexity. The dimensions of visual complexity listed in Table 1 are not exhaustive: one can imagine that the perceived complexity of scenes of a larger volume of space (e.g., urban environments) might require new dimensions better suited to representing these spaces (e.g., perspective). However, the fact that there exists a set of defined properties that most people are sensitive to is appealing for modeling the visual complexity, where each dimension would be represented as a combination of low-level (e.g. contours, junctions) and medium-level features (e.g. connectedness, symmetry, Mack & Oliva, 2004). Furthermore, finding the true meaningful axes in the space generated by a multi-dimensional scaling algorithm, as well as the status of these dimensions (separable, integral, Garner, 1974; Gardenfors, 2000; Maddox, 1992) will be the subject of a follow-up study.

Conclusion

The goal of this study was to characterize the representation of visual complexity and its modulation by task constraints. The complexity ratings provided by observers on 100 pictures of (indoor) real-world scenes are consistent with a multi-dimensional representation of visual complexity. While the contribution of the dimensions are modulated by task constraints, visual complexity is principally represented by the perceptual dimensions of quantity of objects, clutter, openness, symmetry, organization, and variety of colors.

Acknowledgments

This research was partly funded by a graduate research assistantship to M.L.M (NSF-IGERT training grant) and A.O. was partly funded by an NIMH grant. We used the *Isomap* code in Matlab provided by J.B. Tenenbaum. The authors would like to thank Nancy Carlisle, Monica Castelhano, Zach Hambrick, Antonio Torralba as well as two anonymous reviewers for helpful comments about the paper. Correspondence can be addressed to A.O. (oliva@mit.edu), or M.L.M (mackmic1@msu.edu).

References

Gardenfors, P. (2000). *Conceptual spaces: the geometry of thoughts*. Bradford Books MIT Press.

Garner, W.R. (1974). The processing of information and structure. Potomac, MD: Erlbaum.

Feldman, J. (1997) Regularity-based perceptual grouping. Computational Intelligence, 13(4), 582-623.

Heaps, C., & Handel, C.H. (1999). Similarity and features of natural textures. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 299-320.

Heylighen F. (1997). *The Growth of Structural and Functional Complexity during Evolution*. F. Heylighen & D. Aerts (eds.).

Mack, M.L., & Oliva, A. (2004). The perceptual dimensions of visual simplicity. Presentation at the *4th Annual Meeting of Visual Sciences Society*, Sarasota, Florida.

Maddox, W.T. (1992). Perceptual and decisional separability. In Ashby, G.F., ed. *Multidimensional models of perception and cognition*, 147-180. Hillsdale, NJ: Lawrence Erlbaum.

Oliva, A., & Schyns, P.G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34, 72-107.

Oliva, A., & Schyns, P.G. (2000). Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*, 41,176-210.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42,145-175.

Palmer, S.E., (1999). Vision Science: Photons to Phenomenlogy. MIT Press.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509-522.

Rao, A.R., & Lohse, G.L. (1993). Identifying high-level features of texture perception. *Graphical Models and Image Processing*, 55, 218-233.

Schyns, P.G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychological Science*, 5, 195-200.

Tenenbaum, J.B., de Silva, V., & Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2319-2323.

Torralba, A., & Oliva, A. (2002). Depth Estimation from Image Structure. *IEEE Pattern Analysis and Machine Intelligence*, 24 (9), 1225-1238

van der Helm, P.A. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, 126, 770-800.

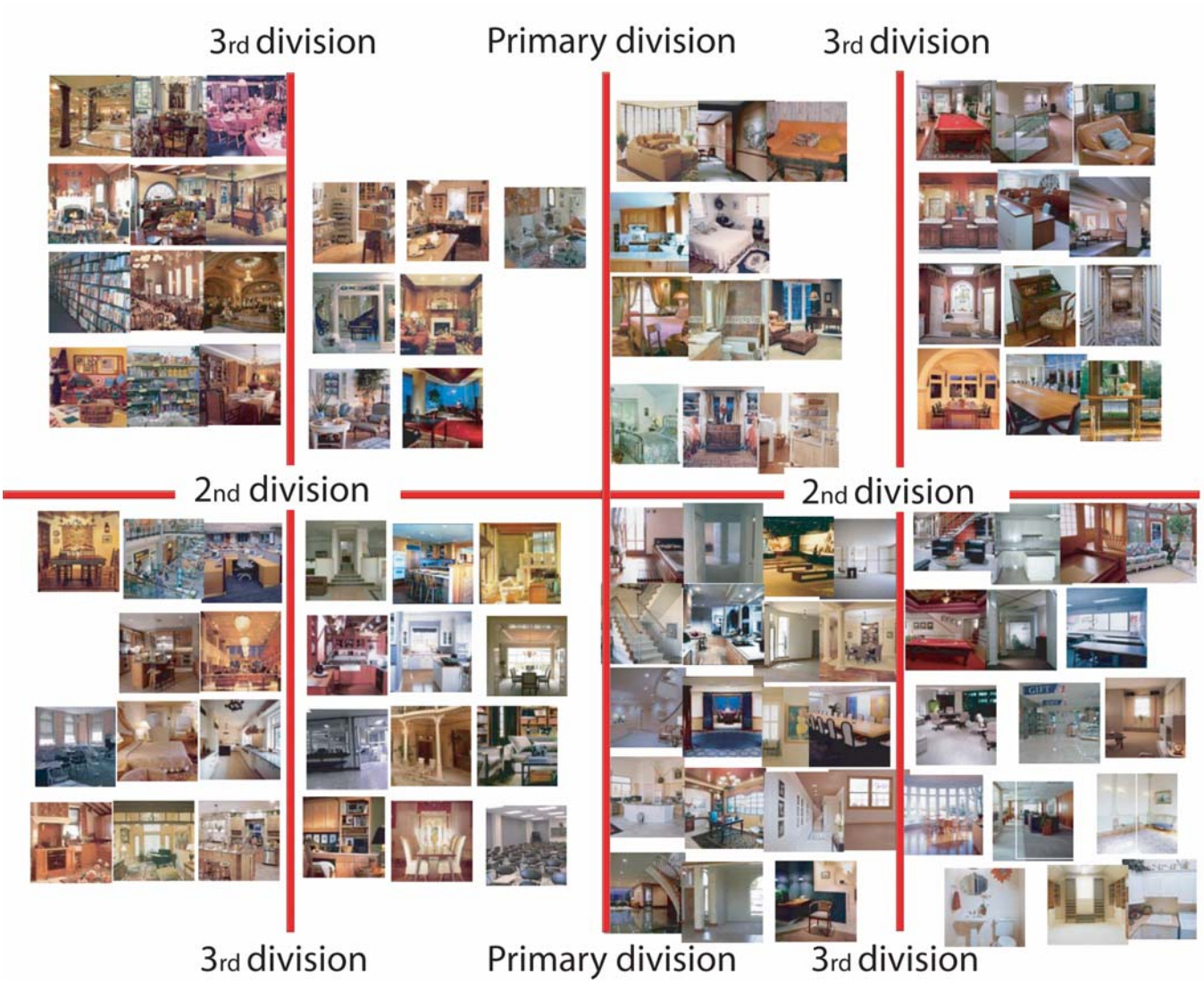


Figure 2: Illustration of the hierarchical grouping task after completion (organization made by subject 1 in the *Structure* group). Most complex scenes are in the top left corner, and most simple scenes are the bottom right corner.

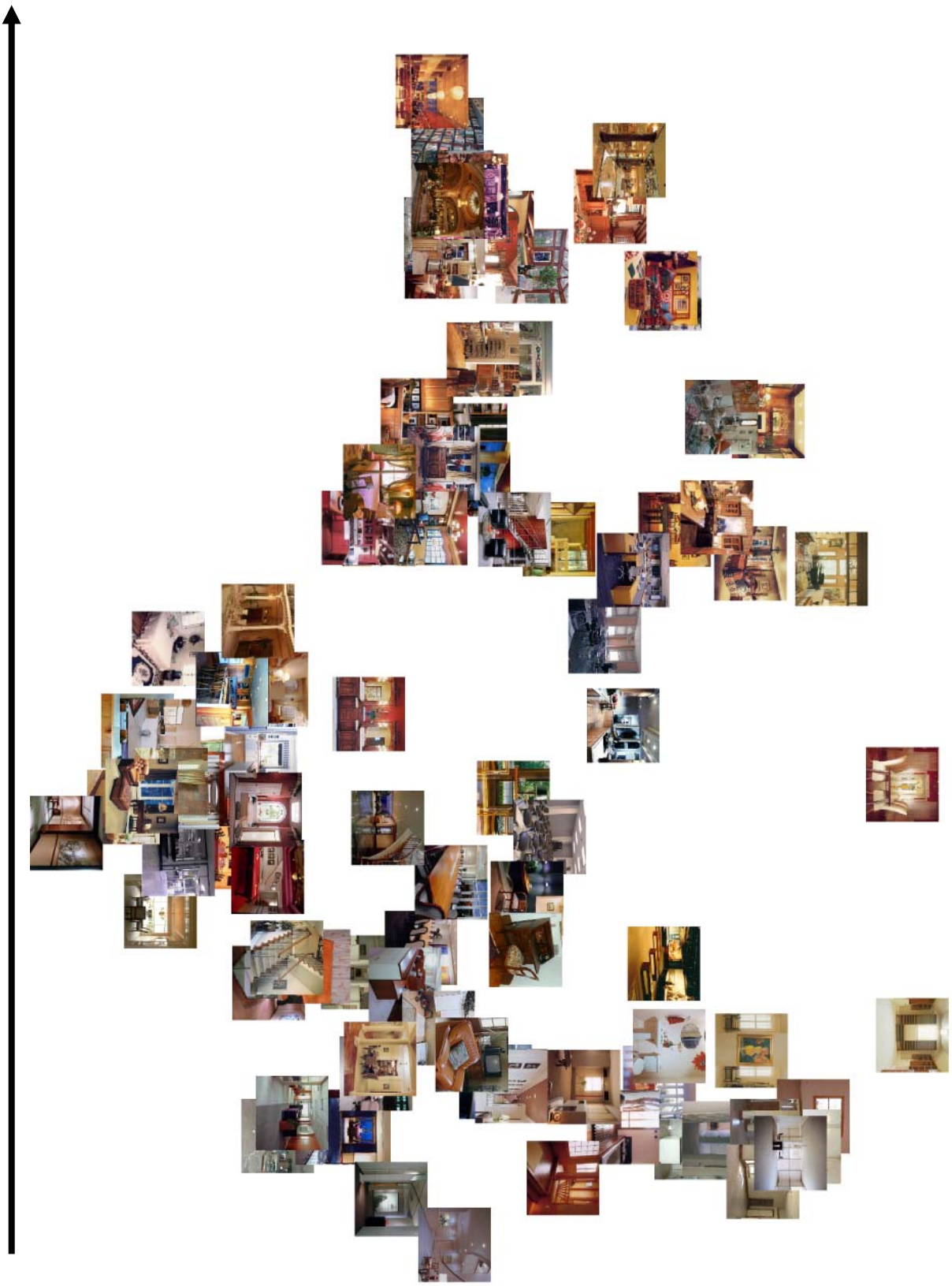


Figure 3: Representation given by Isomap for the structure group. The space shows on the arrow axis, a principal direction corresponding to increasing quantity of objects and clutter. The images that are far away from that direction are images that exhibit the highest amount of variability in how they were grouped in relation to other images. Scenes of medium and low level of clutter exhibit more variations along a second direction, possibly related to symmetry and spatial arrangement (cf. Figure 4).

Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies

Luca Onnis (lo35@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853, USA

Padraic Monaghan (P.Monaghan@psych.york.ac.uk)
Department of Psychology, University of York, York, YO10 5DD, UK

Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853, USA

Nick Chater (nick.chater@warwick.ac.uk)
Institute for Applied Cognitive Science and Department of Psychology, University of Warwick, Coventry, CV47AL, UK

Abstract

An important aspect of language acquisition involves learning the syntactic nonadjacent dependencies that hold between words in sentences, such as subject/verb agreement or tense marking in English. Despite successes in statistical learning of adjacent dependencies, the evidence is not conclusive for learning nonadjacent items. We provide evidence that discovering nonadjacent dependencies is possible through statistical learning, provided it is modulated by the variability of the intervening material between items. We show that generalization to novel syntactic-like categories embedded in nonadjacent dependencies occurs with either zero or large variability. In addition, it can be supported even in more complex learning tasks such as continuous speech, despite earlier failures.

Introduction

Statistical learning – the discovery of structural dependencies through the probabilistic relationships inherent in the raw input – has long been proposed as a potentially important mechanism in language development (e.g. Harris, 1955). Efforts to employ associative mechanisms for language learning withered during following decades in the face of theoretical arguments suggesting that the highly abstract structures of language could not be learned from surface level statistical relationships (Chomsky, 1957). Recently, interest in statistical learning as a contributor to language development has reappeared as researchers have begun to investigate how infants might identify aspects of linguistic units such as words, and to label them with the correct linguistic abstract category such as VERB. Much of this research has focused on tracking dependencies between *adjacent* elements. However, certain key relationships between words and constituents are conveyed in nonadjacent (or remotely connected) structure. In English, linguistic material may intervene between auxiliaries and inflectional morphemes (e.g., *is cooking*, *has traveled*) or between subject nouns and verbs in number agreement (*the books on the shelf are dusty*). The presence of embedding and nonadjacent relationships in language was a point of serious difficulty for early associationist approaches. It is easy to see that a distributional mechanism computing solely neighbouring information would parse the above sentence as …**the shelf is dusty*. Despite the importance of detecting remote dependencies, we know relatively little about the conditions under which this skill may be acquired by statistical means.

In this paper, we present results using the Artificial Language Learning (ALL) paradigm designed to test learning of nonadjacent dependencies in adult participants. We suggest that a single statistical mechanism might underpin two language learning abilities: detection of nonadjacencies and abstraction of syntactic-like categories from nonadjacent distributional information.

Despite the fact that both infants and adults are able to track transitional probabilities among adjacent syllables (Saffran, Aslin, & Newport, 1996), tracking nonadjacent probabilities, at least in uncued streams of syllables, has proven elusive in a number of experiments and the evidence is not conclusive (Newport & Aslin, 2004; Onnis, Monaghan, Chater, & Richmond, submitted; Peña, Bonatti, Nespor, & Mehler, 2002). Thus, a serious empirical challenge for statistical accounts of language learning is to show that a distributional learner can learn dependencies at a distance. Previous work using artificial languages (Gómez, 2002) has shown that the variability of the material intervening between dependent elements plays a central role in determining how easy it is to detect a particular dependency. Learning improves as the variability of elements that occur between two dependent items increases. When the set of items that participate in the dependency is small relative to the set of elements intervening, the nonadjacent dependencies stand out as invariant structure against the changing background of more varied material. This effect also holds when there is no variability of intervening material shared by different nonadjacent items, perhaps because the intervening material becomes invariant with respect to the variable dependencies (Onnis,

Christiansen, Chater, & Gómez, 2003). In natural language, different structural long-distance relationships such as singular and plural agreement between noun and verb may in fact be separated by the same material (e.g. *the books on the shelf are dusty* versus *the book on the shelf is dusty*). We call the combined effects of zero and large variability the *variability hypothesis*.

Very similar ALL experiments tested have failed to show generalization from statistical information unless additional perceptual cues such as pauses between words were inserted, suggesting that a distributional mechanism alone is too weak to support abstraction of syntactic-like categories. On these grounds Peña et al. (2002) have argued that generalization necessitates a rule-based computational mechanism, whereas speech segmentation relies on lower-level statistical computations. However, these experiments tested nonadjacency learning and embedding generalization with low variability of embedded items, which we contend is consistent with the variability hypothesis that learning should be hard. Our aim is to show that at the end-points of the variability continuum, i.e. with either no or large variability, generalization becomes possible. In Experiment 1, we present results suggesting that both detection of nonadjacent frames and generalization to the embedded items are simultaneously achieved when either one or a large number of different type items are shared by a small number of highly frequent and invariant frames. In Experiment 2 we also investigate whether tracking nonadjacent dependencies can assist speech segmentation and generalization simultaneously, given the documented bias for segmenting speech at points of lowest transitional probability (Saffran et al. 1996a,b).

We conclude that adult learners are able to track both adjacent and nonadjacent structure, and the success is modulated by variability. This is consistent with the hypothesis that a learning mechanism uses statistical information by capitalizing on stable structure for both pattern detection and generalization (Gómez, 2002, Gibson, 1991).

Generalising under variability

The words of natural languages are organized into categories such as ARTICLE, PREPOSITION, NOUN, VERB, etc., that form the building blocks for constructing sentences. Hence, a fundamental part of a language knowledge is the ability to identify the category to which a specific word, say *apple*, belongs and the syntactic relationships it holds with adjacent as well as nonadjacent words. Two properties of word class distribution appear relevant for a statistical learner. First, closed class words like articles and prepositions typically involve highly frequent items belonging to a relatively small set (*am*, *the*, *-ing*, *-s*, *are*) whereas open class words contain items belonging to a very large set (e.g. nouns, verbs, adjectives). Secondly, Gómez (2002) noted that sequences in natural languages involve members of the two broad categories being interspersed. Crucially, this asymmetry translates into patterns of highly invariant *nonadjacent* items, or frames, separated by highly variable material (*am cooking*, *am working*, *am going*, etc.). Such sequential asymmetrical properties of natural language may help learners solve two complex tasks: a) building syntactic constructions that sequentially span one or several words; b) building relevant abstract syntactic categories for a broad range of words in the lexicon that are distributionally embedded in such nonadjacent relationships. Frequent nonadjacent dependencies are fundamental to the process of progressively building syntactic knowledge of, for instance, tense marking, singular and plural markings, etc. For instance, Childers & Tomasello (2001) tested the ability of 2-year-old children to produce a verb-general transitive utterance with a nonce verb. They found that children were best at generalizing if they had been mainly trained on the consistent pronoun frame *He`s VERB-ing it* (e.g., *He`s kicking it*, *He`s eating it*) rather than on several utterances containing unsystematic correlations between the agent and the patient slots (*Mary`s kicking the ball*, *John`s pushing the chair*, etc.).

Gómez (2002) found that the structure of sentences of the form $A_iX_jB_i$, where there were three different $A_i\_B_i$ pairs, could in fact be learned provided there was sufficient variability of X_j words. The structure was learned when 24 different Xs were presented, but participants failed to learn when Xs varied from sets of 2, 4, 6, or 12, i.e. with low variability. Onnis et al. (2003) replicated this finding and also found that learning occurred with only one X being shared, suggesting the nonadjacent structure would stand out again, this time as variant against the invariant X.

While Gómez interpreted her results as a learning bias towards what changes versus what stays invariant, thus leading to "discard" the common embeddings in some way, we argue here that there may be a reversal effect in noting that common elements all share the same contextual frames. If several words – whose syntactic properties and category assignment are *a priori* unknown – are shared by a number of contexts, then they will be more likely to be grouped under the same syntactic label, e.g. VERB. For instance, consider a child faced with discovering the class of words such as *break*, *drink*, *build*. As the words share the same contexts below, s/he may be driven to start extracting a representation of the VERB class (Mintz, 2002):

I am-X-ing
dont-X-it
Lets-X-now!

Mintz (2002) argued that most importantly, in hearing a new word in the same familiar contexts, for instance *eat* in *am-eat-ing*, the learner may be drawn to infer that the new word is a VERB. Ultimately, having categorized in such a way, the learner may extend the usage of *eat* as a VERB to new syntactic constructions in which instances of the category VERB typically occur. For instance s/he may produce a novel sentence *Lets-eat-now!* Applying a category label to an word (e.g. *eat* belongs to VERB) greatly enhances the generative power of the linguist system, because the labeled item can now be used in new syntactic contexts where the category applies. In Experiment 1 we tested whether

generalization to new X items in the *A_X_B* artificial grammar used by Gómez (2002) and Onnis et al. (2003) is supported under the same conditions of no or large variability that affords the detection of invariant structure. Hence, if frames are acquired under the variability hypothesis, generalization will be supported when there is either zero or large variability of embeddings. Likewise, because invariant structure detection is poor in conditions of middle variability, generalization is expected to be equally poor in those conditions too.

Experiment 1

Method

Subjects

Thirty-six undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each.

Materials

In the training phase participants listened to auditory strings generated by one of two artificial languages (L1 or L2) of the type $A_iX_jB_i$. Strings in L1 had the form $A_1X_jB_1$, $A_2X_jB_2$, and $A_3X_jB_3$. L2 strings had the form $A_1X_jB_2$, $A_2X_jB_3$, $A_3X_jB_1$. Variability was manipulated in 3 conditions – zero, small, and large– by drawing X from a pool of either 1, 2 or 24 elements. The strings, recorded from a female voice, were the same that Gómez used in her study and were originally chosen as tokens among several recorded sample strings in order to eliminate talker-induced differences in individual strings.

The elements A_1, A_2, and A_3 were instantiated as *pel*, *vot*, and *dak*; B_1, B_2, and B_3, were instantiated as *rud*, *jic*, *tood*. The 24 middle items were *wadim*, *kicey*, *puser*, *fengle*, *coomo*, *loga*, *gople*, *taspu*, *hiftam*, *deecha*, *vamey*, *skiger*, *benez*, *gensim*, *feenam*, *laeljeen*, *chla*, *roosa*, *plizet*, *balip*, *malsig*, *suleb*, *nilbo*, and *wiffle*. The middle items were stressed on the first syllable. Words were separated by 250-ms pauses and strings by 750-ms pauses. Three strings in each language were common to all two groups and they were used as test stimuli. The three L2 items served as foils for the L1 condition and vice versa. The test stimuli consisted of 12 strings randomized: six strings were grammatical and six were ungrammatical. The ungrammatical strings were constructed by breaking the correct nonadjacent dependencies and associating a head to an incorrectly associated tail, i.e. $*A_iXB_j$.. Six strings (three grammatical and three ungrammatical) contained a previously heard embedding, while 6 strings (again three grammatical and three ungrammatical) contained a new, unheard embedding. Note that correct identification could only be achieved by looking at nonadjacent dependencies, as adjacent transitional probabilities were the same for grammatical and ungrammatical items.

Procedure

Six participants were recruited in each of 3 Variability conditions (1, 2 and 24) and for each of two Language conditions (L1, L2) resulting in 12 participants per Variability condition. Learners were asked to listen and pay close attention to sentences of an invented language and they were told that there would be a series of simple questions relating to the sentences after the listening phase. During training, participants in the two conditions listened to the same overall number of strings, a total of 432 token strings. This way, frequency of exposure to the nonadjacent dependencies was held constant across conditions. Participants in set-size 24 heard six iterations of each of 72 type strings (3 dependencies x 24 middle items), participants, in set-size 2 encountered each string 12 times as often as those exposed to set size 24, and so forth. Hence, whereas nonadjacent dependencies where held constant, transitional probabilities of adjacent items decreased as set size increased.

Training lasted about 18 minutes. Before the test, participants were told that the sentences they had heard were generated according to a set of rules involving word order, and they would now hear 12 strings, 6 of which would violate the rules. They were asked to give a "Yes/No" answer. They were also told that the strings they were going to hear may contain new words and they should base their judgment on whether the sentence was grammatical or not on the basis of their knowledge of the grammar. This is to guarantee that participants did not select as ungrammatical all the sentences with novel words simply because they contained novel words.

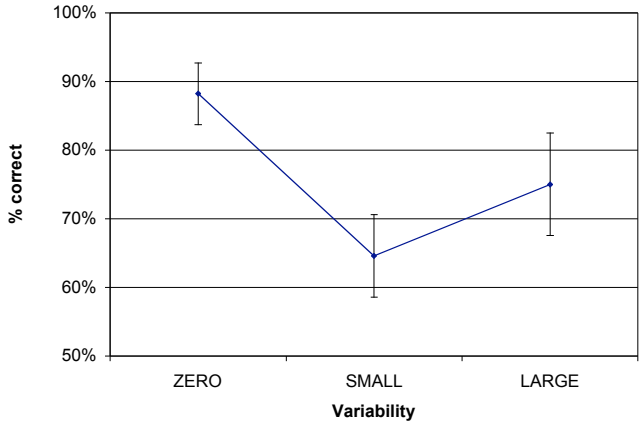


Figure 1. Generalisation under variability - Exp.1

Results and discussion

An analysis of variance with Variability (1 vs. 2 vs. 24) and Language (L1 vs. L2) as between-subjects and Grammaticality (Trained vs. Untrained strings) as a within-subjects variable resulted in a main Variability effect, $F(2,30)= 3.41$, $p< .05$, and no other interaction. Performance across the different variability conditions resulted in a U-shaped function: a polynomial trend analysis showed a significant quadratic effect, $F(1, 35) =7.407$, $p <.01$. Figure 1 presents the percentage of endorsements for total accuracy in each of the three variability conditions. These results add considerable power to the variability hypothesis: not only can nonadjacencies be detected, but generalization too can occur distributionally, and *both* processes seem to be modulated by the same conditions of variability. In addition, generalization with zero variability allows us to disambiguate previous results, in that the high performance obtained by Onnis et al. (2003) could have been due to a simple memorization of the 3 strings repeated over and over

again during training. However, in Experiment 1 correct classification of new strings as grammatical can only be done on the basis of the correct nonadjacencies. Thus, it seems that learning on zero or large variability conditions is supported by a similar mechanism. Finally, we note that A and B words are monosyllabic and X words are bysillabic, participants could simply learn a pattern S-SS-S (where S=syllable). However, because all sentences display such pattern across conditions this cannot explain the U-shape of the learning curve.

Experiment 2

In Experiment 1 the items of the grammar are clearly demarcated by pauses. It can be argued that this makes the task somewhat simplified with respect to real spoken language, which does not contain for instance such apparent cues at every word boundary. In addition, the embedded item *X* was instantiated in bisyllabic words (as opposed to monosyllabic *A* and *B* words), providing an extra cue for category abstraction. In this context, Peña et al. (2002) have argued that generalization and speech segmentation are separate processes underpinned by separate computational mechanisms: statistical computations are used in a segmentation task but this is not performed simultaneously with algebraic computations that would permit generalizations of the structure. Once the segmentation task was solved by introducing small pauses in the speech signal, their underlying structure was learned. Hence it is important to test these claims in the light of the variability hypothesis, which we argue might provide the key to learning nonadjacencies and generalizing altogether, even in connected speech, without invoking two separate mechanisms.

Recent attempts to show statistical computations of a higher order at work in connected speech with a similar AXB language have met with some difficulty: Newport & Aslin (2004), for instance, exposed adults to a continuous speech stream, created by randomly concatenating *AXB* words with 3 *A_B* syllable dependencies and with 2 different middle *X* syllables. A sample of the speech stream obtained would be …$A_1X_3B_1A_2X_2B_2A_3X_1B_3$…. In this case participants were unable to learn the nonadjacent dependencies. Concatenating words seamlessly adds considerable complexity to the task of tracking statistical information in the input for two main reasons: first, transitional probabilities between words of a language containing, say 3 dependencies and 3 Xs, p($B|A$)= 0.5 are higher than within words, p($X|A$) and p($B|X$)= 0.33, and this pressures for segmentation within words (Saffran, Aslin, & Newport, 1996ab). Secondly, assuming the statistical mechanism is sensitive to nonadjacent dependencies as seems the case in Experiment 1, concatenating items entails the additional burden of tracking nonadjacent transitional probabilities across word boundaries, e.g. $X_3\_A_2$, $B_1\_X_2$, and dependencies spanning *n* words away can in principle also be attended to, e.g. two items away ($B_2\_\ \_A_3$…,etc.). One can readily see that if all transitional probabilities of different order were to be computed this scenario would soon create a computational impasse. The insight from Gómez (2002) and Experiment 1 is that variability plays a key role, in that it allows adjacent dependencies to be overcome in favour of nonadjacent ones, but it remains to be seen whether this can be done in connected speech too.

Peña et al. (2002) tested participants on whether they learned to generalize from the rules of an *AXB* language very similar to Newport & Aslin (2004) in unsegmented speech. Again *AXB* items were instantiated in syllables and formed words concatenated one to the other seamlessly. At test, participants demonstrated no preference for so-called "rule-words", new trigram sequences that maintained the $A_i\_B_i$ nonadjacent dependencies but contained a different *A* or *B* in the intervening position (e.g., $A_1B_3B_1$), compared to part-words, i.e., sequences that spanned word boundaries (e.g., $X_2B_1A_3$, or $B_3A_1X_2$). In a further manipulation, 25-ms gaps were introduced between words during the training phase of the experiment, and now participants generalized as indicated by a preference for rule-words over part-words. Peña et al. claimed that altering the speech signal resulted in a change in the computations performed by their participants. Statistical computations were used in a (previously successful) segmentation task but this was not performed simultaneously with algebraic computations that would permit generalizations of the structure. They argued that once the segmentation task was solved by introducing small gaps in the speech signal, the underlying structure would be learned. However, using the same stimuli and experimental conditions as Peña et al. Onnis, Monaghan, Chater & Richmond (submitted) found that rule-words were preferred over part-words in both segmentation and generalization tasks even when the nonadjacent structure was eliminated: participants reliably preferred incorrect rule-words $*A_1B_3B_2$ to part-words B_1A_2X, due to preference for plosive sounds in word-initial position. Hence such preference did not reflect learning of nonadjacent dependencies. Although discouraging at first sight, all these negative results are not inconsistent with the variability hypothesis. In fact, they are all cases structurally similar to the low-variability condition in Gómez (2002) and Experiment 1. Thus, in Experiment 2 we tested whether with sufficiently large variability:

a) tracking higher-order dependencies can be used to segment speech. This is a difficult task because it implies overriding even lower transitional probabilities p($X|A$) than previously tested and this pressures for segmentation within word boundaries (Saffran et al. 1996);

b) generalization of the embeddings can occur *simultaneously* to speech segmentation, i.e. on-line in running speech, and can be done by statistical analysis of the input alone, i.e. without additional perceptual cues such as pauses. We tested this using the same material and training conditions as Peña et al. for their unsuccessful pause-free generalization task, but increasing the variability of the *X* syllables to 24 items as in Experiment 1.

Method

Subjects

20 undergraduate and postgraduate students at the University of Warwick participated for £1. All participants spoke English as a first language and had normal hearing.

Materials

We used the same nine word types from Peña et al.'s Experiment 2 to construct the training speech stream in our Experiment 2. The set of nine words was composed of three groups ($A_i\_B_i$), where the first and the third syllable were paired, with an intervening syllable (X) selected from one of either three syllables (low variability condition) or 24 syllables (high variability condition). The syllables were randomly generated from the following set of consonants: /p/,/b/,/g/,/k/,/d/,/t/,/l/,/r/,/f/,/tʃ/,/dʒ/,/n/, /s/,/v/,w/,/m/,/θ/,/ʃ/,/z/and the following vowels: /ɛi/,/uw/,/a/,/iy/,/au/,/oi/,/ai/,/æ/,/œ/.

Consonants and vowels were permuted, then joined together. No syllables occurred more than once in the set of 33 generated. Each participant listened to a different permutation of consonant-vowel pairings. Notice that the language structure in the two conditions match very closely those of small and large variability in Experiment 1. Unlike Experiment 1 all items were monosyllabic and equally stressed.

Words were produced in a seamless speech stream, with no two words from the same set occurring adjacently, and no same middle item occurring in adjacent words. Hence, adjacent transitional probabilities were as follows: for the small variability condition, and within words, $p(\underline{X}|\underline{A})$ and $p(\underline{B}|\underline{X})= 0.33$; between adjacent words $p(\underline{B_i}|\underline{A_i})= 0.5$. Nonadjacent transitional probabilities were $p(B_i|A_i)= 1$, $p(A_i|X_{previous})= 0.33$, $p(X_j|B_{previous})= 0.33$. For the large variability condition all probabilities were the same except within word adjacent probabilities $p(X|A) = 0.041$.

Therefore, the predicition is that if learners computed adjacent statistical probabilities they should prefer part-words and perhaps significantly more in the large variability condition. Conversely, if they computed nonadjacent dependencies they would rely on the most statistically reliable ones, namely $p(B_i|A_i)= 1$, i.e. they would segment correctly at word boundary.

We used the Festival speech synthesizer using a voice based on British-English diphones at a pitch of 120 Hz, to generate a continuous speech stream lasting approximately 10 minutes. All syllables were of equal duration, and were produced at a rate of 4.5 syllables/second. Words were selected randomly, except that no $A_i\_B_i$ pair occurred twice in succession. The speech stream was constructed from 900 words, in which each word occurred approximately 100 times. The speech stream faded in for the first 5 seconds, and faded out for the last 5 seconds, so there was no abrupt start or end to the stream. In addition, and crucially, for each participant, we randomly assigned the 9 syllables from the first experiment to the A_i, B_i and X_j positions. Thus, each participant listened to speech with the same structure containing the nonadjacent dependencies, but with syllables assigned to different positions. This was to avoid any bias towards choosing a rule-word because of a preference for plosive sounds, as Onnis et al. (submitted) demonstrated. Part-words were formed from the last syllable of one word and two syllables from the following word (B_iA_jX), or from the last two syllables of one word and the first syllable from the following word (XB_iA_j).

Procedure

In the training phase, participants were instructed to listen to continuous speech and try and work out the "words" that it contained. They then listened to the training speech. At test part-words were compared to "rule-words", which were composed of $A_i\_\_B_i$ pairs with an intervening item that was either an A_j or a B_j from another $A_j\_B_j$ pair. Participants were requested to respond which of two sounds was a "word" in the language they had listened to. They were then played a "rule-word" and a part-word separated by 500 ms, and responded by pressing either "1" on a computer keyboard for the first sound a word, or "2" for the second sound a word. After 2 seconds, the next rule-word and part-word pair were played. In half of the test trials, the "rule-words" occurred first. Five participants heard a set of test trials with one set of words first, and the other 5 participants heard the other set of words first.

Results

The results are shown in Figure 2. In line with the original Peña et al.'s experiment, we found no evidence for participants learning to generalize from the nonadjacent structure of the stimuli in the low-variability condition. Participants responded with a preference for rule-words over part-words 41.9% of the times, which was significantly lower than chance, $t(9) = -2.73$, $p < .05$. Conversely, in the high-variability condition participants preferred rule-words 63.3% of the times, significantly higher than chance, $t(9)= -3.80$, $p = .0042$. In addition, there was a significant difference between the low variability and the high variability condition, $t(18) = -4.68$, $p < .001$.

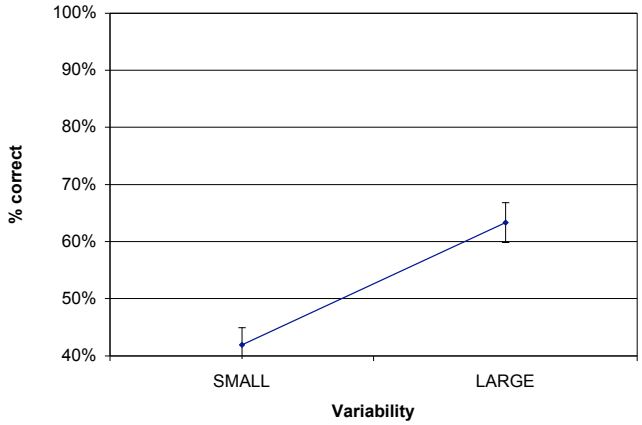


Figure 2. Generalisation in unsegmented speech - Exp. 2

General Discussion

Statistical learning of dependencies between adjacent elements in a sequence is fast, robust, automatic and general in nature. In contrast, although the ability to track remote dependencies is a crucial linguistic ability, relatively little research has been directed toward this problem. Nonadjacent structure in sequential information seems harder to learn, possibly because learners have to overcome the bias toward adjacent transitional probabilities. In fact, a statistical learning mechanism that kept track of all possible adjacent and nonadjacent regularities in the input, including syllables one, two, three away, etc., would quickly

encounter a computationally intractable problem of exponential growth. It would seem that either statistical learning is limited to sensitivity to adjacent items, or there may be statistical conditions in which adjacencies become less relevant in favour of nonadjacencies. It has been suggested that this applies under conditions of large variability of the intervening material (Gómez, 2002) or zero variability (Onnis et al., 2003). This paper contributes some steps forward: first, Experiment 1 shows that variability is the key not only for detection of remote dependencies but also for generalization of embedded material, fostering the creation of abstract syntactic-like classes, which is often assumed to require higher-level algebraic computation. Secondly, in Experiment 2 segmentation and generalization are achieved simultaneously, without the assist of pauses (a difference in signal) as Pena et al. claimed. Consequently, rather than supporting a statistical/algebraic distinction our results suggest specific selectivities in learning patterned sequences. The specific characterization of such selectivities may not be simple to identify: Newport & Aslin (2004) found that nonadjacent segments (consonants and vowels) could be learned but not nonadjacent syllables, and proposed that this accounts for why natural languages display nonadjacent regularities of the former kind but not of the latter. Experiment 2, however, shows that with large variability nonadjacent syllabic patterns can in fact be learned. The key factor for success is again variability. Experiment 2 also shows that learners are indeed able to track nonadjacent dependencies in running speech, despite the well documented bias for adjacent associations and the preference for segmenting continuous speech at points of lowest transitional probabilities.

Overall, the results suggest that the learning mechanism entertains several statistical computations and implicitly "tunes in" to statistical relations that yield the most reliable source of information. This hypothesis was initiated by Gómez (2002) and is consistent with several theoretical formulations such as reduction of uncertainty (Gibson, 1991) and the simplicity principle (Chater, 1996) that the cognitive system attempts to seek the simplest hypothesis about the data available. In the face of performance constraints and way too many statistical computations, the cognitive system may be biased to focus on data that will be likely to reduce uncertainty. Specifically, whether the system focuses on transitional probabilities or nonadjacent dependencies may depend on the statistical properties of the environment that is being sampled.

Our work ties in with recent acquisition literature that has emphasized the constructive role of syntactic frames as the first step for building more abstract syntactic representations (Tomasello, 2003 for an overview). Children's syntactic development would build upon several consecutive stages from holophrases such as *I-wanna-see-it* (at around 12 months), to pivot-schemas (*throw-ball*, *throw-can*, *throw-pillow*, at about 18 months), through item-based constructions (*John hugs Mary*, *Mary hugs John*, at about 24 months), to full abstract syntactic constructions (*a X*, *the Xs*, *Eat a X*).

Statistical learning seems, at least in adults, powerful enough to allow the discovery of complex nonadjacent structure, but simply not any condition will do: we have suggested that variability such as that emerging from the asymmetry between open and closed class words may be a crucial ingredient for understanding the building of language.

Acknowledgments

We thank M. Merkx for running Exp. 2, and R. Gómez for the stimuli in Exp.1 and important insights. Part of this work was conducted while L. Onnis and P. Monaghan were at the University of Warwick. Support comes from European Union Project HPRN-CT-1999-00065, and Human Frontiers Science Program.

References

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103, 566-581.

Childers, J. & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology*, *37*, 739-748.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Gibson, E.J. (1991). *An Odyssey in Learning and Perception.* Cambridge, MA: MIT Press.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-436.

Harris, Z.S. (1955). From phoneme to morpheme. *Language 31*, 190-222.

Mintz, T.H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition* , *30*, 678-686.

Newport, E.L., & Aslin, R.N. (2004). Learning at a distance I. Statistical learning of nonadjacent dependencies. Cognitive Psychology, *48*, 127-162.

Onnis, L., Monaghan, P., Chater, N., & Richmond, K. (submitted). Phonology impacts segmentation and generalization in speech processing.

Onnis, L., Christiansen, M., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from Artificial Grammar Learning. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 887-891.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298,* 604-607.

Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996a). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Self-Explanation Reading Training: Effects for Low-Knowledge Readers

Tenaha O'Reilly (t.oreilly@mail.psyc.memphis.edu)
Rachel Best (r.best@mail.psyc.memphis.edu)
Danielle S. McNamara (d.mcnamara@ mail.psyc.memphis.edu)

Psychology Department, University of Memphis
Memphis, TN 38152 USA

Abstract

This study examined the effects of self-explanation reading strategy training (SERT) and Preview training on high-school students' comprehension of a science text. The students (n=136) were from a middle to lower SES, inner-city Virginia high school. They were assessed in terms of their science knowledge and reading skill. Nine biology classes were then randomly assigned to SERT, Preview, or Control conditions. Science comprehension was assessed both immediately after training and after a one-week interval. The results indicated that after the one-week retention interval, SERT participants outperformed both Preview and Control participants on passage comprehension. This comprehension advantage was particularly enhanced for low-knowledge readers. This result replicates with high-school student findings reported by McNamara (in press) with college students.

Introduction

Many high-school students encounter difficulties comprehending their textbooks, particularly those covering scientific material (Bowen, 1999; Snow, 2002). Problems with comprehension can occur for a variety of reasons. One source of difficulty occurs from text specific factors, such as text cohesion (Beck, McKeown & Gromoll, 1989; McNamara, Kintsch, Songer, & Kintsch, 1996). Another source of problems stems from the reader's aptitudes. Of course, efficient decoding abilities are necessary for the reader to understand the words in the sentences (e.g., Perfetti, 1985). However, comprehension difficulties also occur even for readers who understand the words. These comprehension problems can emerge from the inability to draw inferences (e.g., Long, Oppy, & Seely, 1994) and the failure to apply other higher-level reading skills, such as meta-cognitive reading strategies (Cornoldi & Oakhill, 1996).

The recent RAND report on *Reading for Understanding* (Snow, 2002) documents the pressing need to improve reading comprehension. The RAND report also provides a useful heuristic for conceptualizing reading comprehension, which includes four interactive components: Characteristics of the text, the reader, the comprehension activities, and the socio-cultural context. Accordingly, these factors rarely operate in isolation, and as such, potential interactions between attributes associated with these factors need to be considered in order to develop a more complete understanding of reading comprehension processes.

In terms of text characteristics, research has shown that the structure of the text plays a major role in comprehension. For example, in an analysis of the cohesion of social studies texts, Beck et al. (1989) found that many texts have structures that are far from optimal in terms of promoting deep comprehension. Texts often present too much information with too little detail, contain loose unconnected statements, and have poor integration with previous sections. Overall, the manner in which many informational texts are written present challenges for comprehension.

The knowledge and skills readers bring into the reading situation play an important role in the comprehension of informational texts (McNamara & Kintsch, 1996; Voss & Silfies, 1996). For example, Voss and Silfies found that prior knowledge helps the reader spell out causal links between concepts, whereas reading skill contributes to a better textbase understanding. Indeed, other research has shown that conceptual gaps in text are most easily repaired with domain knowledge (McNamara et al, 1996; cf. McNamara, in press).

Comprehension activities, or the use of higher level reading strategies, also have a major impact on learning (Pressley, Wood, Woloshyn, Martin, King, & Menke, 1992). For example, Chi, De Leeuw, Chiu, and LaVancher (1994) have found positive learning gains for a technique called self-explanation. Students who were asked to explain what they learned from a biology text made more inferences, integrated more information across topics, and developed a deeper understanding of the text than did Control students.

Similarly, McNamara (in press; McNamara & Scott, 1999) developed a comprehensive reading strategy intervention called self-explanation reading strategy training (SERT). SERT helps improve comprehension by encouraging students to make use of various reading

strategies to build connections between the reader's knowledge and the text. It teaches students to use active reading techniques including comprehension monitoring, logic and common sense, elaboration, paraphrasing, bridging inferences, and prediction to improve their ability to explain a text and understand it at a deeper level. McNamara (in press) found that college students who were trained to use SERT outperformed controls on measures of text comprehension. This improvement was especially enhanced for low-knowledge readers on text-based questions (assessing knowledge of information explicitly stated in the text).

While the effects of SERT have been beneficial in improving comprehension with college students, the effects of SERT training on high-school students has not been tested. Examining the influence of learning strategies is especially important in high-school populations because there is a dearth of strategic reading interventions being taught and used in classrooms (Cox, 1997, Garner, 1990). Moreover, in comparison to other countries, students in the United States are falling behind students in other countries on measures of reading comprehension (Snow, 2002).

The purpose of this study was to examine the effect of SERT training on high-school students' ability to comprehend science texts. We compared students' comprehension of science texts after they had received either SERT training, Preview training, or a reading Control condition. Previewing is a reading strategy designed to help students better comprehend texts by encouraging them to preview various subsections of the text before they read. This strategy training is based on, and includes the K-W-L instructional technique developed by Ogle (1986). The use of preview techniques is relatively common and has been associated with comprehension gains (Johnston & Allington, 1991).

Our goal was to examine whether SERT or Preview training helped students to better understand the science passages in comparison to the Control condition. First, we predicted that students trained with SERT and Previewing would outperform students in the reading Control group who did not receive reading strategy training. We expected both types of interventions to facilitate students' science text comprehension. However, we expected the benefits of SERT and Previewing to depend on the students' level of prior knowledge. This prediction arises from the way in which each technique encourages the use of prior knowledge. Previewing focuses on the activation of relevant prior knowledge before reading the text to provide a mental schema of the contents. While prior research has clearly shown the benefits of schemas to improve comprehension, we doubt that they will be particularly useful if the student has little prior knowledge about the topic. As such, we expect the benefits of previewing to depend largely on the amount of prior knowledge the student has about the text topic, primarily benefiting relatively high-knowledge students. In contrast, SERT encourages students to actively process texts (e.g., elaborate and link information contained in the text). The student learns to use whatever knowledge available, including logic and general knowledge, to make sense of the text. Our previous research demonstrated that SERT was primarily beneficial to low-knowledge students (McNamara, in press). We expect similar results here.

Method

Nine biology classrooms within an inner-city school were randomly assigned to one of three conditions: SERT, Preview, or Control. Students in the three SERT classrooms were provided training on how to self-explain text using five reading sub-strategies. Preview participants were taught how to preview the text before they read it. Control classrooms were simply asked to read the text. After reading, they were asked to focus on any strategies they used to help them better remember the text. Comprehension was assessed using two science passages taken from school science texts. The passages differed in topic and length, which allowed us to examine whether the strategies were used across different science texts. We used both text-based and bridging-inference questions to assess students' comprehension of the science passages.

Participants

The sample consisted of 136 ninth and tenth-grade biology classes. Students were of mixed gender and ethnicity. The high school was located in an inner city region of Norfolk, Virginia.

Materials

Student aptitudes were measured with two tests; a modified version of the Gates-MacGinitie Reading Skill Test and a Prior Knowledge Test. The Gates-MacGinitie test is a standardized reading comprehension test, designed for grades 10-12. The test consisted of 40 multiple-choice questions designed to assess student comprehension on several short text passages (Cronbach's Alpha α=.91). Due to time constraints, the vocabulary section of the test was not administered, and the time limit for the comprehension question section was reduced to 15 minutes. The prior knowledge test consisted of 35 multiple-choice items which tap knowledge of different science domains including biology, scientific methods, mathematics, earth science, physics, mathematics, and chemistry (Cronbach's Alpha α=.74).

To examine immediate and long-term retention, there were two comprehension testing sessions; one immediately following training, and again one week after training. A text on viruses, describing the structure, and reproduction of viruses as well as some examples of viruses and how they relate to disease, was used during the immediate test. The passage was 1216 words in length with a Flesch Reading Ease of 45.1 and Flesch-Kincaid Grade level of 10.6. One week later, we administered a text about earthquakes, which described the causes of earthquakes and the conditions under which they occur. The earthquake text was 749 words in length with a Flesch Reading Ease of 65.1 and Flesch-Kincaid Grade level of 7.5. In an ideal laboratory setting, the order of the texts would be counterbalanced between testing sessions. However, this was not possible due to logistical constraints (e.g., high rates of absenteeism and movement of students between classes).

Reading comprehension was assessed with a set of 8 open-ended and 8 multiple-choice questions; half were text-based and half were bridging-inference questions. For each comprehension assessment, students were asked to indicate whether they had finished reading and, if not, how far they had read.

Self-Explanation Reading Training (SERT) was delivered in three main phases; introduction, demonstration, and practice (see McNamara, in press). During the introduction phase, participants were provided with a description and examples of self-explanation. The instructor defined and provided examples for five reading strategies: comprehension monitoring, paraphrasing, elaboration/logic and common sense, prediction, and bridging.

During the demonstration phase, participants watched a video depicting a student reading and self-explaining a text about forest fires. Participants could refer to the accompanying video transcript during viewing. The video was paused at various points, and participants identified and discussed the strategies being used by the student in the video. In the practice phase, the participants worked in pairs to practice self-explanation while reading a chapter from their science textbook. The participants took turns self-explaining, alternating after each paragraph. At the end of each paragraph, the partner who was listening (and not self-explaining) summarized the paragraph.

Preview training was also delivered in the form of three phases. During the introduction, participants were given a description of the basic Preview strategy; a review of subsections in a text that can be previewed (Title, introduction, objectives bold italics, pictures/figures, conclusion and chapter review questions) and strategies for note-taking during reading (questions such as, What I know, What I need to know. What I found out). During demonstration, the instructor demonstrated the Preview strategies to the class with a text on forest fires. Finally, during practice, the students practiced using the preview techniques on a chapter taken from their textbook.

In the Control condition, participants read the science texts to which trained groups were exposed during training. Participants wrote down strategies they used while reading, but did not discuss the strategies.

Design and Procedure

Each class was randomly assigned to one of the three conditions; SERT, Preview, or Control (three classes were assigned to each condition). Experimental sessions were conducted during students' regular classroom time by two experimenters. Prior to training, students completed the prior knowledge and Gates-McGinitie Reading test. A 15-minute time limit was given for each test. The prior knowledge test was administered first.

SERT and Preview training were conducted during two class periods conducted on consecutive days. Participants were told that the purpose of the study was to learn strategies that would help them to better understand and remember what they read. The total amount of time spent learning the experimental interventions during the two days was approximately the same. The Control condition required one class period. Students in the Control condition were told that the purpose of the training was to find out about the strategies they use when reading their textbooks.

Immediately after training and after a one-week interval, participants were given 30 minutes to read the science passage and answer comprehension questions. In the experimental conditions, the experimenter briefly reviewed the strategies for either SERT or Preview before beginning the comprehension test. For both the virus and earthquakes texts, the students did not have the text available when they answered the questions.

Results

Pre-test Scores

Scores on the pretest measures of prior knowledge and reading skill were examined as a function of condition to ensure that the groups were comparable on these measures. To do this, we conducted two univariate ANOVAs, with condition as the between-subjects factor (SERT, Preview, and Control) and student aptitude scores (either prior knowledge or reading Gates-MacGinitie). There were no reliable differences in prior knowledge scores ($F(2,135)=0.74$, $MSE=0.017$, $p>0.05$) or Gates-MacGinitie scores ($F(2,134)=0.77$, $MSE=41.192$, $p>0.05$) as a function of condition.

To examine the effects of individual differences, we used a median split to divide students into high and low-knowledge groups, or high and low reading comprehension skill groups.

Effects of Training Condition

To assess the effectiveness of SERT training on students' comprehension of the virus and earthquake texts, we conducted two sets of mixed model ANOVAs. The within-subjects factors were question type (text-based or bridging-inference), question format (multiple-choice or open-ended) and the between-subjects factors condition (SERT, Preview, or Control) and either reading skill (high or low) or prior knowledge (high or low). A combined analysis of reading skill and prior knowledge could not be performed because there were too few participants. To avoid conflating effects of reading skill with prior knowledge, the participants' scores on the prior knowledge measure were entered as covariates in the reading skill analysis, and vice versa, the participants' scores on the reading skill measure were entered as covariates in the prior knowledge analysis.

There were no reliable main or interaction effects of condition on the immediate comprehension test (virus text). In contrast, the earthquake text administered one week later revealed reliable effects of training condition. Therefore, the present analysis focuses on the latter results.

Reading Skill Analysis

In the first analysis, reading skill was treated as a between-subjects variable, while prior knowledge was included as a covariate. There was a main effect of condition ($F(2,124)=3.72$, $MSE=0.268$, $p<0.05$), indicating that SERT participants ($M=0.43$, $SD=0.15$) outperformed Control ($M=0.36$, $SD=0.16$) and Preview ($M=0.36$, $SD=0.12$) participants (see Figure 1). A post-hoc analysis using Least Significant Difference confirmed this trend showing that SERT participants performed better than Preview ($p<0.05$) and Control participants ($p<0.05$); however, Control and Preview did not differ ($p<0.05$). There was also a main effect of reading skill ($F(1,124)=11.08$, $MSE=0.798$, $p<0.05$), indicating that high reading skill students ($M=0.43$, $SD=0.15$) better understood the earthquake passage than did low reading skill students ($M=0.34$, $SD=0.13$).

Our analysis also showed effects for question format and question type, which were independent of training condition. There was a significant effect for question format ($F(1,124)=16.86$, $MSE=0.664$, $p<0.05$) indicating that more multiple-choice questions ($M=0.48$, $SD=0.19$) were answered correctly than open-ended questions ($M=0.29$, $SD=0.15$). There was also a significant main effect for question type ($F(1,124)=6.63$, $MSE=0.286$ $p<0.05$), indicating that students answered more questions based on the text correctly ($M=0.46$, $SD=0.18$) than questions requiring bridging-inferences ($M=0.31$, $SD=0.17$). No other effects were significant.

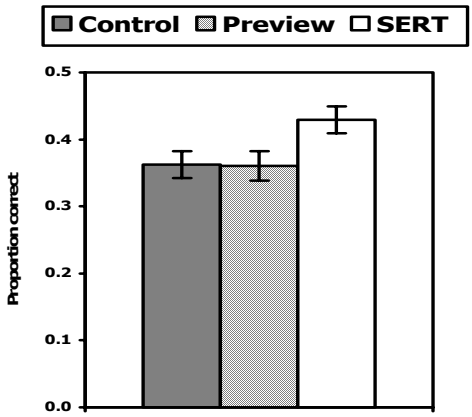


Figure 1. Proportion correct on Earthquakes comprehension test as a function of condition

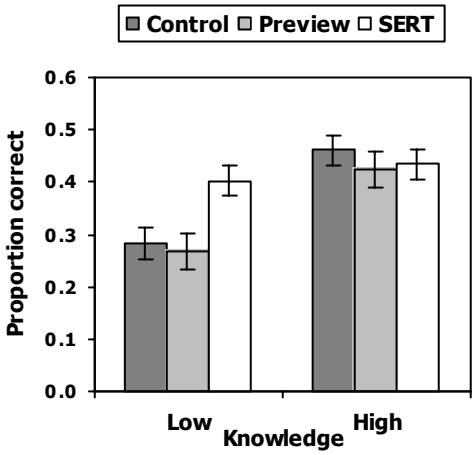


Figure 2. Proportion correct on Earthquakes comprehension test as a function of condition and prior knowledge

Prior Knowledge Analysis

This section examines how the effects of condition depended on students' prior knowledge by including prior knowledge as a between-subjects variable, and reading skill as a covariate. As reported above, the main effects of condition, question format, and question type were reliable There was also a main effect of prior knowledge ($F(1,124)=11.10$, $MSE=0.783$, $p<0.05$), indicating that high-knowledge ($M=0.43$, $SD=0.14$) students better understood the passage than did low-knowledge students ($M=0.34$, $SD=0.13$). This main effect was qualified by a significant interaction between knowledge and condition ($F(2,124)=3.84$, $MSE=0.271$, $p<0.05$). A Post hoc Least Significant Difference

analysis indicated for low-knowledge students, SERT participants (M=0.43, SD=0.15) outperformed Control ($p<0.05$) (M=0.30, SD=0.16) and Preview ($p<0.05$) participants (M=0.30, SD=0.12) (see Figure 2). There was no difference between Control and Preview participants ($p<0.05$). In contrast, the effect of condition was not evident for high-knowledge participants ($F(2,67)=0.41$, MSE=0.036, $p>0.05$).

Discussion

Our findings indicate that SERT training helps students comprehend science texts. Specifically, SERT students performed better on the earthquake comprehension assessment than did Preview or Control students. Moreover, our findings also suggest that SERT training is particularly beneficial to facilitating comprehension among low-knowledge students, as evidenced by the finding that low-knowledge students trained with SERT performed better on the earthquake comprehension test than did low-knowledge students in the Preview and Control conditions.

The finding that the effect of SERT training emerged one week after training is encouraging because it suggests that students remember and use the strategies beyond the time of training. We expect that the lack of training effects on the immediate comprehension test is likely due to a fatigue effect following training.

Overall, our findings replicate those of McNamara (in press), which showed that SERT training facilitated readers' comprehension of scientific texts. Most importantly, both McNamara's study and the present results show that low-knowledge students tend to benefit most from SERT training in terms of comprehension gains. This is a very important finding because it suggests that the SERT method may be an effective technique for supporting comprehension among low-knowledge students most at risk from comprehension problems while reading difficult textbooks.

We predicted that SERT and Previewing would have differential effects on comprehension. We predicted that Preview technique would be less effective for low-knowledge students because using the Preview technique requires students to activate their prior knowledge before reading (i.e., when previewing the titles and subtitles). Thus, we expected Previewing to help high-knowledge students and SERT to benefit low knowledge students. Although we confirmed the latter prediction, we did not find benefits for previewing. This result may be because the students were relatively low knowledge overall. Perhaps the benefits of Previewing depend on a greater availability of prior knowledge than possessed by these inner-city high-school students. Indeed, a majority of students may not have sufficient knowledge to make use of previewing (Snow, 2002).

We expected the use of the SERT method to facilitate comprehension because of the manner in which the student engages with the material during the application of the SERT techniques. Specifically, the reader engages in on-line comprehension monitoring, while at the same time applying techniques known to aid comprehension, such as elaboration and bridging (Chi et al, 1994). This on-line method allows the student to process information effectively, which, in turn, supports comprehension. SERT strategies may provide a scaffold by which the student can integrate new information into their existing knowledge schemas, even in the absence of domain knowledge. Conversely, when using the Previewing method, students do not necessarily engage in on-line comprehension monitoring and reading strategy techniques. Rather, the Preview method encourages readers' to use comprehension monitoring prior to reading. A potential problem with this method is that it does not engage the reader as effectively with the text and the information they are reading.

In terms of understanding the reading process, our study supports the view that the comprehension activities used by the reader play a critical role in reading comprehension (e.g., Pressley et al, 1992). That is, students who have better comprehension skills, as reflected in their use of comprehension monitoring practices and utilization of active reading strategies, are likely to make comprehension gains (Chi et al, 1994).

Of course, to understand the links between reading strategy use (e.g., self-explanations) and comprehension, it is necessary to investigate the associations between strategy use and comprehension. For example, McNamara (in press) investigated the links between styles of self-explanation and performance on a science comprehension test. She found a reliable positive correlation between college students' comprehension scores and their use of logic and common sense self-explanations. In a similar vein, O'Reilly, Sinclair, and McNamara (in press) also found correlations between self reported use of individual SERT strategies and comprehension. However, more research is needed in this area to understand more fully the ways in which students use strategies to facilitate comprehension actually support learning gains.

In conclusion, our findings suggest that SERT training is a useful method of reading strategy training, which can be used to enhance comprehension at the high-school level. Indeed, McNamara and colleagues are currently working on integrating the SERT method of teaching into the high-school classroom. McNamara, Levinstein, and Boonthum (in press) have developed an interactive, automated version of SERT training, designed to be integrated into classrooms. The long-term aim is to develop a SERT intervention which is tailored to the needs and level of the student. In its

current form, the automated trainer, iSTART (Interactive Strategy Trainer for Active Reading and Thinking) teaches students the SERT strategies. The program comprises the three basic components used in the tutor delivered training: introduction, demonstration, and practice. During the practice section, in which students practice using the strategies, the iSTART system assesses the quality of the self-explanations and provides feedback to the student to encourage deep processing of the text (McNamara et al., in press).

Several preliminary examinations of the iSTART system have indicated that the computerized version of SERT training is as effective as live SERT training (O'Reilly, Sinclair, & McNamara, in press). Overall, the present findings and those collected from the iSTART studies indicate that SERT is an effective reading strategy intervention, which can be used in the classroom to help students comprehend difficult textbooks.

Acknowledgements

We would like to thank the members of the ODU Strategy Lab who helped to conduct this study, including Kim Cottrell, Karen Fuller, Erin McSherry, Grant Sinclair, Danny Simmons, and Karen Stockstill. This project was supported by NSF (IERI Award number 0241144). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Beck, I., McKeown, M., & Gromoll, E. (1989). Learning from social studies texts. *Cognition and Instruction*, *6*, 99-158.

Bowen, B. A. (1999). Four puzzles in adult literacy: Reflections on the national adult literacy survey. *Journal of Adolescent and Adult Literacy*, *42*, 314-323.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439-477.

Cornaldi, C. & Oakhill, J. (Eds.), *Reading comprehension difficulties: Processes and Intervention.* Mahwah, NJ: Erlbaum.

Cox, B. (1997). The rediscovery of the active learner in adaptive contexts: A developmental-historical analysis of transfer of training. *Educational Psychologist, 32*, 41-55.

Johnston, P., & Allington, R. (1991). Remediation. In P. D. Pearson, R. Barr, M. L. Kamil, P. Mosenthal (Eds.), *Handbook of reading research* (pp. 984-1012). White Plains, NY: Longman Inc.

Garner, R. (1990). When children and adults do not use learning strategies: Toward a theory of settings. *Review of Educational Psychology*, *60*, 517-529.

Long, D., Oppy, B., & Seely, M. (1994). Individual differences in the time course of inferential processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1456-1470.

McNamara, D. S. (in press). SERT: Self-Explanation Reading Training. *Discourse Processes*.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247-288.

McNamara, D. S., Levinstein, I. B. & Boonthum, C. (in press). iSTART: Interactive Strategy Trainer for Active Reading and Thinking. *Behavioral Research Methods, Instruments, and Computers*.

McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Ogle, D. S. (1986). K-W-L group instructional strategy. In A. S. Palincsar, D. S. Ogle, B. F. Jones, & E. G. Carr (Eds.), *Teaching reading as thinking* (Teleconference Resource Guide, pp. 11-17). Alexandria, VA: Association for Supervision and Curriculum Development.

O'Reilly, T., McNamara, D. S., & Sinclair, G. P. (in press). Reading Strategy Training: Automated versus Live. *Proceedings of the 26th Annual Conference of the Cognitive Science Society. Chicago,Il: Cognitive Science Society*.

Perfetti, C. (1985*). Reading ability*. New York: Oxford University Press.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, *27*, 91-109.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Voss, J. F., & Silfies, L. N. (1996). Learning From History Text: The Interaction of Knowledge and Comprehension Skill With Text Structure. *Cognition and Instruction*, *14*, 45-68.

Reading Strategy Training: Automated Verses Live

Tenaha O'Reilly (t.oreilly@mail.psyc.memphis.edu)
Grant P. Sinclair (gsinclair@mail.psyc.memphis.edu)
Danielle S. McNamara (d.mcnamara@ mail.psyc.memphis.edu)

Psychology Department, University of Memphis
Memphis, TN 38152 USA

Abstract

This study examined the effectiveness of Self-Explanation Reading Training (SERT) and an automated version of this intervention called Interactive Strategy Training for Active Reading and Thinking (iSTART) in improving science text comprehension. College students (N=297) were assigned to one of three conditions: SERT (trained by a human instructor), iSTART (trained by a computer), or no treatment control. Participants read a text on cell mitosis and answered text-based and bridging inference questions. There was a significant overall effect of condition indicating that both iSTART and SERT out performed controls on comprehension. However, this effect was modulated by question type: both SERT and iSTART significantly enhanced comprehension for text-based questions, but the effect was not reliable for bridging inference questions.

Introduction

Many students have difficulty understanding what they read; in particular, many students have trouble comprehending science texts (Bowen, 1999; Snow, 2002). Problems associated with comprehension are augmented by the lack of strategic reading interventions in classrooms: students seldom use high-level comprehension strategies that promote deep comprehension (Cox, 1997; Garner, 1990).

One way to improve comprehension is to teach reading strategies that encourage deeper processing of the text. Interventions that promote deeper processing such as self-explanation and elaborative interrogation have been successful in improving student comprehension (e.g., Chi, De Leeuw, Chiu, & LaVancher, 1994; Pressley, Wood, Woloshyn, Martin, King, & Menke, 1992). For example, McNamara (in press) has reported positive learning gains for her Self–Explanation Reading Training (SERT). SERT is a modified version of the self-explanation learning strategy (e.g., Chi et al., 1994). The SERT training program helps improve comprehension by encouraging students to utilize various sub-strategies to build strong connections between the reader's knowledge and the text. SERT teaches students various reading strategies, including comprehension monitoring, logic and common sense/elaboration, paraphrasing, bridging inferences, and prediction to improve their ability to explain text and understand it at a deeper level.

McNamara (in press) examined the effectiveness of SERT with college students who varied in prior knowledge of science. Half of the participants learned to self-explain and use reading strategies while reading four science texts. The other half of the participants read aloud the texts and answered questions concerning them. After the training phase, the two groups' ability to self-explain was compared. They also answered text-based and bridging-inference questions about the text that they had all self-explained. The results indicated that low-knowledge readers who were trained to use SERT outperformed control participants on measures of text comprehension. However, this advantage only occurred for text-based questions. In addition, protocol analyses of the readers' self-explanation indicated that the low-knowledge readers improved in terms of their ability to paraphrase the text, and more importantly, in their ability to use domain-general knowledge (or logic and common sense) to make sense of the text. Not having the requisite knowledge, they were not able to make inferences requiring domain specific knowledge while reading. Nonetheless, they showed the same level of comprehension on the text-based measures as did the high-knowledge readers, and substantially greater performance than their low-knowledge counterparts. These findings are particularly encouraging because it demonstrates improvement for the students who need the training the most: the low-knowledge students.

The present study compares the effectiveness of SERT and a similar, but automated version of the training, called Interactive Strategy Training for Active Reading and Thinking (iSTART; McNamara, Levinstein, & Boonthum, in press). Automating the core aspects of SERT training has several advantages including self-paced learning and standardized training. iSTART is a computer program that uses automated agents to provide SERT-based training to students. The program, like SERT, has three sections, introduction,

demonstration, and practice. The program has both vicarious and interactive components to enhance learning. The students learn vicariously by watching "agent students" interact and learn strategies taught by a "teacher agent." Later the student interacts with the program, and the system provides feedback on the student's performance.

While automating the SERT intervention has several advantages, one potential problem is that automation may influence the effectiveness of SERT. For example, during live SERT training, students practice with a partner while self-explaining. In iSTART, human peer interaction does not occur. In light of the work on reciprocal teaching (e.g. Palincar & Brown, 1984), removing human peer interaction may diminish the effectiveness of the training. The goal of this study was to examine whether the automated iSTART was as effective at improving comprehension as the live SERT training. Participants were assigned to one of three conditions: live SERT (trained by a human instructor), iSTART (trained by the computer program) and a control condition which had no training (and instead read a text and answered questions concerning it). One week after the training phase, participants read a passage on cell mitosis (see McNamara, 2001; McNamara, in press). The dependent measure was the total proportion of correct answers (both text-based and bridging-inference questions) based on the passage. It was expected that SERT would out perform controls because previous research showed a facilitative effect of SERT on comprehension (McNamara, in press). It was expected that iSTART training would also improve comprehension compared to controls. This prediction was made based on research that has shown the benefits of automated agents on learning (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Du Boulay, 2000; Graesser & Person, 1994).

Finally, we were also interested in uncovering any spontaneous strategies used by control students in our study. Prior research has indicated that the average student is not particularly strategic when learning from text (Cox, 1997; Garner, 1990). After reading the passage on mitosis, participants in all three conditions were asked to indicate what strategies they used to help them understand what they read. We predicted that both the SERT and iSTART conditions would indicate strategies such as those taught in the training session. A higher reported use of SERT strategies in the training condition serves as a manipulation check for whether the trained participants actually used the strategies to comprehend the text. Moreover, our secondary goal was to examine what, if any, strategies would be reported by the untrained participants.

Method

Participants

The sample consisted of 297 biology college students from Old Dominion University. There were 87 males and 210 females and the average age was 21 years old (SD=4.58). The students participated during the laboratory sessions of their Introductory Biology Course. Each lab was randomly assigned to one of the three conditions. The students received extra credit in the course for participation.

Materials

Two sets of individual difference measures were used to gauge students' cognitive ability: reading skill and prior knowledge. Reading skill was measured by Nelson-Denny Reading Skills Test. The test consisted of 38 multiple-choice questions designed to assess comprehension on several short text passages. Prior knowledge was measured by a 54-item multiple-choice test on general science knowledge and the humanities. The test consisted of questions drawn from biology, art, literature, history, geology, political science, and psychology.

Participants in the SERT condition were given a short list of five reading strategies (i.e., comprehension monitoring, paraphrasing, elaboration, prediction, and bridging), a video transcript and note sheet (used during the video segment of training), and a booklet with more detailed descriptions of the strategies and examples of their use in self-explanations. Participants in the iSTART condition were given written instructions on how to use the iSTART system, and a short list of the reading strategies.

A passage on cell mitosis described the sequential stages involved in cell division, and included all the information required to subsequently answer a set of comprehension questions. The text was 650 words in length and had a Flesch Reading Ease 52 and a Flesch-Kincaid Grade Level 9.1. Comprehension was assessed using a set of 12 open-ended questions: six text-based and six bridging-inference questions. The answers to the text-based questions could be found in a single sentence within the passage, while the bridging-inference questions required the reader to integrate information from two or more sentences within the passage. Participants were also given a sheet of paper which asked if they had finished reading the text and which, if any, strategies they used to help them understand the text. Finally, participants were given a 260 word text on thunderstorms with a Flesch Kincaid Grade level of 8.6 and a Flesch Reading Ease 56. For control participants the text was accompanied by 8 open-ended questions.

Design and Procedure

The experiment had three phases: pre-testing, training and post-testing. All participants were given the individual difference measures during the pre-test phase. Participants first took the prior knowledge test (20 minutes) followed by the Nelson-Denny reading skill test (15 minutes). The following day the experimental training and control phases were conducted in a 2-hour session. The post-test phase occurred after a one-week retention period, during which all participants read and answered questions for the mitosis passage. Participants were given 30 minutes to read the passage and answer the questions. Students did not have the text available to help them once they began answering the questions.

Training

SERT: The SERT training session was conducted in a 2-hour session. SERT participants were told that the purpose of the study was to teach them strategies that would help them to better understand and remember what they read. Participants were first provided with a description and examples of self-explanation. The instructor then defined and provided examples for five reading strategies: comprehension monitoring, paraphrasing, elaboration, prediction, and bridging.

Participants then watched a video depicting a student reading and self-explaining a text about forest fires. Participants could refer to the accompanying video transcript during viewing. The video was paused at various points, and participants identified and discussed the strategies being used by the reader in the video.

Finally, the participants worked in pairs to practice self-explanation while reading the thunderstorms text. The participants took turns self-explaining, alternating after each paragraph. At the end of each paragraph, the partner who was listening (and not self-explaining) summarized the student's self-explanation.

iSTART: The iSTART training session was conducted in a 2-hour session. Participants in the iSTART condition were told that the purpose of the study was to teach them strategies that would help them to better understand and remember what they read. Participants were then given instructions on how to use the iSTART system, and they proceeded to go through the three sections of the program: introduction, demonstration, and practice. The practice section involved reading and self-explaining a text about thunderstorms one sentence at a time. Participants typed their self-explanations into the computer.

Control: Participants were told that the purpose of the study was to determine the types of strategies students use when they read. Participants read the thunderstorms text and indicated the strategies they used while reading. The participants answered corresponding questions to assess comprehension.

Testing

One week following training, all three groups of participants were asked to read the science text about cell mitosis (i.e., the low-coherence version used in McNamara, 2001). The participants were asked to use the strategies that they had learned or talked about the previous week. After reading the text, participants were asked to indicate what strategies they used to help them understand the text. They were then given the 12 open-ended questions to assess comprehension. Participants were given 30 minutes to read the text and answer the questions. The text was not available to the students once they began answering the questions.

Results

The effect of training on reported strategy use.

Our first question was whether training condition affected students' reported use of strategies one week after training when reading the cell mitosis text. Table 1 lists the percent of participants who indicated using SERT strategies, while Table 2 indicates the percent of participants who reported using non-SERT strategies. Students' self-reports of strategy use during reading were tabulated into 17 categories. The categories were devised based on a combination of a priori strategies reported in the literature, and strategies that were frequently mentioned in the students' responses. It is important to note that category membership is not mutually exclusive. That is, a participant could have listed more than one strategy, and therefore the percentages per condition will not sum to 100%.

Strategy	Control	iSTART	SERT
Bridging	0%	39.8%	38.5%
Prior Knowledge	7.3%	15.1%	17.7%
Elaboration	0%	17.2%	21.9%
Prediction	0%	14%	13.5%
Self-explanation	0.9%	23.7%	19.8%
Paraphrase	0%	46.2%	46.9%

Table 1 Percent of self-reported SERT/iSTART strategies by condition

Further calculations revealed that 77.2% of iSTART and 72.9% SERT participants reported using at least one of the reading strategies taught by the iSTART/SERT, whereas only 8.3% of control participants reported using these strategies ($\chi(2,297)=123.30$, $p<0.05$). A chi-square also revealed that more control participants (58.9%) indicated using more non-SERT strategies (i.e., 10 strategies listed in table 1 other than the six SERT strategies) than iSTART (10.9%) or SERT (30.25%) participants ($\chi(2,297)=60.79$, $p<0.05$). Although this measure does

not indicate whether they actually *used* these strategies (nor whether their self reports were all inclusive, i.e., they may not have reported some strategies) it does reveal that iSTART and SERT participants become sensitive to the notion of active reading strategies through training, and that this was retained one week later.

Strategy	Control	iSTART	SERT
Imagery	7.3%	0%	4.2%
Re-Read to understand	11.9%	1.1%	1.0%
Summarize	1.8%	0%	2.1%
Mnemonics	6.4%	1.1%	4.2%
Skim text	4.6%	2.2%	2,1%
Note taking	5.5%	0%	1.0%
Memorization	3.7%	0%	2.1%
Repetition	30.3%	8.6%	25%
Focus	9.2%	0%	2.1%
Key points	14.7%	1.1%	7.3%
No Strategy	16.5%	4.3%	1.0%

Table 2 Percent of self-reported non-SERT/iSTART strategies by condition.

The effect of strategy training on comprehension.

Our second question was whether SERT and iSTART training successfully improved comprehension for those students who reported using the strategies. All control participants (N=109) were included in this analysis. However, in the iSTART and SERT conditions, only participants who explicitly stated that they used one or more of the SERT strategies (comprehension monitoring, self-explanation, paraphrasing, prediction, bridging, elaboration, and prior knowledge use) were included in the analysis. We restricted this analysis to training participants who reported using the strategies after training because our question regarded the effects of training for those participants who attempted to use the SERT strategies to read and understand the cell mitosis text. Therefore, this manipulation check reduced the number of participants in the iSTART condition from N=92 to N=71, and reduced the number of participants in the SERT condition from N=96 to N=70. While it is interesting in itself that about 25% of the training participants did not attempt to use the reading strategies after training, there are many reasons why they might not (e.g., lack of motivation in the laboratory setting, lack of sufficient learning of the strategies, preference for other strategies). However, we cannot identify these reasons and those participants are not the focus of this particular analysis. The exclusion of participants who did not report any SERT strategies is a conservative effort. The self-report measure includes both participants who actually used the SERT strategies and those who said they used them, but did not use them in practice. In a similar vein, by retaining all the control participants, the control condition has an advantage because the analysis includes the participants who may use higher-level strategies and who may, therefore, be expected to score as well as the trained participants. A similar set of analyses which also excluded control participants who did not report any strategies produced a similar pattern of results as the analysis used here that did not exclude any control participants. Hence, the analyses reported below included all control participants

Differences in pre-test abilities

A one-way between-participants ANOVA was conducted on the student's pre-test level of prior knowledge to determine whether the groups differed as a function of pre-treatment knowledge. The results indicated that there were no differences between conditions, $F(2,233)= 1.47$, $MSE= 72.25$, $p>.05$, indicating that any difference between the conditions are unlikely due to pre-treatment levels of knowledge. Likewise, a between-participants analysis was conducted on the students' reading skill scores to determine whether the pre-treatment reading skill differed as a function of condition. The analysis revealed that there was no significant effect of pre-test reading skill, $F(2,233)= 2.05$, $MSE= 71.31$, $p>.05$, and thus, differences among the conditions are unlikely due to pre-treatment differences in reading skill.

Effects of condition

A repeated measures analysis of variance was conducted on comprehension scores including the within-participants variable of question type (text-based, bridging inference) and the between-participants variables of training condition and knowledge with reading skill as a covariate. There was a significant effect of question type, $F(1,229)= 23.57$, $MSE= 0.602$, $p<.05$, indicating that more text-based questions ($M=0.52$, $SD=0.26$) were answered correctly than bridging questions ($M=0.22$, $SD=0.17$). There was also an effect of knowledge, $F(1,229)= 35.14$, $MSE= 1.85$, $p<.05$) indicating that high-knowledge students ($M=0.28$, $SD=0.15$) scored higher than low-knowledge students ($M=0.45$, $SD=0.17$). The analyses revealed a significant effect for training condition, $F(2,229)= 3.94$, $MSE= 0.207$, $p<.05$) indicating that both iSTART ($M=0.39$, $SD=0.21$) and SERT ($M=0.39$, $SD=0.19$) participants scored higher than controls ($M=0.33$, $SD=0.22$). This effect was qualified by a significant interaction of question type and condition, $F(2,229)= 2.98$, $MSE= 0.076$, $p<.05$). Post hoc Least Significant difference tests revealed that SERT ($M=0.55$, $SD=0.23$, $p<.05$) and iSTART ($M=0.54$, $SD=0.24$, $p<.05$) participants scored higher than controls ($M=0.45$, $SD=0.27$) on text-base questions, but not bridging

questions (see Figure 1). No other effects were significant.

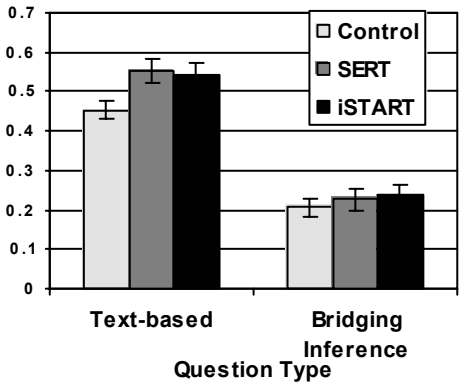


Figure 1. Proportion correct on the mitosis passage score as a function of condition and question type

To examine effects of reading skill, a second repeated measures analysis of variance was conducted on comprehension including the within-participants variable of question type (text-based, bridging-inference) and the between-participants variables of training condition and reading skill, with knowledge as a covariate. The analysis revealed a significant effect for question type, $F(1,229)=5.07$, $MSE=0.147$ $p<.05$, indicating that more text-based questions ($M=0.52$, $SD=0.26$) were answered correctly than bridging questions ($M=0.23$, $SD=0.17$). The main effect of training condition was reliable, $F(2,229)=4.69$, $MSE=0.239$ $p<.05$, indicting that iSTART ($M=0.39$, $SD=0.21$) and SERT ($M=0.41$, $SD=0.19$) participants scored higher than control participants ($M=0.33$, $SD=0.22$). This main effect was qualified by a significant interaction with question type, $F(2,229)=2.98$, $MSE=0.076$ $p<.05$. Post hoc Least Significant Difference tests revealed that iSTART ($M=0.54$, $SD=0.24$, $p<.05$) and SERT ($M=0.58$, $SD=0.23$, $p<.05$) participants scored higher on text-based questions compared to control participants ($M=0.46$, $SD=0.27$). But this effect was not found for bridging-inference questions. In sum, both SERT and iSTART improved students' comprehension compared to controls, particularly at the textbase level of understanding.

Finally, correlations between each of the 18 self-reported strategies and the total proportion correct on the mitosis passage. Six strategies were significantly correlated to comprehension. Five of the strategies were taught by the SERT/iSTART technique bridging, $r=.25$, self-explanation, $r=.19$, elaboration, $r= .14$, paraphrasing, $r=. .25$, predictions, $r=.12$, and one non-SERT/iSTART strategy, reread to understand, $r=.15$.

Discussion

The results of the present study are congruent with research demonstrating beneficial effects of reading-strategy training on understanding and learning (Chi et al., 1994; Pressley et al., 1992). First, the majority of SERT and iSTART participants reported the use of SERT strategies such as elaboration, using prior knowledge and making bridging inferences. Research has shown that high-level strategies such as prior knowledge use (Spilich, Vesonder, Chiesi, & Voss, 1979) and elaboration (Pressley et al., 1992) are much more effective than low-level strategies such as repetition. The current results suggest that both SERT and iSTART training encourage the use of higher-level strategies during reading, and that the self reported use of these strategies correlates with comprehension. The present findings also seem to support the views of Cox (1997) and Garner (1990): average untrained students are spontaneously unlikely to use higher-level strategies to help them better understand what they read.

In a related study, Best, Ozuru, and McNamara (2004) analyzed the content of students' self-explanations while interacting with iSTART. The researchers found that several of the participants indicated the use of high quality elaborations including logic/common sense and scientific reasoning. Many of these elaborations were knowledge building, which helps the reader more effectively explain the current sentence. However, the quality of the elaborations depended upon both the sentence difficulty and individual differences. In short, iSTART seems to promote the use of both high-level and high-quality comprehension strategies.

Second, and more importantly, the self-report data is bolstered by the findings from the comprehension data. Participants in both the SERT and iSTART conditions answered more text-based questions correctly than did control participants. Hence, the current study suggests that SERT and iSTART training encouraged many of the learners to use higher-level strategies, and when they did, comprehension for text-based information was facilitated. These findings are congruent with results reported by McNamara (in press) showing positive learning gains for students who were given SERT training. In that study, McNamara (in press) found evidence that SERT helped participants by encouraging them to use logic, common sense and general world knowledge. Moreover, the beneficial effects of training were most prominent for low-knowledge readers; that is, for the readers who need the training the most.

As in McNamara (in press), the facilitative effect of training in the current study did not extend to bridging-inference questions. One possible explanation is that participants did not have the specific domain knowledge required to make effective bridging inferences (cf., McNamara, in press). This interpretation is in accordance with the finding that

prior knowledge is important in generating inferences (e.g., Singer & Ritchot, 1996), particularly when text cohesion (the degree to which relations are made explicit) is low (McNamara, 2001). The mitosis text used here was taken from McNamara (2001), who manipulated text cohesion as an independent variable. The current study utilized the low-cohesion version of the mitosis text. Because text cohesion was low, readers require a greater degree of specific domain knowledge to generate the necessary inferences.

As found in previous studies (e.g., Anderson et al., 1995; Du Boulay, 2000; Graesser & Person, 1994), this research also confirms the effectiveness of computerized training, particularly those using automated agents as tutors. The benefits of automated tutors include self-paced learning, standardized training, and feedback tailored to the individual's progress. The results of the present work support the effectiveness of an automated tutoring system by showing that a computerized presentation of the SERT strategies (iSTART) was as effective as a presentation of the strategies by a human instructor. Given the current trend towards increasing classroom size and cutbacks in educational funding, this result suggests that automated trainers may provide a means for reducing the load on resource-strained educators.

In sum, this study adds to the literature by demonstrating that SERT and iSTART training increase the reported use of higher-level strategies during reading, and when used, the SERT and iSTART strategies enhance comprehension, at least for text-based information. More encouraging is the finding that the self-paced computer version of the training is as effective as the human delivered training.

Acknowledgements

This project was supported the NSF (Award number: 0241144). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Anderson, J.R., Corbett, A. Koedinger, K. & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4,* 167-207.

Beck, I., McKeown, M., & Gromoll, E. (1989). Learning from social studies texts. *Cognition and Instruction,* 6, 99-158.

Best, R., Ozuru, Y., & McNamara, D.S. (2004). Self-explaining Science Texts: Strategies, Knowledge and Reading Skill. *Proceedings of the Sixth International Conference of the Learning Sciences,* Monica, CA.

Bowen, B. A. (1999). Four puzzles in adult literacy: Reflections on the national adult literacy survey. *Journal of Adolescent and Adult Literacy, 42*, 314-323.

Chi, M. T. H., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Cornoldi, C., & Oakhill, J. (1996). In C. Cornoldi and J. Oakhill (Eds.), *Reading comprehension difficulties*, New Jersey: Lawrence Erlbaum Associates, Publishers.

Cox, B. (1997). The rediscovery of the active learner in adaptive contexts: A developmental-historical analysis of transfer of training. *Educational Psychologist,* 32, 41-55.

Du Boulay, B. (2000). Can we learn from ITS? International Journal of *Artifical Intelligence in Education, 11*, 1040-1049.

Garner, R. (1990). When children and adults do not use learning strategies: Toward a theory of settings. *Review of Educational Psychology*, 60, 517-529.

Graesser, A. & Person, N (1994). Question asking during tutoring. *American Educational Research Journal, 31,* 104-137.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51-62.

McNamara, D.S. (in press). SERT: Self-Explanation Reading Training. *Text and Discourse*.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2003). iSTART: Interactive Strategy Trainer for Active Reading and Thinking. *Submitted to Behavioral Research Methods, Instruments, and Computers.*

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction, 2*, 117-175.

Perfetti, C. (1985). Reading ability. New York: Oxford University Press.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist, 27* (1), 91-109.

Singer, M., & Ritchot, K. (1996). The role of working memory capacity and knowledge access in text inference processing. *Memory & Cognition, 24*, 733-743.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND

Spilich, G., Vesonder, G., Chiesi, H., & Voss, J. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 275-290.

Case Interpretation and Application In Support of Scientific Reasoning

Jakita N. Owensby, jowensby@cc.gatech.edu
Georgia Institute of Technology, College of Computing
801 Atlantic Drive Atlanta, GA 30332-0280

Janet L. Kolodner, jlk@cc.gatech.edu
Georgia Institute of Technology, College of Computing
801 Atlantic Drive Atlanta, GA 30332-0280

Abstract

Scientific reasoning involves the use of scientific skills, practices, and domain knowledge to solve science problems. A little emphasized tool that experts use to help them reason is to refer back to previous problem solving experiences, interpreting and applying those experiences as they solve problems. Results from a pilot study conducted Fall Semester 2002 suggest that improvement in interpreting and applying expert cases to solve a problem may also lead to improvement in certain scientific reasoning skills. In this paper, we seek to explore the connection between case application and scientific reasoning skills, namely, using evidence to justify a claim, generating hypotheses, making predictions, and explaining scientific phenomena.

Introduction

Scientific reasoning involves the use of scientific skills, practices and domain knowledge to solve science problems. Much research has been done to understand how students can develop more expert-like scientific reasoning skills (e.g. Kuhn, 1993; Schauble et. al., 1995), and much research has been done to promote more expert use of scientific reasoning skills in educational settings (e.g. Bell & Davis, 2000; Reiser, et. al., 2000). However, little attention has been given to the role case interpretation and application might play in learning to reason scientifically. There is evidence that scientists use cases extensively in their reasoning. For example, when trying to analyze a series of unexpected results, scientists will refer to cases that may seem unrelated but that have similar in order to explain why those unexpected results may have occurred (Blanchette & Dunbar, 2001).

In educational settings, it is often difficult to support students as they attempt to acquire and carry out expert-reasoning processes. In many cases, the expert-reasoning process may be too complex to pare down in such a way that students can engage in it without getting lost in all of the complexity (Reiser, 2000). In other cases, because the expert-reasoning process is not fully understood, it becomes difficult to assess where students may experience difficulty, and when they do, it is difficult to know what kind of help to provide.

We have sought to address these difficulties for one complex skill: case application. Fall semester 2002, we conducted a study to understand the effectiveness of the Case Application Suite (CAS) (Owensby & Kolodner, 2004), a set of tools designed to support middle-school students in project-based inquiry classrooms as they interpret and apply the experiences of experts to solve design problems. In particular, we were interested in understanding how effective our system of scaffolds was at supporting students as they interpreted and applied expert cases, whether the distribution of scaffolding responsibilities across teacher and software was effective, how well students were able to use case application skills in the absence of the scaffolding, and whether the distribution of scaffolding responsibilities could be articulated in a cognitive apprenticeship (Collins, Brown & Newman, 1989) framework.

Analysis of the data showed that CAS was effective at supporting students in case application, showing significant differences for interpretation and trends for application. In addition, the trends in the data suggested an unexpected finding—that case application supports the learning of scientific reasoning skills. Our analysis of this phenomenon suggests that this is because case application and scientific reasoning share foundational skills, namely using evidence to support a claim, generating hypotheses, making predictions, and explaining phenomena scientifically. This paper seeks to explore the connection between case application and scientific reasoning skills to suggest that improvement of certain case application skills will promote improvement in these aspects of scientific reasoning. As part of our exploration, we will show how we've used software-realized scaffolding (Guzdial, 1994) to support the acquisition of case application skills among middle-school students in project-based inquiry science (Blumenfeld, et al., 1991).

Case Application and Scientific Reasoning

Case application is the process of interpreting, analyzing, and applying experiences to address challenges or solve problems (Owensby & Kolodner, 2003; the CBR literature, e.g., Kolodner, 1993). It involves three high-level steps: interpretation, application, and assessment. Interpretation involves, at the time of encountering the case, it, focusing on extracting the connections between its criteria and constraints and the solution chosen to address its challenge, making connections between the solution chosen and the outcomes that happened, and identifying what can be learned from the experience, and at the time of working toward applying it, making connections between the case (acting as a source case) and the new situation (target). Application involves applying those lessons to the new situation or target case, either directly or via adaptation. Assessment involves analyzing the applicability and quality of the proposed solution either by making predictions about the target case's solution or by testing the target case's solution and analyzing the outcomes that result.

Case application is integral to the practices of experts. Medical experts use cases to diagnose as well as to refine treatments for patients. Architects keep file cabinets of cases to go back to when working on new projects. Lawyers refer to previous cases and decisions when constructing a strategy to prosecute someone or to defend a client.

Analogical reasoning has long been recognized as an important aspect of scientific reasoning (e.g. Gentner, 1999; Anderson, 2000, Blanchette & Dunbar, 2001). Case application extends standard analogical reasoning. In addition to mapping the solution for one problem onto the solution for another problem, we include in case application the analysis and interpretation of a case at the time it is encountered that allows its application. We also include in case application the identification of those nuggets of an encountered case that might apply in a new situation. When the cases being used are those of others, this interpretation process involves significant reading for understanding. While reading is taught in schools, rarely does science class focus on helping learners read. Yet real science practice is impossible without the skills involved in reading a scientific case for understanding and reasoning through its application.

Understanding requires identifying claims, the evidence used to support its claims, and the quality of explanations put forth, while applying what is in a science case requires making predictions based on those claims and finding particularly useful information in a big document.

In order to use evidence to support a claim, one must interpret the experience from which the claim arose in such a way that he/she recognizes that the evidence applies. Then, one must interpret the evidence in such a way that the aspects that apply to the claim can be identified. Next, one must be able to articulate how the relevant aspects of the evidence support the claim and make predictions for future use of the concept, skill, or claim. Understanding the experience from which the claim and evidence put forth involves interpreting the experience and drawing out the lessons that can be learned from the experience. Articulating how the evidence supports the claim involves articulating the lessons learned from the evidence and the experience that the claim rises from and then applying those lessons to explain how the evidence supports the claim and then making predictions about how the claim might be useful in the future. It does make sense, then, that supporting students as they learn how to interpret and apply cases illustrating the evidence of scientific phenomena and the application of scientific principles could help those same students become better scientific reasoners.

Our approach to supporting the development of case application skills

To help middle schoolers interpret cases and apply them in new situations, we have designed a suite of software tools called the Case Application Suite (CAS) to play the role of coach within a cognitive apprenticeship framework (Collins, Newman & Brown, 1989). In a cognitive apprenticeship approach to learning complex skills, the teacher models the skills and explains his/her reasoning to the students and then coaches and hints as students begin to carry out parts of that reasoning. As students become more capable, they, in turn, model for their peers and coach them to their next levels of capability. But when students work in small groups, there may not always be a group member expert enough to be able to apply that coaching to the rest of the group. CAS supports students as they work in small groups by asking the kinds of questions and making the kinds of suggestions that a teacher or more able student might make if he/she were available.

The design of CAS was informed by suggestions made by the skills acquisition, case-based reasoning, transfer, and cognitive apprenticeship literatures (Anderson, et. al, 1981; Anderson, 2000; Kolodner 1993; Branford, Brown & Cocking, 1999; Collins, Brown & Newman, 1989, respectively). CAS contains three tools. The Case Interpretation Tool helps students identify problems the experts encountered in achieving their goals, solutions they attempted and why they chose those, criteria and constraints that informed those solutions, results they accomplished and explanations of those, and any lessons learned, or rules of thumb, that can be extracted from the experience. The Case Application Tool guides students through attempting to apply the rules of thumb gleaned from the case, prompting them to consider whether a rule of thumb is applicable and then helping them explore ways they can apply it to their solution. The Solution Assessment Tool helps students make predictions about the success of their solution, analyzing the impacts they expect their solution to make as well as where they expect their solution to fall short.

The system of scaffolds in CAS includes five different types of scaffolds: (1) the structure of the suite serves as a scaffold as each tool corresponds to a major step in the case application process; (2) the prompts in each tool's center frame focus students' attention on important aspects of the case; (3) hints are provided with each prompt to give more specific help; (4) examples are provided with each prompt to help students see what they need to be accomplishing; and (5) charts and templates serve as organizers to help students with creating an analyzing the applicability of the rules of thumb they have gleaned.

Each tool is divided into three frames (Owensby & Kolodner, 2003; Owensby & Kolodner, 2004). In the left frame is the expert case and interpretations that have already been done of it. The middle-frame shows the prompts for the tool the group is currently working on. The right frame shows hints and examples (Figure 1).

Use of CAS in the Classroom

We've tried CAS out in classrooms engaging in the Learning by Design (LBD; Kolodner et al., 2003) project-based inquiry unit called *Tunneling Through Georgia*. In this challenge, student teams serve as consultants for the design of several tunnels needed for a transportation system that will run across the state of Georgia. Four tunnels need to be designed, each for a different geological area of the state—mountainous, sandy, and so on. Students need to address several issues—at what depth to dig the tunnel, what methods to use for the digging, and what support systems are needed in the tunnel's infrastructure. Cases are used extensively in the unit to suggest which geological characteristics of the tunnel location they need to learn more about to address the challenge, to

introduce students to different kinds of tunneling technologies, and to give them an appreciation of the complexity of tunnel design. For example, the story about the design and construction of the Lotchberg Tunnel in Switzerland, shows

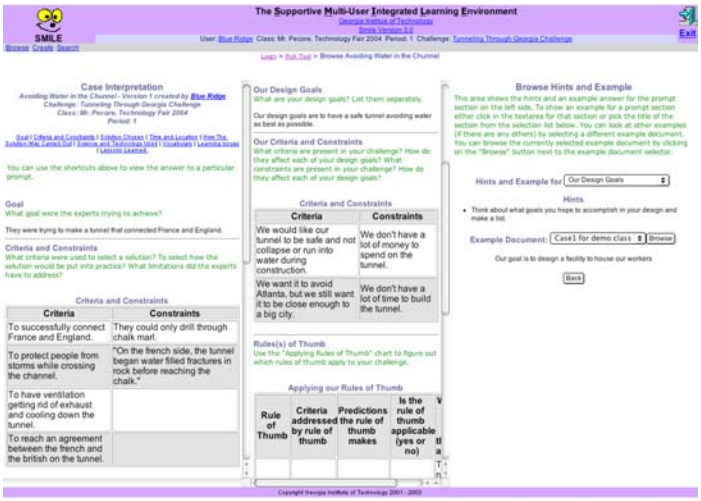


Figure 1: Case Application Tool

some of the problems the experts faced trying to tunnel through the summit of a mountain that has two peaks separated by a river and suggests understanding the composition of a mountain by using test shafts and core sampling can help to identify and possibly avoid problems like crumbling rock, flooding, and cave-ins.

The Tunneling unit is preceded by another unit that requires case application. In that unit, students learn about earth's surface processes as they engage in the challenge of designing and constructing (in a stream table) a way of managing erosion around a basketball court. They read two cases during this unit, one about the dustbowl and another about landslides on the U.S.'s West Coast. The teacher helps them read and understand the cases together as a class and moves around the room coaching them as they work in small groups to apply what they've learned to their challenge. In addition, students use a template to keep track of important aspects of the cases they are reading about. The template, created by the teacher and based on the My Case Summary Design Diary page (Puntambekar & Kolodner, 1998), organizes a page into columns representing *Case Summary, Problems, Ideas, Learning Issues, and Questions*.

As they get started with the Tunneling unit, the teacher again models case application for students as they analyze the Lotschberg Case together as a class. After analyzing the Lotschberg Case, student groups are assigned one of four tunnel cases to interpret and present to the rest of the class. They are introduced to CAS's Case Interpretation Tool to support them as they interpret the case on their own in small groups. This is followed by presentation of their interpretations of their cases to the class and discussion of the lessons that can be pulled from them. When it is time to apply what's been learned from the cases to their own tunnel challenge, students use the Case Application Tool to create a solution. This sequence is repeated a second time as groups read another set of four cases. Later, they sometimes use the Solution Assessment Tool to make predictions about how well their proposed designs might work, what they might have overlooked, and what they would do differently if given another chance

Our Study

We were interested in learning how to help students learn to interpret and apply cases to project challenges and in understanding the effects of adding software designed to augment the teacher's modeling and coaching to a cognitive apprenticeship. Our study collected data to answer three questions: (1) How are students' abilities to interpret and apply cases to their project challenge affected by such scaffolding? (2) To what extent would students' ability to apply cases in the absence of the suite be influenced by its use during a project? (3) To what extent does the suite enable students to articulate the processes involved in case application? When we noticed that some students' scientific reasoning capabilities were improved, we also analyzed to answer a fourth question: To what extent does case application capability predict scientific reasoning capability?

Methods

Procedures

We report here on a study where we used CAS in the classrooms of an 8th grade teacher (Mrs. K) (Owensby & Kolodner, 2003) who had only 4 computers available for her class. Because of this, only a subset of the students were able to use the software; the rest engaged in all of the same activities but had available only the template as scaffolding as they were interpreting and applying cases. All students in the study engaged together in solving the erosion challenge and in doing *Tunneling Through Georgia* activities, and all were exposed to the same teacher modeling. Overall, students engaged in case interpretation and application activities five times – twice during the erosion challenge, once with the teacher at the beginning of the Tunneling unit (the Lotchberg Case), and twice more in small groups. Each time, groups work together to interpret a case and draw out the lessons it teaches; they present their case interpretations to the class, and they lead discussion about their case. Comparison students (n=33 students; 9 groups) used the template to scaffold their case interpretation and application as they interpreted and applied cases after the Lotschberg Case, while experimental students used CAS (n=14 students; 4 groups). We compared the capabilities of students who had the software available to those who did not as students engaged in the unit and after its completion.

Software groups were videotaped as they used the software, and software and non-software groups were videotaped as they presented their interpretations to the class. In addition, templates and logs of CAS use were collected for analysis.

At the end of the unit, a performance assessment was given. Called the Bald Head Island Challenge, students worked in their *Tunneling Through Georgia* groups to make recommendations about the design of two subdivisions on an island off the coast of Georgia. They read a case about Bald

Head Island and used it to give advice. They were asked to identify the risks involved with the project, identify possible management methods, create rules of thumb (Part 1), design a plan for designing and constructing the subdivisions, and make final recommendations about whether the project should move forward with the given time and budget constraints (Part 2). Groups were videotaped as they discussed their ideas. All groups had only template scaffolding available as they engaged in this activity, organized into columns representing *Risk, Why Is This A Risk, Ways To Manage The Risk, Pros, and Cons*.

Analysis

Video data was analyzed using a coding scheme that described the data for specific interpretation and application dimensions. Two coders analyzed video-recorded group performance for interpretation on dimensions shown in Table 1 and for application and assessment on dimensions shown in Table 2, treating each of the two parts of the performance assessment as an episode. A five-point Likert scale was used for each, with one representing no evidence of presence of the quality being rated and 5 representing that the group fully displayed the quality being rated. Differences in ratings were negotiated by discussion, and inter-rater reliability was calculated.

Results

The results that follow provide evidence that case application can be supported in educational settings despite its difficulties, that distribution of scaffolding responsibilities across teacher and software in a cognitive apprenticeship framework seems to be a viable approach for promoting case application, and that particular scientific reasoning skills among students who used the software tools seem to be more sophisticated. We first discuss the differences between students who used the software and those who did not as they were engaging in classroom activities of the Tunneling challenge. We then discuss student capabilities while engaging in the performance assessment, completed by all students after the Tunneling unit was completed and without software scaffolding. The results are discussed with respect to using evidence to justify a claim, generating hypotheses, making predictions, and explaining scientific phenomena.

Case Application During Class Activities

Examination of student artifacts and presentations of case interpretations for groups using CAS vs. the case study template showed three major differences. First, the software groups better identified the reasons for positive and negative outcomes. For example, in learning about the Queens Midtown Tunnel, one software group told us: "They wanted to build [the tunnel] straight [through the city] but couldn't, so they continued it further underground in an S-shape under First Avenue and they took different core samples". This group was specific about the goals of the experts, the constraints that kept them from achieving those goals if they tried the obvious solution, what they did instead, and the activities they had to engage in to do that successfully. The typical non-software group, on the other hand, provided general descriptions about the experts' goals, neither mentioning the constraints' impact on the outcomes nor alternatives. For example, one non-software group told us: "The Manhattan side [was] on a large bluff higher than Queens[, so they] continued tunnels underground in a slope under First [Avenue]."

Second, the groups who used the software included more sophisticated causality in their rules of thumb. For example, the non-software groups' rules of thumb are in the form of simple imperative statements (e.g., "Control water problem", "Take core samples"), while the software groups' rules of thumb explain why (e.g., "Take core samples—they can save your life because if you hit the wrong kind of rock, you can get hurt", "You should always have an oxygen pass so the toxic fumes can get out."

Case Application at Completion of the Unit

In the performance assessment, groups discussed their answers in preparation for writing individual recommendations. We analyzed the video for interpretation and application capabilities.

Table 1 shows results for case interpretation (reliability 89%). Software groups tended to be better at all case interpretation capabilities and significantly better at specifying expert problems, identifying relevant aspects of the case to apply, and using the case to understand the context in which the risks/problems arose.

Table 1. Performance Assessment Results for Part 1 - Interpretation

Coding Characteristic (**bold** denotes significant difference, $p<0.05$)	Software group	Standard Dev. (software group	Non-Software group	Standard Dev. (non-software group)
Recognizes that the case should be used to solve the challenge	**3.88**	**0.25**	**2.66**	**0.71**
Makes direct reference to the case to justify an argument or position	3.135	0.25	2.33	0.82
Able to identify expert problems	**3.00**	**0.00**	**2.42**	**0.61**
Able to identify expert "mistakes"	2.63	0.75	2.00	1.17
Able to identify relevant aspects of the case that can be applied to the challenge	**3.88**	**0.25**	**1.83**	**0.98**
Identifies risks based on prior experience with another LBD/software case	1.88	1.44	1.33	0.41
Able to identify criteria and constraints	3.38	1.11	1.58	0.66
Uses the case to understand the context of the risks	**2.88**	**0.25**	**1.67**	**0.61**
Identifies rules of thumb	1.00	0.00	1.00	0.00

Software groups tended to describe expert problems on a finer-grained level than non-software groups

(3.00 vs. 2.42, $p<0.05$). For example, non-software groups identified "sand" as a risk, while software groups identified the "incompatibility of the old sand and the beach with the new sand dug when the channel was deepened" as a risk, or expert problem. In the case, there are a number of risks or problems that involve sand, so being able to distinguish between those problems is important.

Software groups tended to discuss whether a management method made sense for their challenge, analyzing how the management method would play out in their challenge and questioning each other about the feasibility of a proposed management method (3.88 vs. 1.83). Non-software groups tended to discuss management methods only if they were different from what they expected.

Software groups tended to use the case not only to identify the problems the experts encountered, but also to understand the context in which those problems arose (2.88 vs. 1.67, $p<0.05$). They sought to understand what was happening in the environment that caused the problems to occur or to grow worse. Non-software groups tended to look for keywords that they were familiar with when identifying problems and management methods. For example, while flipping through the case, one non-software student declared , "Oh!! I see erosion here—erosion is a problem." In a similar incident in which one software group member stated that erosion was a problem, another member of that group declared, "but it says here that the problem is the shoreline eroding." This discussion resulted in the software group providing more detail about the erosion problem. In addition, for interpretation, we looked specifically at how well software students used evidence (the case) to justify a claim, and found that software students tended to do a better job than non-software students.

Table 2. Performance Assessment Results for Part 2 – Application and Assessment

Coding Characteristic (**bold** denotes significant difference, $p<0.05$)	Software group	Stadard Dev. (software group)	Non-Software group	Standard Dev. (non-software group)
Identifies issues or problems not explicitly stated in the case	2.88	0.25	2.00	1.02
Able to identify relevant aspects that can be applied to the challenge	2.50	1.08	1.67	1.03
Suggests incorporating a solution found in the case	2.50	0.58	1.92	0.92
Notices that a management method used by the experts cannot be applied as is but must be adapted	1.63	0.58	2.08	0.88
Notices that a solution used by the experts cannot be applied as is but must be adapted	2.38	0.95	2.33	0.68
Justifies use, modification, or abandonment of an expert solution based on criteria and constraints of group's challenge	2.75	0.25	2.25	0.76
Applies a solution used by the experts directly to their challenge	1.75	1.03	1.33	0.82
Suggests that an expert solution should be abandoned	1.25	0.50	1.25	0.61
Applies the case to the challenge using rules of thumb	1.00	0.00	1.00	0.00

For the video-recorded data for Part 2, application and assessment, reliability was 86% and results show trends toward better performance by software groups on several dimensions. First, software groups tended to suggest that a solution from the case would be good to incorporate into their challenge solution (2.50 vs. 1.92). This seems to result from the fact that software groups tended to refer back to the risks and solutions they identified in the expert case in Part 1. They would discuss those solutions to figure out whether they made sense to use in their challenge solution.

Second, software groups tended to justify the use, modification, or abandonment of an expert solution based on the criteria and constraints of the group's challenge (2.75 vs. 2.25). For example, one software group member suggested that the group build a sea wall out of an expensive material. His fellow group member pointed out that that particular material would be very expensive and given that they only had 2 million dollars to work with, they should consider another material. Few non-software groups even mentioned criteria and constraints when deciding whether an expert solution or management method should be used. Again, justification of a claim using evidence was analyzed directly and software groups showed better performance than non-software groups.

Discussion

The goals of this paper are two-fold: (1) to show that through repeated use of scaffolding that supports case interpretation and application students do indeed become better users of cases and (2) to point out the connection between interpretation and application of expert cases and scientific reasoning. The first is shown in the data that has been reported. The second can be seen by connecting what students did while interpreting and applying cases to scientific reasoning.

It seems that using evidence to support a claim and explaining scientific phenomena is important in both case application and scientific reasoning, while

analysis of the data suggests that certain case application skills (i.e. understanding the context of problems, understanding criteria/ constraints, identifying relevant aspects of the case to apply) may be important in generating hypotheses and making predictions. For example, understanding the connection between addressing criteria/constraints and the outcomes that result seems to involve the same reasoning as generating a hypothesis and analyzing the results to determine whether the hypothesis is supported or rejected. This seems to suggest several things:

1. Understanding how to better support case application may lead to understanding how to better support certain scientific reasoning skills.
2. Students can be supported in case application despite its complexity, and students can improve case application skills. As such, support that leads to improvement in case application skills may also lead to improvement in certain scientific reasoning skills.
3. Using a cognitive apprenticeship framework and distributing scaffolding responsibilities across teacher and software seems to be effective at supporting case application skills that seem to be connected to certain scientific reasoning skills. As such, this same approach may be useful in supporting other scientific reasoning skills.

To make these suggestions stronger or to make stronger claims about the connection between case application and certain scientific reasoning skills, the data would need to be coded using dimensions to describe more specifically what is happening with students' scientific reasoning skills as their case application skills are improving. Though this was not the focus of this study, the trends that emerged and the suggestions that arose certainly suggest that this connection between case application and scientific reasoning is worthy of further exploration.

Acknowledgements

We would like to thank the National Science Foundation and the National Physical Science Consortium for their support in this research effort.

References

Anderson, J.R., Greeno, J.G., Kline, P.K. & Neves, D.M. (1981). Acquisition of problem solving skill. In J.R. Anderson (Ed.) *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Anderson, J.R. (2000). *Cognitive Psychology and Its Implications: Fifth Edition*. New York: Worth Publishing.

Bell, P. & Davis, E.A. (2000). Designing Mildred: Scaffolding Students' Reflection and Argumentation Using a Cognitive Software Guide. In B. Fishman (Ed.), *Proceedings of ICLS '00: The Fourth International Conference on the Learning Sciences*.

Blanchette, I. & Dunbar, K. (2001). Analogy use in naturalistic settings: The influence of audience, emotion, and goals. *Memory & Cognition*, 29, pp. 730-735.
Mahwah, NJ: Lawrence Erlbaum Associates.

Blumenfeld, P.C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M. & Palincsar, A. (1991). Motivating project-based learning: Sustaining the doing, supporting the learning. Educational Psychologist, Vol. 26 (Nos. 3 & 4), pp. 369-398.

Bransford, J.D., Brown, A.L. & Cocking, R.R. (1999). *How People Learn*, Washington, D.C. National Academy Press, 41-66.

Collins, A., Brown, J.S., & Newman, S.E. (1989). Cognitive Apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L.B. Resnick (Ed.) *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaiser*. Hillsdale, NJ: Erlbaum.

Gentner, D. (1999). Analogy. In W. Bechtel & G. Graham (Eds.), A companion to cognitive science (pp. 107-113). Oxford: Blackwell.

Guzdial, M. (1995). Software-realized scaffolding to facilitate programming for science learning. *Interactive Learning Environments*, 4(1), 1-44.

Kolodner, J.L. (1993). *Case-Based Reasoning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Kolodner, J.L., Crismond, D., Fasse, B. B., Gray, J., Holbrook, J., Puntembakar, S. (2003). Putting a Student-Centered Learning By Design™ Curriculum into Practice: Lessons Learned. *Journal of the Learning Sciences*, Vol. 12 No. 4..

Kuhn, D. (1993). Science as argument: implications for teaching and learning scientific thinking. *In Science Education*, 77(3), 319-337.

Owensby, J.N. & Kolodner, J.L. (2004). Case Interpretation Tool: Collaboratively Coaching Students' Understanding of Second-hand Experiences in Learning by Design Classrooms. To be presented as a paper at *International Conference of the Learning Sciences 2004 (ICLS 2004)*. Santa Monica, California, June 2004.

Owensby, J.N. & Kolodner, J.L. (2003). *Case Application Suite: A Study of Teacher Use in Learning By Design™ Classrooms*. Paper presented at the American Educational Researchers' Association (AERA) 2003 Conference, Chicago, IL.

Puntembekar, S. & Kolodner, J. L. (1998). The Design Diary: Development of a Tool to Support Students Learning Science By Design. *Proceedings International Conference of the Learning Sciences '98*, pp. 230-236.

Reiser, B.J., Tabak, I., Sandoval, W.A., Smith, B., Steinmuller, F., L & Leone, T.J. (2001). Bguile: Strategic and Conceptual Scaffolds for Scientific Inquiry in Biology Classrooms. In S.M. Carver & D. Klahr (Eds.) *Cognition and Instruction: Twenty five years of progress*. Mahwah, NJ: Earlbaum.

Schauble, L. Glaser, R., Duschl, R., Schulze, S. & John, J. (1995). Students' Understanding of the Objectives and Procedures of Experimentation in the Science \ Classroom. In *The Journal of the Learning Sciences* 4(2), 131-166.

Contribution of Reading Skill to Learning from Expository Texts

Yasuhiro Ozuru (y.ozuru@mail.psyc.memphis.edu)
Rachel Best (r.best@mail.psyc.memphis.edu)
Danielle S. McNamara (d.mcnamara@ mail.psyc.memphis.edu)

Psychology Department, University of Memphis
Memphis, TN 38152 USA

Abstract

Our study investigated the importance of reader aptitudes (prior knowledge and reading skill) in the processing of an expository text. We analyzed self-explanations produced by 42 high-school students while reading an expository text about thunderstorms. Specifically, we focused on students attempts to paraphrase information (i.e., restate sentences in their own words) and elaborate on the sentence content (e.g., connect information in different sentences to build a global representation of the text). Our findings suggest that reading skill is important for the active processing of expository texts. Skilled readers produced more elaborations than less skilled readers, and also a more diverse range of strategies, which may be crucial for supporting learning. Implications for learning from text are discussed.

Introduction

Successful reading comprehension requires the efficient co-ordination and integration of a number of underlying processes (Vellutino, 2003). These processes may include identification of words, decoding of word meanings, integrating the meaning of individual words into a coherent sentence level meaning using sentence syntax, and finally building a more global level representation of the text by continuously integrating the individual sentences into meaningful discourse level representation.

Our focus is on the higher level processes involved in reading comprehension, in particular, on how readers construct global mental representations of situations, and/or topics described by written material (van Dijk & Kintsch, 1983). Further, we focus on the impact of two reader-related factors on reading comprehension at this higher level of processing.

The first of these factors concerns the readers' knowledge of the topic of the text. As readers comprehension processes move from lower levels (e.g., word meaning) to higher levels (e.g., sentence/discourse processing), comprehension becomes increasingly influenced by the reader's knowledge about the topic of the text (Kintsch, 1988). General knowledge clearly influences the comprehension of narrative text (e.g., McNamara & McDaniel, 2004). Moreover, general and domain-specific knowledge is particularly critical for successful comprehension of expository texts (e.g., chemistry, biology). A great deal of research has shown that readers' knowledge facilitates comprehension and learning from expository texts (e.g., Chiesi, Spilich, & Voss, 1979; McNamara & Kintsch, 1996).

However, there are reasons to believe that other individual difference factors influence higher level processing of texts. In particular, factors associated with general reading skill, such as working memory capacity (Just & Carpenter, 1992), reading strategy knowledge (Guthrie, Anderson, Alao, & Rinehart, 1999), and meta-cognition (Baker, 2002) appear to be important. For example, McNamara and colleagues (McNamara, in press; O'Reilly, Best, & McNamara, in press; O'Reilly, Sinclair, & McNamara, in press) have shown that meta-cognitive reading strategy training improves science text comprehension, particularly for low-knowledge readers. Thus, active, strategic processing of text is particularly important to the comprehension of expository texts.

We explore here how reading skill is beneficial to comprehension of expository texts beyond the impact of prior knowledge found in various past research (e.g., Chiesi, Spilich, & Voss, 1979) A possibility explored here is the notion that general reading ability is associated with a set of skills that facilitate active processing of expository texts.

We hypothesized that knowledge level and reading ability would be associated with different aspects of the reading comprehension of expository texts. Specifically, whereas "knowledge" should be important for how easily one can comprehend the material, "reading ability" or "skill" should influence active, deep level processing of the material. Thus, our goal in this paper is to examine how reading ability as well as prior knowledge contributes to the "process of learning" from expository texts. We use the phrase "process of learning" here to refer to the processes engaged by students to acquire new

information from a text and integrate it into their own knowledge structures.

According to constructivist framework, learning occurs as a function of constructive activity one engages at the time of processing the materials (Cobb, 1994). We hypothesized that reading ability is associated with a set of skills that help readers engage in constructive activities such as elaborative inferences during reading, hence contributing to learning from the materials.

To tap into the effect of reading ability on "the process of learning," we analyzed self-explanations students produced when reading an expository text. Think aloud protocols are known to be useful for obtaining an insight into thought processes associated with problem solving (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Ericsson & Simon, 1993) and reading (Chi, 2000). Evidence further indicates that eliciting self-explanations during reading leads some readers to engage in active reading strategies that improve text comprehension (Chi, de Leeuw, Chiu, & Lavancher, 1994). Finally, there is evidence showing that the number of self-explanations (e.g., elaborative inferences) is correlated with learning (Chi, 2000). These studies reveal strong relations between self-explanation and the "process of learning." Thus, the analysis of self-explanations should afford an effective method of gaining insight into the "process of learning" that occur while students are reading expository texts.

Assuming skilled and less skilled readers possess different degrees of reading related aptitudes (e.g., motivation, working memory, effort, strategies) to deal with the reading comprehension situation, we expect variations to emerge in the type of reading strategies observed in the self-explanations generated while reading an expository text. In particular we hypothesize that reading ability will be associated with high frequency of constructive activities such as elaborative inferences.

The self-explanation protocols analyzed in this paper were collected during computerized reading strategy training, in which students typed their self-explanations to expository texts (see also, Best, Ozuru, & McNamara, 2004). The computerized trainer, called iSTART (Interactive Strategy Trainer for Active Reading and Thinking), is an automated reading strategy training program developed by McNamara, Levinstein, and Boonthum (in press). iSTART has been shown to be as effective in training reading strategies as its parallel, live version, called SERT (McNamara, in press; O'Reilly et al., in press).

Method

Participants

The sample consisted of 42 eighth and ninth grade children from an east coast suburban school. The students were enrolled in a learning program, called Learning Bridge, designed to provide summer school to students from under-privileged backgrounds.

Design and Materials

Individual differences were measured with two tests; a modified version of the Gates-MacGinitie Reading Skill Test and a Prior Knowledge Test. The Gates-MacGinite test is a standardized reading comprehension test, designed for grades 10-12. The test consisted of 40 multiple-choice questions that assess students' comprehension on several short text passages (Cronbach's Alpha α=.91). Due to time constraints we omitted the vocabulary comprehension section. The prior knowledge test had 35 multiple-choice items, which tap knowledge of different science domains, including biology, scientific methods, mathematics, earth science, physics, and chemistry (Cronbach's Alpha α=.81).

The iSTART system (McNamara et al., in press), in which the self-explanations were collected, consists of three phases: Introduction (introduces concept of self-explanation and reading strategies), Demonstration (shows users examples of self-explanation) and Practice (requires students to generate self-explanations).

In the practice section, students are presented with science texts one sentence at a time on the computer screen. For each sentence, they are asked to type a self-explanation. iSTART assesses the quality of the self-explanation and provides the feedback to students via the pedagogical agent, Merlin. The feedback is largely based on the degree of argument overlap between the students' self-explanation and the target sentence. The system is designed such that it encourages students to use information that is not in the target sentence (e.g., elaboration based on commonsense and previous sections of the text.). For example, Merlin might respond with "Try adding some more information that explains what the sentence means" when the self-explanation is too similar to the target sentence. Thus, feedback differs for each user, depending on the quality of self-explanations produced.

The self explanations analyzed in this paper are the final versions that the students provided for each sentence. Thus, they have been affected by the feedback of the system in that the final self-

explanations reflect better quality protocols than would otherwise have been provided under spontaneous circumstances. However, the system tends to reduce the difference between high and poor quality self explanations because there is a certain threshold for the acceptance. In this sense, the effects of individual differences reported in this paper are unlikely to be artifacts of the system's feedback.

Students self-explained two texts in the practice phase, "Stages of Thunderstorm Development" and "Origin of Coal." The present analysis focuses on self explanations for the thunderstorm text. This text, which was extracted from a school textbook, has 13 sentences and 197 words, with Flesch-Kincaid Grade level of 9.4.

Procedure

Individual difference measures were collected shortly before iSTART training. Students completed the prior knowledge test followed by the Gates-MacGinite reading test. They were given 15 minutes to complete each assessment. The students then completed the training with iSTART program. During the practice phase, they provided explanations to the Thunderstorms text.

Coding

Students typed their self-explanations, which were automatically recorded in the database. Two independent coders analyzed the self-explanations in terms of the following five dimensions: 1) presence of comprehension monitoring; 2) presence of paraphrasing (none, topic identification, repetition, and paraphrasing); 3) distance of paraphrasing from the target sentence; 4) accuracy of the paraphrasing; and 5) presence of elaborations.

Coding of comprehension monitoring assessed whether self-explanations incorporated the monitoring of students' understanding. Explanations were coded for the presence or absence of comprehension monitoring statements (e.g., 'I don't understand X'). The presence of paraphrasing was judged on students' attempt to restate the target sentence in their own words. For this coding, a self explanation was categorized as one of the following: 1) a paraphrase that was a restatement of the sentence using different words, 2) a repetition of the sentence that was lexically too similar to the target sentence, 3) a simple topic identification (e.g., 'this is about storms'), or 4) no paraphrase, repetition, or topic identification. If the explanation was categorized as a paraphrase, the paraphrase was further coded for *accuracy* and *distance* from the target sentences. *Accuracy* has three levels (inaccurate, partially accurate, and accurate); and *distance* has two levels (distant and close). Close paraphrases were closely aligned to the original sentence in terms of sentence structure and/or content words. Distant paraphrases contained the same semantic content as the target sentence, but did not have the same sentence structure or content words.

Coding of the elaborations was based on whether the self-explanations included any ideas that were not explicitly present in the target sentence. Once a self-explanation was found to contain an elaboration, it was further coded for the nature of its *contribution:* 1) relevant to the comprehension of neither the current sentence nor the overall text; 2) relevant and contributes to the comprehension of only the target sentence; and 3) relevant and contributes to a global level of comprehension that goes beyond the current sentence (e.g., actively building the large picture depicted by overall text). We also coded the elaborations in terms of their source and accuracy. However, since the present analysis does not focus on these aspects, they are not described here (see Best et al., 2004).

Reliability of the coding was evaluated using Cohen's Kappa and simple agreement (when the coding is binary). Reliability between the coders was 85% or above for all coding dimensions. Disagreements were resolved via a discussion between the coders.

Results

To explore the role of the reading skill and prior knowledge, we adopted a median split method; students were divided into high and low reading comprehension skill, or high and low-knowledge groups, using the median scores of The Gates-MacGinite test or Prior Knowledge test. The correlation between the Gates-MacGinite test and the prior knowledge test was high, $r = 0.604$, $p < .001$.

Our data indicated that students' often attempted to paraphrase (91.0%) and elaborate (41%), but seldom expressed comprehension monitoring (4%). Thus, the subsequent analysis focuses on paraphrase and elaborations.

We used analysis of variance (ANOVA) to analyze the students' use of the strategies. The univariate ANOVAs comprised reading skill (high or low) or knowledge (high or low) as the between-subjects factor and strategy type (e.g., paraphrases) as the dependent variable.

Frequency distribution of strategy use

Our first analysis investigated the frequency of four types of strategy used by different types of readers (low and high reading skill or knowledge students). For this analysis, we classified all the self-explanation into one of four categories: 1 =

Repetition or topic identification only; 2 = Paraphrasing only; 3 = Elaboration I (irrelevant or current sentence); and 4 = Elaboration II (knowledge building).

It is important to note that elaboration (both type I and II) may or may not contain paraphrasing. Elaboration I and Elaboration II were treated separately because knowledge building elaborations are indicative of the investment of a greater effort to understand the text (i.e., building a more global representation of the text). Irrelevant elaborations were coded under Elaboration I because they indicate that the student is investing an effort to integrate the information into their knowledge structures.

The levels of the classification scheme reflect an increase in the construction of integrated and global representations of the situation described by the text. For example, repetition does not require any integration because it only involves rewriting the text information. In contrast, paraphrasing requires the restatement of the situation using different words or sentence structure, revealing how students understood the meaning of the sentence. Similarly, Elaboration I indicates that students are making an effort to understand the target sentence by relating, integrating, or comparing the information in the sentence with what they already know. Finally, Elaboration II involves effort to integrate information appearing on a multiple sentences into a coherent model. Therefore, this analysis can examine the degrees of constructive activities carried out by the student.

The first analysis investigated the self-explanation strategies used by skilled and less skilled readers. As shown in Table 1, skilled readers were more likely to use elaboration strategies whereas less skilled readers were more likely to repeat or paraphrase only.

We performed two separate ANOVAs, assessing the frequency with which skilled and less skilled readers used elaboration types I and II. Skilled readers produced more current sentence focused elaborations (elaboration I), $F(1, 39) = 6.75$, $MSE = .053$, $p = .01$, and knowledge building elaborations (elaboration II), $F(1, 39) = 7.6$, $MSE = .003$, $p < .01$, than did less skilled readers. The difference between skilled and less skilled readers in current sentence focused elaboration (Elaboration I) is not solely attributable to irrelevant elaborations included in Elaboration I because the difference remained marginally significant after excluding irrelevant elaborations: (less skilled readers $M=.22$ $SD=.22$) and (skilled readers $M=.33$ $SD=.20$), $F(1, 39) = 2.95$, $MSE = .043$, $p = .094$.

Table 1. Strategy use and comprehension skill

Strategy use	Less Skilled	Skilled
Repetition/Topic identification	.08 (.07)	.03 (.07)
Paraphrasing	.62 (.20)	.46 (.23)
Elaboration I	.29 (.23)	.46 (.22)
Elaboration II	.01 (.02)	.05 (.07)

Note: Standard deviations are in parentheses.

The previous analysis was repeated with knowledge as the between-subjects variable. As shown in Table 2, there were few differences in the distribution of strategies used by high and low-knowledge readers. ANOVAs confirmed this conclusion: there was no difference in the frequency with which low-knowledge and high-knowledge students used Elaboration I, $F(1,39) = 1.04$, p = ns, or Elaboration II, $F(1, 39) = .017$, p = ns. There was no difference, again, in Elaboration I even after excluding irrelevant elaborations: low-knowledge ($M= .28$ $SD =.24$) and high-knowledge ($M= .28$ $SD = .19$) students.

Table 2. Strategy use and knowledge

Strategy use	Low Knowledge	High Knowledge
Repetition/Topic identification	.07 (.06)	.04 (.09)
Paraphrasing	.55 (.27)	.53 (.20)
Elaboration I	.35 (.26)	.40 (.23)
Elaboration II	.03 (.06)	.03 (.10)

Note: Standard deviations are in the parentheses.

Variety of strategy use

Next, we were interested in how skilled and less skilled readers and high and low-knowledge readers differed in terms of the diversity of the strategies they use in self-explaining the 13 sentences of the text.

We hypothesized that less skilled readers would be relatively uniform in their strategy use (i.e., use only one type of strategy). In contrast, skilled readers would frequently change their strategy use (i.e., adapt their strategies to different sentences). On the other hand, we predicted that knowledge level would make little difference in terms of the variety of strategies students' used.

To conduct this analysis, we counted how many of the aforementioned strategies (paraphrase, type I elaboration, and type II elaboration) each student used to self-explain the 13 sentences. We did not count repetition/topic identification as a strategy because it is a default technique. Accordingly, the

score each student obtained varied from 0 (no strategy used) to 3.

The analysis showed that reading skill was an important determiner with regard to the variety of self-explanation strategies used. An ANOVA indicated that skilled readers used a greater variety of self-explanation strategies (M = 2.4, SD = 0.6) than did less skilled readers (M = 2.0, SD = 0.5), F(1, 39) =5.9, MSE = .277, p = .02.

On the other hand, the analysis based on readers' knowledge indicated that there was no reliable differences in the range of self-explanation strategies used by high (M = 2.3, SD = 0.5) and low-knowledge students (M = 2.1, SD = 0.3), F(1,39) = 1.45, ns.

Effect of prior knowledge and reading skill on the quality of paraphrases

Thus far, our analyses have indicated that reading skill, rather than prior knowledge is an important factor in determining the frequency of elaborative inferences and variety of self-explanation strategies students employ. Does this mean that prior knowledge does not play an important role in the reading comprehension process? According to our argument, and existing literature, prior knowledge should have a large influence on comprehension level of expository texts. In order to examine the effect of prior knowledge on comprehension, we analyzed the quality of paraphrases. Given that paraphrasing involves describing the gist of a sentence using one's own words, the quality of paraphrases should reflect the students' understanding of sentence meaning.

Our analysis focused on two different qualities related to successful paraphrasing: 1) accuracy of the paraphrase and 2) distance of the paraphrase. For accuracy, a score of 0, 0.5 and 1.0 was assigned for inaccurate, partially accurate, and accurate paraphrases, respectively. The two sets of univariate ANOVAs on the accuracy scores indicated that there were no effects of individual differences. This rather disappointing result is possibly due to the fact that the target sentence was available for reference while students self-explained the sentence.

Turning to the distance of paraphrase, we compared the frequency of distant paraphrases produced by high and low-knowledge and/or by skilled and less skilled students. Two sets of univariate ANOVAs on the frequency of distant paraphrases indicated that distant paraphrases occurred more frequently for high-knowledge (M = 0.63, SD = 0.22) than low-knowledge students (M = 0.44, SD = 0.24), F (1, 39) = 8.153, MSE = .051, p <.01. Also the main effect of reading comprehension skill was marginally significant, with more distant paraphrases produced by skilled readers (M = 0.61, SD = 0.29) than less skilled readers (M = 0.47, SD = 0.24), F (1, 39) = 2.891, MSE = .057, p =.097.

We also examined accuracy of distant and close paraphrases by dividing all of the items into distant or close paraphrases, and analyzing whether the accuracy differed across distant and close paraphrases. The analysis revealed that accuracy does not differ between distant (M =.62, SD = .22) and close paraphrases (M =.69, SD = .20), F(1, 38) = 2.215, p > 0.1, suggesting that distant paraphrases are not necessarily less accurate. Overall, these analyses indicate that students' prior knowledge has a larger effect on the production of distant paraphrases. Given that producing a distant paraphrase without distorting the meaning of the sentence requires accurate comprehension of the sentence, this finding confirms previous findings for effects of knowledge on comprehension level of the expository text.

Discussion

Overall, the analyses support our prediction that reading comprehension ability is closely associated with the effort and strategies that readers expend to understand the expository text, whereas knowledge is more closely associated with the actual comprehension level of the material (as indicated by the distance of the paraphrase analysis). Skilled readers' self-explanations tend to include more constructive activities, such as elaborative inferences and linking different parts of the text to obtain a "larger picture." This finding is remarkable because it demonstrates that skilled readers' are able to generate these elaborative and bridging inferences even when they are dealing with relatively unfamiliar materials (i.e., an expository text about thunderstorm as opposed to narrative texts used in Gates-MacGinite test). One interpretation of this finding is that skilled readers possess skills/strategies to effortfully activate relevant information from relatively unfamiliar text-based information (McNamara, in press). This ability is associated with constructive activities such as bridging inferences, and elaborative inferences. Use of these types of strategies, either naturally or after being trained to do so, contributes positively to learning from expository texts.

One limitation of our study is that it focuses solely on the processes readers employ while reading the text, and hence, does not directly show learning gains for skilled readers. However, research supports the assumption that greater use of elaboration is associated with learning gains (Chi, 2000; McNamara, in press).

There are two contrasting views regarding how elaborations contribute to learning: the incomplete text view and the self-repair view (Chi, 2000). These views differ critically with respect to whether (or to

what extent) the accuracy of readers' elaborations affect leaning. In line with Chi's (2000) view, our assumption here is that elaborations facilitate learning, regardless of whether they are accurate (see also, McNamara, in press). But, of course, we do not rule out the likelihood that there are several different ways in which elaborations can contribute to leaning. Accuracy of elaborations may well have important implications for the learning process, particularly when the reader does not have the opportunity to repair inaccurate elaborations based on information encountered later in the text.

In conclusion, the present research highlights the important role active reading strategies play in the comprehension of and learning from expository texts. Given that active strategies are beneficial, future work should explore, in more detail, the reader related factors (e.g., metacognition, knowledge on the strategies, etc) that underlie the use of the strategies, and the text-related factors (e.g., sentence difficulty, that affect active reading.

Acknowledgements

We would like to thank the members of the ODU Strategy Lab who helped to conduct this study. This project was supported by the NSF (IERI Award number 0241144). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Baker, L. (2002). Metacognition in comprehension instruction. In C.C. Block & M. Pressley (Eds.), Comprehension instruction: Research-based best practices (pp. 77-95). New York: Guilford Press.

Best, R. M., Ozuru, Y. & McNamara, D. S. (2004). *Self-explaining science texts: strategies, knowledge and reading skill*. Proceedings of the International Conference for the Learning Sciences, CA, LA.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In Glaser, R. (Ed.). Advances in Instructional Psychology, Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M. T. H., de Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self explanations improves understanding. *Cognitive Science*, 18, 439-477

Chiesi, H. I., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 275-290.

Cobb, P. (1994). Where is the mind? Constructivism and sociocultural perspectives on mathematical development. *Educational Researcher, 23,* 13-20.

Ericsson, K. A., & Simon, H. (1993). Protocol Analysis (Rev.ed.). Cambridge, MA: MIT Press.

Guthrie, J. T., Anderson, E., Alao, S., & Rinehart, J. (1999). Influences of concept oriented reading instruction on strategy use and conceptual learning from text. *Elementary School Journal, 99*, 343-366.

Just, M. A., & Carpenter, P.A. (1992). A capacity hypothesis of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122-149.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction integration model. *Psychological Review, 95*, 163-182.

McNamara D. S. (in press) SERT: Self-Explanation Reading Training. *Discourse Processes*.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-288.

McNamara, D. S., & McDaniel, M. A. (2004). Suppressing irrelevant information: knowledge activation or inhibition? *Journal of Experimental Psychology, Learning, Memory, and Cognition, 30*, 465-482.

McNamara, D. S., Levinstein, I. B. & Boonthum, C. (in press). iSTART: Interactive Strategy Trainer for Active Reading and Thinking. *Behavioral Research Methods, Instruments, and Computers*.

O'Reilly, T., Best, R. & McNamara, D. S. (in press). Self-explanation reading training: Effects for low-knowledge readers. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

O'Reilly, T., Sinclair, G. P. & McNamara, D. S. (in press). Reading strategy training: Automated versus live. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Vellutino, F. R. (2003). Individual differences as sources of variability in reading comprehension in elementary school children. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking Reading Comprehension*. New York, NY: Guilford.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies in discourse comprehension*. New York: Academic press.

The Social Circle Heuristic: Fast and Frugal Decisions Based on Small Samples

Thorsten Pachur (pachur@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94
14195 Berlin, Germany

Jörg Rieskamp (rieskamp@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Lentzeallee 94
14195 Berlin, Germany

Ralph Hertwig (ralph.hertwig@unibas.ch)
University of Basel, Department of Psychology, Missionsstrasse 60/62
4055 Basel, Switzerland

Abstract

Whereas reliance on information from one's proximal social environment for generalizing about the population has often been associated with erroneous judgments, this information is often valuable and can be exploited for making accurate inferences. The social circle heuristic is a judgment mechanism in which the content and structure of people's social networks are used for making inferences about frequencies in the population in a paired comparison task. Because the heuristic has a stopping rule, judgments generated by it will often be based on small samples sizes. In this paper we present experimental evidence that shows both that the social circle heuristic can compete with a more thorough strategy, and that people actually apply it.

Samples as Reflections of the Environment

Scarcity of information is one of the central properties of everyday decision making. For many judgment problems in the real world, neither direct knowledge of the to-be-judged values nor complete knowledge of all relevant facts that might help predict the correct value are available. Instead, inferences have to be made under uncertainty, based on information that is more or less predictive of the criterion. What processes underlie people's inferences in such situations? Recent approaches to judgment under uncertainty that acknowledge the bounded rationality of humans have advanced the notion of *fast and frugal decision making* (Gigerenzer, Todd, & The ABC Research Group, 1999). The heuristics proposed by this program are based on Brunswik's (1955) idea that judgments are made on the basis of cues that are probabilistically related to the target criterion. As Gigerenzer et al. (1999) have shown, such mechanisms can be astonishingly accurate despite using only a limited amount of information.

Typically, fast and frugal heuristics rely on cues that are qualitatively different from the criterion (e.g., considering whether or not there is rent control in a city to predict which of two cities has a higher homeless rate). In the case of frequency judgments, however, the target criterion can be inferred by sampling instances of it from a population. For example, which first name occurs more often in the population: Martin or Simon? Or, does bladder cancer or renal cancer have a higher annual incidence rate? For these inference problems concerned with environmental frequencies, it is possible that—rather than accessing proximal cues—samples consisting of instances of the criterial event are drawn from the proximal environment. As such, samples can also serve as "keys to assessing the distal environment" (Fiedler, 2000, p. 661), and in the absence of direct knowledge about the environment, these "reflections" of the environment are used to infer its latent properties.

That humans use proximal samples when making inferences about the entire population has been argued in various forms. For instance, in one interpretation of the availability heuristic (Tversky & Kahneman, 1973; see also Sedlmeier, Hertwig, & Gigerenzer, 1998), frequencies in the environment are judged by accumulating easily accessible instances of the target event (e.g., Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). Judgment phenomena such as the false consensus effect (Ross, Green, & House, 1977) or the optimistic bias (Weinstein, 1980), have been attributed to the employment of the availability heuristic, and the use of such a small-sample-based heuristic has therefore been associated with the fallibility and irrationality of human decision making.

In contrast, we ask how humans can achieve fairly accurate judgments in spite of the scarcity and cognitive boundedness they have to face in the real world, and examine what processes contribute to this achievement (cf. Krueger & Funder, in press). Elaborating on the idea of sample-based judgments, we propose and test a simple heuristic for paired comparisons that exploits frequency information in one of people's most proximal environment: their social network.

The Social Environment as Sample Space

People's inferences have been shown to be strongly sensitive to information in their social environment. Prominent examples are attitude formation (e.g., Fishbein & Ajzen, 1975), conformity behavior (Hirshleifer, 1995; Latané, 1981) or risk frequency judgments (Benjamin, Dougan, & Buschena, 2001; Hertwig, Pachur, &

Kurzenhäuser, 2003). Furthermore, there is evidence that information obtained from individuals is accessed and used more readily than is the same information obtained in an abstract, statistical format when making judgments (Borgida & Nisbett, 1977)—even when it is pointed out that the concrete individual represents a highly unrepresentative instance (Hamill, Wilson, & Nisbett, 1980).

Apart from the well-known vividness argument, information obtained about concrete individuals could receive special prominence for several reasons: First, as no mediating factor can distort it, information directly obtained or observed about the members of one's own social network is highly *reliable*. Second, the observations of instances of the target criterion are per se a *valid* indicator of the criterion. Further, the information is easily *accessible*, as information about social network members represents a constantly recurring and thus well-rehearsed event. Finally, observations of criterial events in one's social environment are *naturally sampled*, that is, encountered sequentially and represented as natural frequencies. This format has been shown to foster probabilistic reasoning (Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000).

Based on these reasons, we propose that people use their social network as a sample space to search for information they use to draw numerous inferences. Specifically, we propose one heuristic, the social circle heuristic, which makes inferences about which of two events occurs more frequently in the entire population: With the heuristic, instances of the events in question are sampled from a person's social network.

How Social Circles Guide Search and Stop Search

In light of the computational limitations of human cognition and the fact that inferences often have to be made without an exhaustive search of available information (Simon, 1956), the question of when to stop information search arises. In other words, when does one stop sampling from one's social network?

An individual's social network is no homogenous entity of identical types of relationships. Rather, one can argue that social networks have a hierarchical structure, with the relationships that a person has to the members of his or her social network differing in genetic relatedness, frequency of contact, emotional closeness, content of contact, and function (e.g., Milardo, 1992). Collapsing across these dimensions, we will differentiate among the following social circles: family, friends, and acquaintances.

A central idea of the social circle heuristic is that the structure of the social network is used during the sampling process, that is, the heuristic exploits the hierarchical structure of the social environment to guide and stop the sampling process. A popular notion in social network research has been to represent the hierarchical structure of a social network as concentric circles (Moreno, 1936; Kahn & Antonucci, 1980), with the person whose network is described in the focal circle, and persons of increasing "distance" to the person occupying increasingly peripheral circles. For instance, one's family might fall in the circle second closest to the middle, friends in the third circle, and one's acquaintances in the outer circle. As described in the next section, the social circle heuristic works by sequentially sampling instances of the events in question from the different circles, starting with the focal circle. As soon as the search of a complete circle favors one of the two alternatives, the sampling process is terminated and no further circles are looked up. Note that moving outwards, the number of the circles' members increases monotonically, and, as a consequence, so too does the sample size on which an inference can be based.

The Social Circle Heuristic

After defining the sample space and its structure, we are now in the position to describe the social circle heuristic in more detail. Consider the following inference problem: Which disease occurs more often in the population, hepatitis or tuberculosis? The social circle heuristic is a heuristic for such pair comparisons in which events (or characteristics) are judged according to their population frequency.

The heuristic consists of four building blocks and starts with the recognition heuristic (Goldstein & Gigerenzer, 2002).[1] The social circle heuristic has a search rule, which specifies where to search, a stopping rule, which specifies when to stop sampling, and a decision rule, which specifies how to make an inference based on the information gathered through sampling (for different building blocks of heuristics see Gigerenzer et al., 1999).

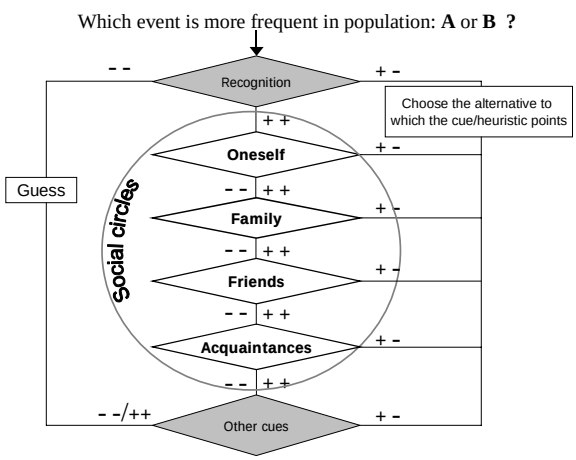


Figure 1: Flow chart of the social circle heuristic and the relationship of the sampling process to the recognition principle and inferences based on other cues (such as Take the Best; Gigerenzer & Goldstein, 1996).

It operates as follows:

Step 0 – Use the *recognition heuristic*. If the name of only one of the two events is recognized, then predict that the recognized one is more prevalent than the

[1] This latter point also illustrates how fast-and-frugal heuristics can be combined by nesting.

unrecognized one. If the names of both events are recognized, recruit the social circle heuristic.

Step 1 – *Search rule*: Search the social circle for instances of the events, running sequentially through the circles, starting with the focal circle.

Step 2 – *Stopping rule*: If search within a circle favors one event, stop search. If; within a circle; the same number of instances is found for both events, continue the search in the next circle.

Step 3 – *Decision rule*: Predict that the event for which a higher number of instances is found is the more prevalent in the population. If the sampled information does not discriminate between the alternatives (and no other information is known), then guess after the last circle is searched.

Due to the stopping rule, the search process will often be terminated early, and an inference based on information gathered with the social circle heuristic will be derived from samples of small sizes. Note that this also implies that as soon as search is stopped at a particular circle, information in more peripheral circles that might overturn the decision, is not considered. In this sense, the heuristic is non-compensatory. To rely on small samples has often been seen as unreasonable ("belief in the law of small numbers"; Tversky & Kahneman, 1971), and only recently a few authors have highlighted the possible value of such a strategy (e.g., Dawes, 1989; Fiedler & Kareev, 2004; Kareev, 2000).

How Accurate Is the Social Circle Heuristic?

In order to test the accuracy of the social circle heuristic, we conducted a computer simulation where the task of the heuristic was to judge which of two events, A or B, occurs more frequently in the entire population. For this task, the heuristic could search for instances of the events in its spatial vicinity. The population consisted of 2,500 agents, represented in a 50×50 matrix in which each cell represented one agent (see Figure 2, which shows the environment simplified to a population with 100 agents in a 10×10 matrix). We used the city block metric to define the distance between the agents. For instance, in Figure 2—which shows the social network of agent #45 in a population of 100—agent #44 is at a distance of 1 from agent #45, agent #34 is at a distance of 2 from agent #45, agent #33 is at a distance of 3 from agent #45 etc. It is assumed that each agent's social network consists of 40 other agents that differ with regard to their distances to the agent. Thus, an agent could maximally sample information about 41 agents (including himself). This social network is divided into different social circles: Circle 1, that only includes the agent itself, Circle 2, including all neighboring agents with a distance of 1 (4 agents), Circle 3, including all neighboring agents with a distance of 2 (8 agents), and Circle 4, including all neighboring agents with a distance of 3 or 4 (28 agents).[2]

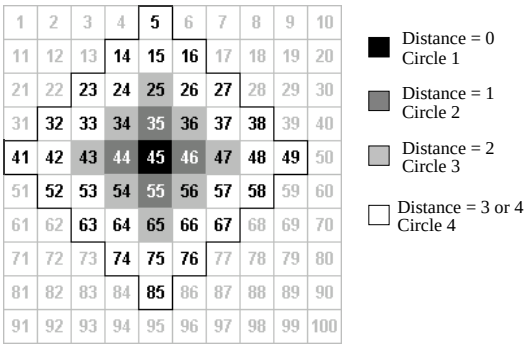


Figure 2: Environment in the computer simulation (here simplified as a 10×10 population).

Two environments were used to test the performance of the social circle heuristic. In the first environment, instances of 10 event categories were distributed randomly across the 2,500 agents (see Figure 3). The 10 events mimicked the frequency distribution of a real world environment used in the experiment (discussed further below): occurrences of infections in Germany. As can be seen from Figure 3, the distribution of the proportions of the infections is very skewed and falls into a J-shaped distribution, a pattern found in many real-world domains (Hertwig, Hoffrage, Martignon, 1999). The proportions of the 10 most frequent infections (from a set of 24) were chosen because their proportional distribution could be represented in a population of 2,500 agents. The most frequent event was set at a frequency of 2,000; the 9 other events were distributed according to this anchor and the proportions reported by the Robert Koch Institute (for details see Pachur, 2002).

Secondly, we constructed an environment in which the same overall number of instances in the population was distributed across the 10 event categories such that the frequency across the 10 events was linearly increasing. As a result, this linear environment and the skewed environment differed substantially with regard to the dispersion of the frequency distribution. We were interested in the effect of the frequency distribution of the events as this property can have an effect on the success of a strategy (Hertwig, Hoffrage, Martignon, 1999).

To make an inference, the social circle heuristic starts with Circle 1 and looks whether event A or B is present. If one is and the other not, no further circles will be looked up, irrespective of what information is present in the other circles, and it will be inferred that the sampled event is more frequent in the population. If neither of the events is present in Circle 1, the agents in Circle 2 will be looked up. If one

[2] The matrix was a wrapped environment, that is, the agents at the borders had neighbors at the opposite side. For instance, in Figure 2, the left-hand neighbor of agent #41 is agent #50.

event occurs more frequently in Circle 2, search is stopped and an inference is made after looking up only four agents. The same rule applies to Circle 3. Only if the number of instances in the first three circles does not discriminate, then Circle 4, and thus the maximum number of 41 agents will be looked up. If even circle 4 does not discriminate, one of the events is picked randomly.

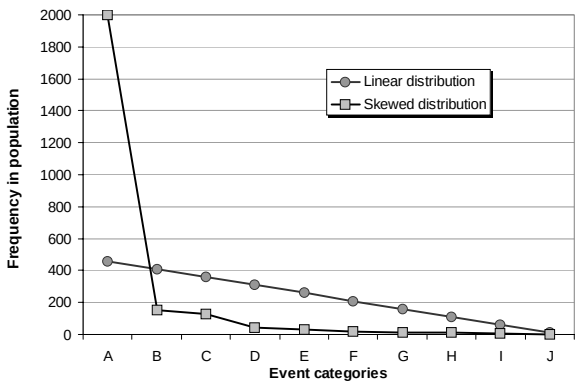


Figure 3: Distribution of the 10 events in the two environments used in the computer simulation.

As a benchmark for the social circle heuristic, its performance was compared with the performance of an exhaustive sampling strategy. For an inference of whether event A or event B occurs more often in the entire population, this strategy, normatively more appropriately, always looks up all 41 agents in the social network (that is, this strategy aggregates information across all circles). The event for which more instances can be sampled is inferred to be more frequent in the entire population. If an equal number of instances is sampled for both events, or if no instances can be sampled at all, one of the events is picked randomly.

For each of the two environments, the random distribution of the 10 events (totalling around 2,400 instances) was repeated 100 times, and each time 100 agents were picked randomly as starting points for the two strategies. At each run, the 10 events were combined in a pair comparison (yielding 45 pairs) and the task was to infer which event is more frequent in the entire population.

How well does the simple social circle heuristic perform compared to the exhaustive sampling strategy? In the skewed environment, derived from a real-world distribution, surprisingly, both strategies showed an identical proportion of correct inferences with a median of 77.8% (arithmetic means: social circle heuristic 76.3%, exhaustive sampling strategy 77.5%). Showing a similar level of performance, the social circle heuristic looked up, on average, only 24.7 agents, which is approximately 55% of the amount of information that the exhaustive sampling strategy used (which always looks up all 41 agents).

In the linear environment, the picture was different: here the exhaustive strategy clearly achieved a higher accuracy than the social circle heuristic. Whereas the social circle heuristic a median of 75.6% correct choices (mean 75.2, *SD* =8.9), the exhaustive strategy 84.4% (mean 83.1, *SD*=6.3).

Table 1 shows the performance of the social circle heuristic in more detail. Because of the stopping rule, the social circle heuristic terminated for some inferences the search at Circle 1, for some at Circle 2, for some at Circle 3, for some at Circle 4, and for some a guess had to be made. The second column of Table 1 reports for each circle the percentage of choices for which search was stopped at the circle. The social circle heuristic had to guess in 31.6% of the cases (whereas the exhaustive sampling strategy had to guess in 33.6% of all choices). The rightmost columns shows the percentage of correct inferences for these choices. Note that in the skewed environment, contrary to normative expectations, the accuracy *decreases* from Circle 1 to Circle 3. In the linear environment, in contrast, the accuracy increased.

Table 1: Proportion and accuracy of choices after search was terminated for the SCH in the two environments. The ns in the first column refer to the number of agents looked up.

Circles	% of choices stopped at each circle		% of correct choices	
	Skewed	Linear	Skewed	Linear
Circle 1 (n=1)	19.2	18.8	95.7	69.7
Circle 2 (n=5)	13.2	38.4	88.2	74.0
Circle 3 (n=13)	13.8	26.2	84.2	78.1
Circle 4 (n=41)	22.3	13.7	85.5	85.8
Guessing	31.6	3.0	50.0	50.0

Thus, we have accumulated a number of arguments for the usefulness of the social circle heuristic. First, it is a simple strategy that can be assumed to be easily performed by a boundedly rational agent. By restricting the search process and the amount of information on which an inference is based to a minimum, the social circle heuristic allows for very quick judgments. Second, as we have seen, it performs equally well as a more thorough strategy that takes much more information into account, and this performance seems to hinge on the statistical structure of the environment. Overall, the social circle heuristic achieves an astonishingly high proportion of correct inferences. But can we find evidence for people's use of such a simple and efficient strategy for making inferences about event frequencies?

Do People Use the Social Circle Heuristic?

The 24 infectious diseases (the proportions of 10 of these were also used in the computer simulation) for which official records are kept by the Robert Koch Institute were combined in a complete paired comparison (yielding 276 pairs), and 40 participants were asked to choose the infectious disease that has a higher annual incidence rate in Germany. After this test, participants indicated for each infection and each of their circles (self, family, friends, and

acquaintances) how many, if any, people in their circles had been affected by the infection. They also indicated whether they recognized the name of the infection. From this information, we calculated how often participants had an opportunity to choose in accordance with the social circle heuristic and determined which prediction the social circle heuristic made in each of these cases (only pairs where both infections were recognized and the reported number of instances in the social network discriminated between the two infections were included). Overall, only relatively little occurrences of the infections were reported by the participants, which is not surprising given the rarity of infections. The social circle heuristic made predictions for 33 participants, and was applicable (i.e., discriminated between the infections), on average, with 11.1% of all choices. Figure 4 shows how often these 33 participants made a choice in accordance with the prediction of the social circle heuristic.

For each participant, the bar indicates the percentage of choices that were in line with the prediction of the social circle heuristic. Overall, the median proportion of inferences in accordance with the social circle heuristic was 79.5% (mean 77%, *SD*=15.9). It seems fair to conclude that the social circle heuristic did quite a good job in describing participants' judgments (focusing on those in which it was applicable).

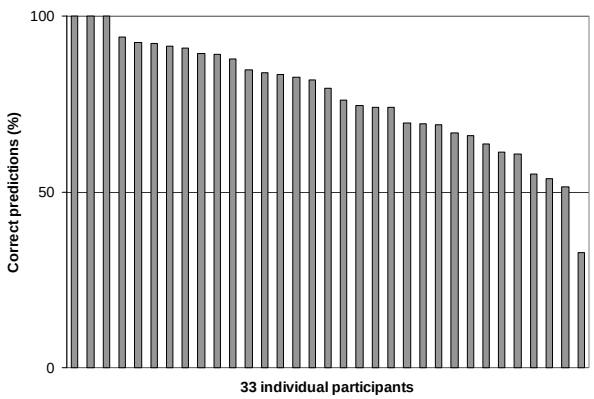


Figure 4: How often the 33 participants who reported instances of the infections in their social network made choices in accordance with the social circle heuristic.

How Ecologically Rational Is the Social Circle Heuristic?

The social circle heuristic is a psychologically plausible strategy: people appear to use it when trying to infer with of two risk events is more frequent. But how accurate a strategy is the heuristic when applied to the infections and based on the occurrences of the infections recalled by our participants? In other words, how ecologically rational are the inferences of the social circle heuristic in the task that our participants solved? In a second analysis, the predictions of the social circle heuristic for each individual were compared with the correct choices, that is, according to the actual incidence rates (averaged values from a 5-year period were used to eliminate year-to-year fluctuations).

An index for the ecological validity was defined as the number of correct inferences made by the social circle heuristic divided by the number of pairs where it was applicable. This index was calculated separately for each participant. The median ecological validity was .83 (mean .78), indicating that, overall, strictly following the social circle heuristic when it was applicable would have led to an accuracy of over 80% correct choices.

In contrast, how did the non-adherence to the social circle heuristic affect participants' performance? As indicated by the ecological validity index, strictly following the social circle heuristic would have yielded over 80% correct choices (which is far above the performance the participants achieved overall). Analyzing the choices that were in line with the predictions of the social circle heuristic and those that were not in line with it, it turned out that when the participants could apply the heuristic and did, they achieved on average 83.4% (*SD*=18.5) correct choices, whereas when they could but did not apply the heuristic, they achieved on average only 45% (*SD*=25.5) correct choices.

To be able to evaluate the accuracy of the social circle heuristic for the domain of infections, we also tested the strategy that always takes all instances that our participants reported into account. The exhaustive strategy showed a very similar fit with our participants' choices (median 81.8% mean 77.6% of choices in line with the predictions of this strategy), but was applicable in slightly fewer cases. In terms of ecological validity, the predictions of this strategy achieved no higher accuracy than the social circle heuristic (median ecological validity of .83, mean .79), which is in line with the results of the computer simulation.

Discussion

In the real world, inferences from small samples need must not be less accurate than inferences from larger samples. In this paper we investigated, both in a computer simulation and in an empirical study, a simple decision mechanism that exploits a person's social network as an easily accessible sample space for judging event frequencies in paired comparisons. The results show that the social circle heuristic allows one to judge accurately, with simple search, stopping, and decision rules, the environmental frequencies of randomly distributed events in a paired comparison task. At the same time, this mechanism describes people's choices rather well. Thus, the performance of the social circle heuristic provides another instance for the argument that small samples can be an efficient basis for judgments in the real world (cf. Fiedler & Kareev, 2004; Kareev, 2000).

This paper was intended to explore the appropriateness of this heuristic in an environment that has a naturally occurring statistical structure, and it was shown that the heuristic works particularly well in such an environment. A question for future research is why the heuristic works so well under these conditions and how it performs in environments in which events occur in clusters.

By virtue of its reliance on the structure of social networks, the social circle heuristic represents another example of a judgment policy in which the mind is a mirror image of the environment. The social circle heuristic thus follows in the footsteps of the pioneering work by Egon Brunswik (1955), John Anderson (e.g., Anderson & Schooler, 1991), and Roger Shepard (e.g., 1994).

Acknowledgments

The authors wish to thank the Max Planck Institute for Human Development for financial support.

References

Anderson, J. R., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396-408.

Benjamin, D. K., Dougan, W. R., & Buschena, D. (2001). Individuals' estimates of the risks of death: Part II—New evidence. *Journal of Risk and Uncertainty*, *22*(1), 35-57.

Borgida, E., & Nisbett, R. E. (1977). The differential impact of abstract vs concrete information on decisions. *Journal of Applied Social Psychology*, *7*(3), 258-271.

Brunswik, E. (1955). Representative design and probabilistic theory in functional psychology. *Psychological Review*, *62*, 193-217.

Dawes, R. M. (1989). Statistical criteria for a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*, 1-17.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*, 659-676.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650-669.

Gigerenzer, G. & Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002) Models of Ecological Rationality: The Recognition Heuristic. *Psychological Review*, *109*(1), 75-90.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology*, *39*, 578-589.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick Estimation: letting the environment do the work. In: G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (p. 209-234). New York: Oxford University Press.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2003). The accuracy and processes of judgments of risk frequencies. Manuscript submitted for publication.

Hirshleifer, D. (1995). The blind leading the blind. In M. Tommasi & K. Ierulli (Eds.), *The New Economics of Human Behavior*. Cambridge: Cambridge University Press.

Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. (2000). Communicating Statistical Information. *Science*, *290*, 2261-2262.

Kahn, R. L., & Antonucci, T. C. (1980). Convoys over the life course: Attachment, roles, and social support. *Life Span Development and Behavior*, *3*, 253-286.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, *107*(2), 397-402.

Fiedler, K., & Kareev, Y. (2004) Does decision quality (always) increase with the size of information samples? Some vicissitudes in applying the law of large numbers. Manuscript submitted for publication.

Krueger, J. I., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*: 596-610.

Krueger, J. I., & Funder, D. C. (in press). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences.*

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*, 343-356.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human learning and memory*, *4*(6), 551-578.

Milardo, R. M. (1992). Comparative methods for delineating social networks. *Journal of Social and Personal Relationships*, *9*, 447-461.

Moreno, J. L.(1936). Organization of the social atom. *Sociometric Review*, *1*, 11-16.

Pachur, T. (2002). *Judgements of health risk frequencies: On people's sensitivity to information validities, the effect of personal experience, and the plausibility of ecological judgement models.* Unpublished Diploma Thesis. Berlin: Freie Universität.

Ross, L., Green, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279-301.

Sedlmeier, P., Hertwig, R. & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 754-770.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin and Review*, *1*, 2-28.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129-138.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *6*, 105-110.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207-232.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806-820.

Symbolizing Quantity

Praveen K. Paritosh (paritosh@cs.northwestern.edu)
Qualitative Reasoning Group, Department of Computer Science,
Northwestern University, 1890 Maple Ave,
Evanston, IL 60201 USA

Abstract

Quantities are ubiquitous and an important part of our understanding about the world – we talk of engine horsepower, size, mileage, price of cars; GDP, population, area of countries; wingspan, weight, surface area of birds, and so on. In this paper, we present cognitively plausible symbolic representations of quantity and principles for generating those representations. Bringing together evidence in linguistics and psychology, we argue that our representations must make two kinds of distinctions – *dimensional*, those that denote changes of quantity, e.g., large and small; and *structural*, those that denote changes of quality, e.g. boiling point and poverty line. We present results of a pilot experiment that suggests that there is a significant agreement between people about the dimensional distinctions. We then describe a computational model CARVE, which is a system that learns to make dimensional and structural distinctions on quantities by being exposed to examples.

1 Introduction

Our knowledge about quantities is of various kinds – we understand that there are `Expensive` and `Cheap` things, that Canada is `larger` (in area) than the USA, that basketball players are usually `tall`, that the `boiling point` of water is `100 degrees Celsius`. A key part of such knowledge seems to be a *symbolization* of the space of values that a quantity can take. By symbolization, we mean identifying and naming intervals and points in the space of values of a quantity. Some examples include `tall` and `short` for the quantity of height of people; `poverty line, lower class, middle class` and `upper class` for income of people; `freezing point` and `boiling point` for the temperature of water.

These symbolizations and their mapping onto quantitative values seem to be determined by a mixture of personal experience (e.g., what I consider to be `spicy` in regards to food), society (e.g., middle class), science (e.g., phase transitions). Some are task-specific – one makes more distinctions than freezing and boiling for bath water. Furthermore, some of these symbolizations have been said to be vague [Varzi, 2003], in the sense that it is not possible to tell exactly at what value of height one becomes tall, and is not tall if any less than that. Given these concerns, finding systematic principles behind such symbolizations seems to be a daunting task, and has not been tackled head-on in cognitive science. That said, there is a vast literature that bears on these issues. In this paper, we address the following two fundamental questions about people's knowledge of quantities –

1. *Representational*: What do our representations of quantity look like? Or, what representational machinery is needed to make the distinctions that we do?
2. *Computational*: How are these representations built with experience?

Large scale knowledge representation efforts like Cyc [Lenat and Guha, 1989] refer to quantities either purely numerically, or using ad hoc representations. Most existing computational models of retrieval and similarity cannot use numerical representations [Falkenhainer et al, 1989; Holyoak and Thagard, 1989; Hummel and Holyoak, 1997; Goldstone and Rogosky, 2002], leading to quantitative information being ignored in computation of similarity. There are models in case based reasoning [Ashley, 1990; Leake, 1996; Ram and Santamaria, 1997] that use numeric information, but they employ ad hoc similarity metrics that are not psychologically grounded. A major motivation of this work is to generate cognitively plausible symbolic representations of quantity that will enhance computational models of similarity, retrieval and generalization.

The rest of the paper proceeds as follows: We next present relevant research from Linguistics, Psychology, Qualitative Reasoning, and models of similarity and retrieval, which provide both background and motivation. Section 3 reports results of a pilot experiment measuring just how vague our notion of large, medium and small is. Section 4 proposes an answer to the representational question above. Section 5 describes CARVE, a computational model for building such representations. We conclude with future work in section 6.

2 Background and Motivation

2.1 Linguistics

In language, one of the ways these symbolizations get represented is by relative adjectives like `large` and `tall`. Relative adjectives are different from absolute adjectives like `rectangular`, `red` and `married` in the sense that (1) they can imply varying degrees of the property in question, as opposed to all-or-none for the absolute adjectives, and (2) their meaning varies with context, e.g., `tall` means different things in context of men and buildings.

These adjectives have been variously called degree, relative, gradable or dimensional adjectives [Bierwisch 1987]. Here we will stick to the term dimensional adjectives, emphasizing our focus on those that denote quantity. It has been proposed that dimensional adjectives denote measure functions that maps from objects to quantity values/ intervals [Kennedy, 2003]. It has long been

recognized by linguists that dimensional adjectives convey an implicit reference to a norm or a standard associated with the modified noun [Sapir, 1944]. This implies two steps in interpreting a phrase like "a large *x*" where *x* can be a country/ insect/ etc.: (1) *x* establishes a *comparison class*. A comparison class is a set of objects that are in someway similar to *x*. For instance, in some cases, this comparison class might be the immediate superordinate of the subject [Bierwisch, 1971]. How to obtain the comparison class is an open question. Staab and Hahn (1998) propose a computational model that uses knowledge about correlations to determine comparison classes on the fly. (2) Once the comparison class has been found, a *standard of comparison* is computed for the class. It is usually believed that this is the norm value of the property for the comparison class, but Kennedy (2003) observes that it can also be the minimum or maximum (e.g., full and open).

The norm in step 2 has not been spelled out in this literature. In cases where we are referring to stable taxonomic categories like insects and countries, it is believed to be some kind of central tendency. But clearly, it is more than a central tendency, since that would imply that most things in this world will be either large or small, as not many will be exactly equal to the norm.

2.2 Psychology

2.2.1 Context sensitivi ty

Rips (1980) considers two hypotheses about how absolute and relative adjectives might be stored in memory – Pre-Storage and Computational model. For absolute adjectives like `married` and `pink`, he accepts the pre-storage model, where these predicates are stored with the concept they apply to. But because of context dependence of relative adjectives like `big`, e.g., in, "Flamingos are big", he argues against storing these predicates in memory. We might have a predicate `pink` attached to flamingo, but in order to decide a flamingo is larger than an eagle, we might need a predicate `is-larger-than-an-eagle` associated with flamingo, which then deescalates into having infinitely many of those like `is-larger-than-turnips` and so on. He also observes that relative adjectives don't propagate in a isa hierarchy – e.g., Grasshoppers are large insects does not imply Grasshoppers are large animals, but if you replace 'large' by 'green', the implication is right. He then shows reaction time and error rates for verifying the truth of statements containing relative adjectives which supports a different model. In his 'computational model' no relative information is stored. Attached to every predicate is a normal value, e.g. with insects, a normal size of quarter inches. An object is called large if it is bigger than this normal size. Once again the problem is that just storing the norm doesn't tell you when the object can be classified as large. The representation that we propose in section 4 solves his concerns with pre-storage models.

2.2.1 Reference Points

The psychological reality of such special reference points on the scale of quantity has been shown in various domains. Rosch (1975) argued for the special status of such "cognitive reference points" by showing an asymmetry – namely that a non-reference stimulus is judged closer to a reference stimulus (e.g., the color off-red to basic-red) than otherwise, while such relationship between two non-reference stimuli is symmetric. Existence of landmarks to organize spatial knowledge of the environment, similar asymmetries [Holyoak and Mah, 1984 among others]. Other relevant psychological studies that support the existence of reference points come from categorical perception [Harnad, 1987] and sensitivity to landmarks [Cech and Shoben, 1985]. Brown and Siegler (1993) proposed the *metrics and mappings* framework for real-world quantitative estimation. They make a distinction between the quantitative, or metric knowledge (which includes distributional properties of parameters), and ordinal information (mapping knowledge).

2.2.2 Models of Retrieval, Similarity and Generalization

There is converging psychological evidence for structured models of retrieval, similarity and generalization.

The structure-mapping engine (SME) [Falkenhainer *et al*, 1989] is a computational model of structure-mapping theory [Gentner, 1983]. Given two structured propositional representations as inputs, the *base* (about which we know more) and a *target*, SME computes a *mapping* (or a handful of them). MAC/FAC [Forbus *et al*, 1995] is a model of similarity-based retrieval, that uses a computationally cheap, structure-less filter before doing structural matching. It uses a secondary representation, the content vector, which summarizes the relative frequency of predicates occurring in the structured representation. The dot product of content vectors for two structured representations provide a rough estimate of their structural match. SEQL [Kuehne *et al*, 2000] provides a framework for making generalizations based on comp uting progressive structural overlaps of multiple exemplars.

One limitation of these models – and of other models of analogical processing (e.g., ACME [Holyoak and Thagard, 1989, LISA [Hummel and Holyoak, 1997], ABSURDIST [Goldstone and Rogosky, 2002]) – is that they do not handle numerical properties well:

Retrieval: Just as `Red` occurring in the probe might remind me of other red objects, a bird with wing-surface-area of 0.272 sq.m. (that is the Great black-bucked gull, a large bird) should remind me of other large birds. This will not happen in the current model, unless we abstract the numeric representation of wing-surface-area to a symbol, say, `Large`.

Similarity: A model of similarity must be sensitive to quantity.For example, in current matchers, two cars which are identical in all dimensions have the same similarity as two that differ in some dimensions, if other aspects of their representations are identical.

Generalization: A key part of learning a new domain is acquiring the *sense of quantity* for different quantities. E.g., from a trip to the zoo, a kid probably has learnt something about sizes of animals.

A symbolic and relational representation of the kind we propose here would make models of analogical processing more quantity-aware.

2.3 Qualitative Reasoning

Qualitative reasoning research seeks to understand human-like commonsense reasoning without resorting to differential equations and real-valued numbers. There is a substantial body of research in QR that has shown that one can, indeed, do powerful reasoning with partial knowledge. Qualitative reasoning has explored many different representations: status algebras (normal/abnormal); sign algebra (– , 0, +), which is the weakest representation that supports reasoning about continuity; quantity spaces, where we represent a quantity value by ordinal relationships with specially chosen points in the space; intervals and their fuzzy versions; order of magnitude representations; finite algebras, among others. While these representations are very promising for cognitive modeling, there has been little psychological work to date on this.

3 Experiment

We conducted a pilot experiment to see how much people agreed on what they would call large, small or medium. We expected agreement across subjects on their labeling. Furthermore, we expected to find out how people go about mapping these symbols to quantity values in a specific scenario – being presented with all the examples at once. And if people indeed agreed on their partitioning, then we expected to gain insight about where they drew the boundaries.

Method

The experiment consisted of two tasks – *Size Labeling* task and *Country Naming* task. In the size labeling task, subjects were presented with an outline political map of Africa. The countries were numbered from 1 through 54, and at the bottom of the map were 54 numbered blanks. They were given the following instruction – "On the following page you will find a map of Africa. All the different countries are shown and numbered. For each country, we want you to think if you will call it LARGE, MEDIUM or SMALL on the basis of size (land area) as shown in the map. Below the map you will find numbered index of all the countries on the map. Please place your answer (LARGE/ MEDIUM/ SMALL) in the blank next to it. Please fill out all the blanks."

At the end of this task, they did the country naming task. Here they were presented another copy of the map, and were told to name as many of the countries as they could. The participants were 19 graduate students at Northwestern University.

Results and Discussion

We found significant agreement across the subjects. Subjects could correctly name very few countries (mean 6 out of 54 countries, *sd* = 6.5). This suggests that prior knowledge should be irrelevant, and their judgments were based on examining the map.

To see how much subjects agreed about their choices, we extracted the most frequent choice for each country, and the percentage of times that was chosen across subjects (e.g, for both Seychelles and Algeria this is 100%, as the most frequent choice was always picked, for Kenya it is 79% which is how often it was called medium). In figure 1 we show the most frequent, second most frequent and the least frequent choice and how often they were chosen. The most frequent choice was chosen an average of 81.2% of the times, and the second most frequent choice was chosen 18.5%, and the least frequent 0.3% of the times. The difference between most frequent choice and the second most frequent choice is statistically significant ($t(53)=12.92$, $p<0.01$).

Subjects seem to do the task in a clustering fashion. They would pick either small/large and start marking out the clearly small/large countries, then countries at the other end of size and then consider the cases in between.

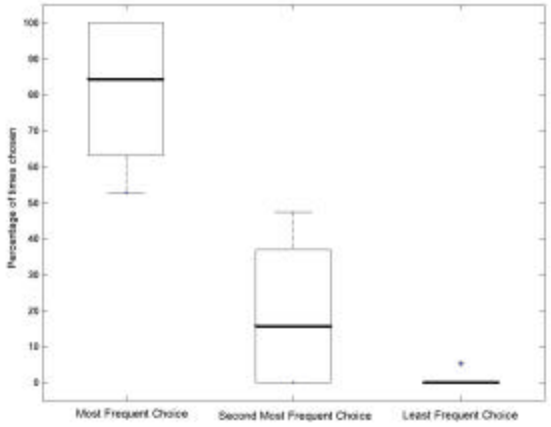


Figure 1. Agreement across subjects on their most frequent choice. The most frequent choice is 81.2%, significantly higher than the second and third chosen size labels.

4 Representation

A representation of quantity allows us to make certain distinctions – numbers allow us to make too many, and dividing the range of values into two equal sized parts doesn't necessarily provide useful distinctions. Representations do not arise in vacuum. They are molded by the kinds of reasoning tasks we perform with them (reasoning constraints), and the things we are trying to represent (ecological constraints). We propose representations based on existing evidence and arguments from these constraints.

4.1 Reasoning Constraints

The three distinct kinds of reasoning tasks involving quantities are –

1. Comparison: These involve comparing two values on an underlying scale of quantity, e.g., "Is John taller than Chris?" Our knowledge of how the quantity varies (its distribution), and linguistic labels like `Large` and `Small`, are but a compressed record of large number of such comparisons. The semantic congruity effect [Banks and Flora, 1977] is the fact that we are better and faster at judging the larger of two large things than the smaller of two large things. Part of the account from experiments involving adults learning novel dimension words, by Ryalls and Smith (2000) is the fact that in usage, we make statements like "X is larger than Y" more often than "Y is smaller than X", if X and Y are both on the large end of the scale.

2. Classification: These involve making judgments about whether a quantity value is equal to, less than or greater than a specific value, e.g., Is the water boiling?, Will this couch fit in the freight elevator?, etc. Usually, such classifications involve comparisons with interesting points (called *limit points* in QR) in the space of values for a quantity, where conditions on either side are qualitatively distinct. The metaphor of *phase transitions* describes many such interesting points, although such transitions in everyday domains are not as sharply and well defined as in scientific domains (consider poverty line versus freezing point).

3. Estimation: These involve inferring a numerical value for a particular quantity, e.g., How tall is he? What is the mileage of your car? This is the activity that has the strongest connection to quantitative scales – one can go a long way in accounting for the above two without resorting to numbers, but estimation involves mapping back to numbers [Subrahmanyam and Gelman, 1998]. Knowledge of interesting points on the scale might play an important role in estimation, for example in providing *anchors* to *adjust* from [Tversky and Kahmenan, 1974].

These tasks are not completely distinct – classification involves comparison, and estimation might be used in the service of classification. Two interesting aspects of our representations follow from these constraints:

1. Our representations must keep track of interesting points on the scale of quantity, to classify, as well as to estimate.
2. Labels like `large` ease making comparisons, as they setup implicit ordinal relationships (it is larger than most objects).

4.2 Ecological Constraints

Our representational framework must be capable of capturing the interesting ways in which a quantity varies in real-world instances of it. Below we present two different kinds of constraints on values a quantity can take –

1. Distributional Constraints: Most quantities have a range (a minimum and a maximum) and a distribution that determines how often a specific value shows up. For example, the height of adult men might be between 4 and 10 ft, with most being around 5-6.5ft. More than just the norm, we can usually talk about the `low, medium, high` for many quantities, which seems to be a qualitative summary of the distributional information. There is psychological evidence that establishes that we *can* and *do* accumulate distributions of quantities [refer to Malmi and Samson, 1983; Fried and Holyoak, 1984; Kraus *et al*, 1993; among others, for more]. Given a distribution of values for a quantity, the next question of how we partition these distributions has not been raised at all.

2. Structural Constraints: Quantities are constrained by what values *other* quantities in the system take, its relationship with those other quantities, via its relationships with them[1]. For instance, for all internal combustion engines – as the engine mass increases, the Brake Horse Power (BHP), Bore (diameter), Displacement (volume) increases, and the RPM decreases. These constraints represent the underlying mechanism, or causal model of the object. Limit points decompose values into regions where the underlying causal story is different (e.g., ice starting to melt, at the freezing point), which induces extremely important and interesting distinctions of *quality* on the space of quantity.

These two ecological constraints point us to the two different kinds of information about quantities, which must be parts of our representations –

1. Distributional information about how the quantity varies.
2. Its role in and relationship to the underlying structure/mechanism, and the points at which there are changes in underlying structure.

4.3 Proposed Representation

There are two kinds of distinctions that our representation of quantity must make –

1. *Dimensional partitions:* Symbols like `Large` and `Small`, which arise from distributional information about how that quantity varies.
2. *Structural Partitions:* Symbols like `Boiling Point` and `Poverty Line`, that denote changes of *quality*, usually changes in the underlying causal story and many other aspects of the objects in concern.

These partitions may manifest as intervals centered around a norm, or by boundaries demarcating transitions. Let's look at dimensional partitions in more detail. Dimensional adjectives like `large` depend upon the context. Consider area of African countries – in our experiment, people agree that Algeria is large, and Swaziland is small, Kenya is medium sized. We represent this as follows –

```
(isa Algeria
     (HighValueContextualizedFn
          Area AfricanCountries))
```

`High/Medium/LowValueContextualizedFn` are functions that take two arguments – a quantity and a context argument and return a collection of objects. So in the above example `HighValueContextualizedFn` denotes the

[1] Comic books, mythology, and fantasy, for example, have the freedom to relax this constraint – a character can be arbitrarily strong, large, small or be able to fly, even though the physical design of the character might not be able to support it.

collection of large African countries, and the isa statement says that Algeria is an instance of that collection. The `LowValueContextualizedFn` similarly lets us represent the negative end, for instance small and cheap.

4.3.1 Relationship to Fuzzy Logic

The dimensional partitions are reminiscent of linguistic variables in fuzzy logic [Zadeh 1965]. Fuzzy variables can take on values like `Large`, `Medium` and `Small`; and allow us to represent overlapping range of values for these symbols. Fuzzy logic thus provides a framework to represent what `Large` means. The specific mapping of `Large`-ness to area of countries, for instance, is a choice of the person building the representation, and is not in the scope of fuzzy logic. Our focus here is that mapping. So dimensional partitions are the answer to the question – what do people mean when they say "a large country," specifically, what is the mapping between `Large` and the values of area?

5 Computational Modeling

We are developing a computational model, called CARVE, as an account of the generation of both dimensional and structural partitions. At this writing, CARVE is partially implemented. The input to CARVE is a set of examples represented as collections of facts in predicate calculus. Countries are an interesting domain for testing CARVE as there are many quantitative parameters with rich causal and structural relationships[2]. The cases for each of these countries were built by extracting facts about them from the Cyc knowledge base. Additional quantitative facts about attributes like population, literacy, etc., were extracted from the CIA Factbook knowledge base [Frank *et al*, 1998] and added to these cases. There were on average 108 facts per case.

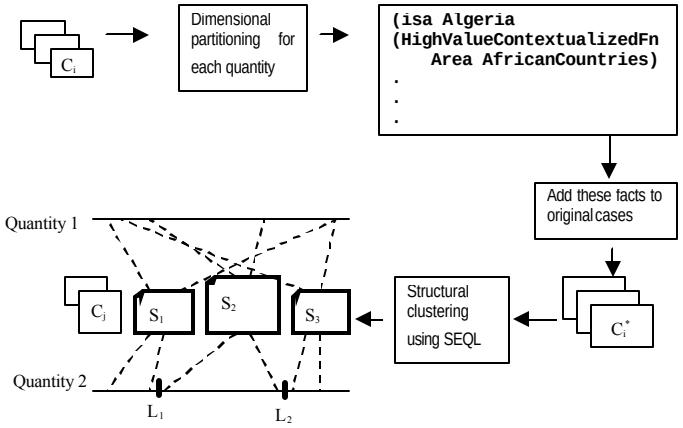


Figure 2. A schematic overview of how CARVE computes dimensional and structural partitions.

[2] Alas, not all of this rich structural knowledge is already represented in our knowledge bases.

Dimensional Partitioning

CARVE takes as input a set of cases. For each quantity, we extract all the numeric values for it in our input cases. Given these values, the job of the dimensional partitioning step is to find three partitions, corresponding to `Low`, `Medium` and `High` ranges of the values that the quantity takes.

These partitions are currently generated using a k-means clustering algorithm. It is possible to plug in different heuristics that partitions the values into ranges of values. Heuristics based on central tendency and percentiles do not work for zipf like distributions which we see in many of the quantities (e.g., GDP, population, area) associated with countries. For such distributions, means and variances are not intuitively meaningful at all.

The k-means clustering algorithm fits with what people did in our pilot experiment. On an average across subjects, the dimensional partitions computed by CARVE agreed with people 74% of the times (sd=27). More empirical data is needed to conclude what set of heuristics people use to make these partitions, and when they work. We believe that depending upon the distribution of data, people will use different partitioning strategies. The clustering scheme used is useful across different kinds of distributions and can be used incrementally without a priori knowledge of distributions.

For each fact about the value of a quantity, we then add a `High/Medium/LowContextualizedValueFn` to the case depending upon which range that numeric value fell in. These facts are used in the next step.

Structural partitioning

SEQL [Skorstad *et al*, 1988; Kuehne *et al*, 2000] provides a framework for making generalizations based on computing progressive structural overlaps of multiple exemplars. The goal of structural partitioning is to find the structural clusters in the cases (for instance, groups of developing and underdeveloped nations) and project these clusters on to various quantity dimensions. The cases produced at the end of the dimensional partitioning step are given as input to SEQL. In figure 2, we see the output of SEQL as three generalizations S_1, S_2 and S_3 and some leftover cases that did not fit any of those. Let's consider two quantities $Quantity_1$ and $Quantity_2$. The projection of a cluster on a quantity is the range of values for that quantity in the cluster. For $Quantity_1$, we see that the projections from all the three generalizations overlap. On the other hand, the projections of the generalization on $Quantity_2$ are non-overlapping. We have marked by L_1 and L_2 the boundaries for these ranges. Notice the predictive power of knowing that for a specific case the value of $Quantity_2$ is less than L_1. We not only know about the quantity value, but about the generalization to which the case belongs, and so can predict a lot of other causal properties of it. For instance, when you know that a country is a developing country, there are rich causal predictions you can make.

The algorithm above has been implemented in CARVE. Unfortunately, because of the lack of rich causal/ relational

knowledge in the cases, it does not yet find any interesting structural partitions. Structural partitions are a reflection of our deep understanding of the causal and correlational structure of examples. In science, phase transitions, and structural distinctions in socio-economic dimensions were not easily discovered. We hope that by adding more knowledge we will get better structural partitions.

6 Conclusions and Future Work

Based on cognitive and linguistic evidence, and arguments from reasoning and ecological constraints, we presented symbolic representations for quantity. We find significant agreement between subjects on dimensional partitions. We presented a computational model for automatically generating these representations.

Currently all the cases are given as input to CARVE. One important way to extend this will be for it to incrementally build and update its representations. Further, we need to create rich structured cases with causal and correlational information and test CARVE.

Acknowledgements

This research is supported by the Computer Science Division of the Office of Naval Research. The author would like to thank Ken Forbus, Dedre Gentner, Chris Kennedy, Lance Rips, Jason Jameson, Tom Hinrichs, Sven Kuehne and Julie Saltzman for insightful comments and discussion on the work presented here.

References

Ashley, K.D. (1990). Modeling Legal Argument, MIT Press, MA.

Banks W. P., and Flora J. (1977). Semantic and Perceptual Processes in Symbolic Comparisons. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 278-290.

Bierwish, M. (1967). Some Semantic Universals of German Adjectivals. *Foundations of Language*, **3**, 1-36.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review,* **100**(3), 511-534.

Cech, C. G. and Shoben, E. J. (1985). Context Effects in Symbolic Magnitude Comparisons. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **11**, 299-315.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence,* **41**, 1-63.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science,* **19**(2), 141-205.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence,* **24**, 85-168.

Fried, L. S., and Holoyak, K. J. (1984). Induction of Category Distributions: A Framework for Classification Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 234-257.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Frank, G.; Farquhar, A.; & Fikes, R. Building a Large Knowledge Base from a Structured Source: The CIA World Fact Book. Knowledge Systems Laboratory, 1998.

Goldstone, R. L. and Rogosky, B. J., (2002). Using relations within conceptual systems to translate across conceptual systems, Cognition, 84, 295-320.

Harnad, S. (1987). *Categorical perception.* Cambridge: Cambridge University Press.

Holoyak, K. J., and Mah, W. A. (1984). Cognitive Reference Points in Judgments of Symbolic Magnitude. *Cognitive Psychology*, **14**, 328-352.

Holyoak, K. J. and Thagard, P. R. (1989). Analogical Mapping by Constraint Satisfaction, Cognitive Science, 13, 295-355.

Hummel, J.E and Holyoak, K. J. (1997). Distributed representations of structure: a theory of analogical access and mapping, Psychological Review, 104, 427-466.

Kennedy, C. (2003). Towards a Grammar of Vagueness. Presented at the *Princeton Semantics Workshop*, May 17, 2003

Kraus, S., Ryan, C. S., Judd, C. M., Hastie R., and Park, B. (1993). Use of mental frequency distributions to represent variability among members of social categories. *Social Cognition*, **11**(1), 22-43.

Kuehne, S., Forbus, K., Gentner, D. and Quinn, B.(2000) SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of CogSci 2000*.

Leake, D. (Ed.) 1996. *Case-based Reasoning: Experiences, Lessons and Future Directions*, MIT Press.

Lenat, D. B. and Guha, R. V. (1989). Building large knowledge-based systems: Representation and inference in the Cyc project, Addison-Wesley, Reading, MA.

Malmi, R. A., and Samson, D.J. (1983). Intuitive Averaging of Categorized Numerical Stimuli, *Journal of Verbal Learning and Verbal Behavior*, **22**, 547-559.

Malt, B. and Smith, E. (1984). Correlated Properties in Natural Categories. *Journal of Verbal Learning and Verbal Behavior,* **23**(2), 250-269.

Paritosh, P.K. and Forbus, K.D. (2003). Qualitative Modeling and Similarity in Back of the Envelope Reasoning. In *Proceedings of the 25th Cognitive Science Conference.*

Ram, A. and Santamaria, J.C. (1997). Continuous case-based reasoning. *Artificial Intelligence,* **90**, 25-77

Rips, L. J., and Turbull, W. (1980) How big is big? Relative and absolute properties in memory. *Cognition*, **8**, 145-174.

Rosch, E. (1975). Cognitive Reference Points. *Cognitive Psychology*, **7**, 532-547.

Ryalls, B. O. and Smith, L. B. (2000). Adults Acquisition of Novel Dimension Words: Creating a Semantic Congruity Effect, *Journal of General Psychology*, **127**(3), 279-326.

Staab, S. and Hahn, U. (1998). Grading on the Fly. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society,* Madison, WI.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases, *Science*, **185**, pp 1124-1131.

Varzi, A. C. (2003). Vagueness, In Encyclopedia of Cognitive Science, Macmillan and Nature Publishing Group, London.

Zadeh, L. (1965). Fuzzy Sets, *Information and Control*, **8**, 338-353.

Distortions of perceptual judgement in diagrammatic representations

David Peebles (D.Peebles@hud.ac.uk)
Department of Behavioural Sciences, University of Huddersfield,
Queensgate, Huddersfield, HD1 3DH, UK

Abstract

An experiment is reported which investigates the distorting e ects of various graphical features in three different diagrammatic representations of the same information. The experiment revealed signi cant distortions in users' perceptual judgements of distance both between the di erent diagrams and within each diagram. The results of the experiment are interpreted as indicating the crucial role of anchor points such as axis tick marks along dimensions.

Introduction

Presenting large, complex data sets to a broad and non-specialist audience is a challenging task if one is to be sure that the information is to be interpreted in an appropriate manner. Choosing the most suitable representation for a particular communicative goal is an important aspect of the task and designers of information artefacts must be aware of how the low-level visual features of individual graphical representations can facilitate or hinder the interpretation of information.

Many studies have shown that the perception of one or more graphical elements in a figure can be distorted by the relationships between them (see, e.g., Deręgowski, 1980; Schiffman, 1995). A famous example is the Müller-Lyer illusion shown in Figure 1a which illustrates how perceptual judgements of line length can be distorted by the acuity of angles subtended by connecting lines. In this illusion, the line on the right of Figure 1a is perceived to be shorter than that on the left, although their lengths are actually the same.

Distortions in the perception of line length can also be caused by a number of so-called *contrast illusions*, for example the parallel lines illusion (Jordan & Schiano, 1986; Schiano, 1986) in which viewers of two parallel lines can perceive the lengths of the lines to be closer than they actually are (assimilation) or more different than they actually are (contrast), depending on the ratio of the line lengths and the distance between them. An example is shown in Figure 1b in which the perceived lengths of the lines in each pair are distorted by the length of the line next to it so that viewers see the lengths of the two paired lines as being more similar than they actually are. This has the effect of distorting the perceived length of the right-most line in each pair to make that on the right of the figure seem shorter than that on the left when, in fact, their lengths are the same.

Previous research has shown that visual illusions can have a strong effect on people's perceptual judgements in commonly used diagrams (e.g., Poulton, 1985). Zacks, Levy, Tversky and Schiano (1998) studied, among other things, the effect of the length of neighbouring elements on judgements of bar height and magnitude comparison in bar charts and their experiments demonstrated that the accuracy of participants' judgements depended of the relative height of the target bar and neighbouring graphical elements.

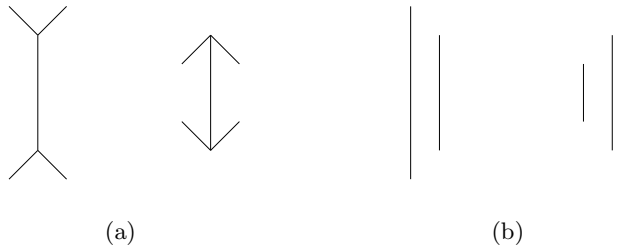


Figure 1: Two visual illusions affecting the perception of line length: (a) The Müller-Lyer illusion, (b) The parallel lines illusion.

In addition to issues relating to visual features, diagram designers must also be mindful of user familiarity. Using a form of diagram that is unfamiliar to a target audience can be problematic as users must expend additional cognitive effort to learn the new representation, something that could discourage engagement with the diagram or result in misinterpretation. Employing a novel representation can be justified, however if it can be demonstrated that the particular representation of information provided by the diagram facilitates a specific set of interpretive tasks. For example, in a series of experiments, Peebles and Cheng (2001, 2002, 2003; Peebles, Cheng, & Shadbolt, 1999) compared the representational and computational properties of two types of Cartesian coordinate (x,y) graph which, according to participant ratings, varied significantly in terms of their familiarity. Our experiments demonstrated that users of the less familiar graph type were able to retrieve information and solve certain problems significantly faster than users of the more familiar form. Eye movement and modelling analysis showed that this was because the graphical format of the unfamiliar representation facili-

tates certain basic reading and lookup procedures.

These issues are not only of academic concern. They have been brought into the public arena recently by the decision of the UK government to publish its national police performance data in the form of a set of diagrams that are relatively unfamiliar to the general public. With much media attention and at a reported cost of £70,000 (approx. US $128,688), the UK government developed the *performance monitor* (Police Standards Unit, 2003, 2004[1]), a variation of a diagram otherwise known as a *spidergram*, *radar* or *kiviat* chart. The purpose of the performance monitor is to present in summary form performance data for individual police forces in five key areas or "domains" (citizen focus, promoting public safety, resource usage, investigating crime, and reducing crime) and to allow easy comparison with average performance computed from a set of police forces most similar to the individual force in terms of socio-economic, demographic and geographic makeup.

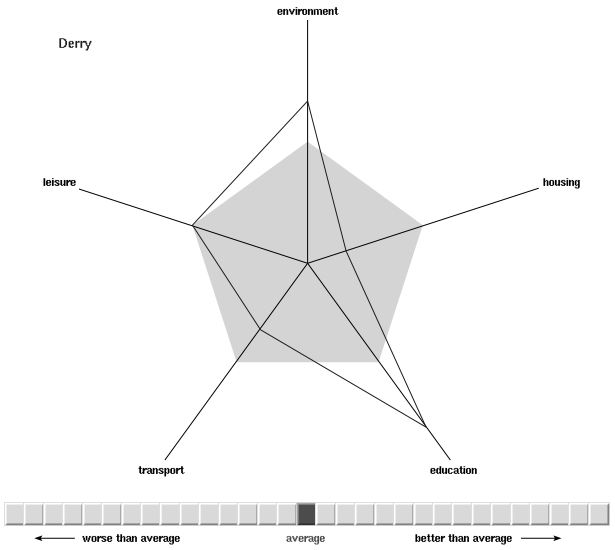


Figure 2: Kiviat chart used in the experiment.

An example of such a diagram is shown in Figure 2. This diagram is taken from the experiment reported here but it is identical in form to the police performance monitor. The subject matter of the diagrams was changed for the experiment to the (fictitious) performance of UK local authorities in five domains (the environment, housing, education, transport and leisure), each of which is indicated by a point on a spoke. The points are connected by straight lines to form a pentagon and the regular shaded pentagon represents the average performance of a set of most similar local authorities. Better performance is shown further out from the centre.

In the most recent version of the police performance report, the central performance monitor diagram has been augmented with five bar charts that illustrate the spread of performance for the most similar police forces in each of the domains. The bar charts are similar in form to the one displayed in Figure 3. In the police performance bar charts, each bar represents the value on that domain of one of the forces from which the average has been computed.

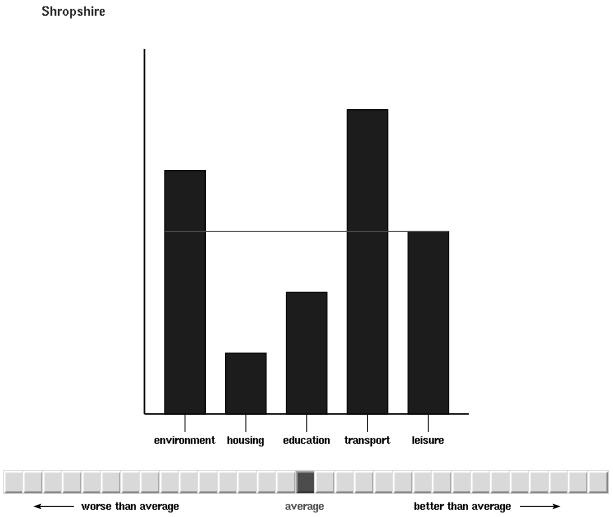


Figure 3: Bar chart used in the experiment.

One striking feature of both the kiviat and bar charts used in the police performance report is the lack of any tick marks on the spokes or axes. Usually the purpose of a scale of numbered tick marks on a chart or graph is to provide numerical values relating to locations in the chart. When numerical values are not deemed necessary, (perhaps because the purpose of the chart is simply to display relative magnitudes), tick marks still provide an objective reference frame within which to compare lengths. Without such a reference frame, it may be the case that perceptual judgements of quantities such as line length become more susceptible to distortion by visual illusions.

For example, in kiviat charts, the two lines connecting a point on a spoke with the two points on the adjacent spokes form a wide range of shapes and angles. In the absence of anchoring tick marks on the spokes, it may be the case that perceptual judgements of distance will be distorted by these angles and shapes by processes analogous to those involved in the Müller-Lyer illusion. Similarly, in bar charts, the lack of tick marks may also permit distortions in perceptual judgements of distance to occur because of the parallel lines illusion.

Another widely used diagram that shares many properties with bar charts is the line graph (see Figure 4), the main obvious difference being that points plotted against the y-axis are joined by lines rather than being represented as the top of a column. In the context of this study, however, an important consequence of this difference is that judgements of distance would not be

[1] A hypertext version of the current Home Office document is available on the Web as `performancemonitors.html` at `http://www.policereform.co.uk/docs/`

susceptible to the parallel lines illusion in line graphs.

Experiment

The primary purpose of the performance monitors and bar charts is to allow a rapid visual comparison of an individual institution's performance with a meaningful average. This could be either at a global level (i.e. to determine how much better or worse than average the institution is overall), or at the level of specific domains. The purpose of the experiment reported here is to determine whether the perceptual judgement of this distance for a particular target domain is affected by the values of the surrounding domains.

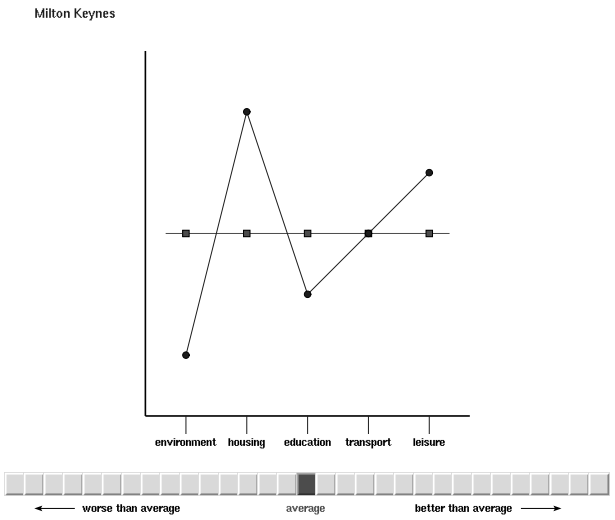


Figure 4: Line graph used in the experiment.

Method

Design The experiment was a mixed design with one between-subjects variable and two within-subjects variables. The between-subjects variable was the type of diagram used (kiviat chart, bar chart, or line graph). The within-subjects variables were the value of the target domain that subjects were required to rate and the values of the two domains adjacent to the target domain.

Participants Sixty-three members of staff from the University of Huddersfield took part in the experiment. The occupations of participants varied from academic, clerical and technical positions to graduate students.

Materials The experiment was conducted using three identical PC computers with 17-inch (43-cm) displays. The stimuli used in the experiment were diagrams similar to those in Figures 2–4. The information content of the diagrams was the performance of 150 UK local authorities across five domains.

In order to generate a manageable range of values, the spokes of the kiviat chart and the y axes of the bar chart and line graph were divided into six equally sized sections numbered 0 to 6 (although these divisions or numbers were not visible to the participants). The numbers 0 and 6 were situated at the bottom and top of the y axes and the centre and outermost points of the kiviat spokes respectively. Only the numbers 1 to 5 were used in the experiment and the locations of these on the diagrams can be seen in Figures 2–4. For example in Figure 2, Derry council has a housing value of 1, a transport value of 2, a leisure value of 3, an environment value of 4, and an education value of 5. The locations of the values on the y axes of the bar chart and line graph are illustrated in Figure 3, where Shropshire council has a housing value of 1, an education value of 2, a leisure value of 3, an environment value of 4, and a transport value of 5.

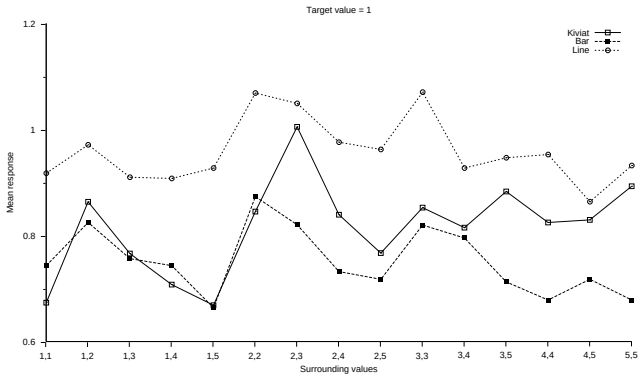


Figure 5: Mean response to target 1, all three diagrams.

The average value was the number 3 located at the centre of the y axes and kiviat spokes. In the bar chart this was represented by a horizontal red line and in the line graph as the same red line with red squares as markers (to conform to the format of the line graph). In the kiviat chart, the average was represented by a red regular pentagon formed by joining the centre points on the five spokes. This produced a kiviat chart identical to those used in the police performance document.

Below each diagram was a scale consisting of 31 buttons. The centre button in the scale was the same red colour as the average marker on the diagrams and underneath it was written the word "average" in red. The 15 buttons on either side of the centre button allowed the scale to be divided into six equally sized units, each containing four buttons. Below the scale were two arrows indicating that increases in performance were represented by buttons further to the right of the scale.

To test the full range of target and neighbouring lengths, each of the five target values was combined with the 15 possible permutations of two adjacent values (1,1; 1,2; 1,3; 1,4; 1,5; 2,2; 2,3; 2,4; 2,5; 3,3; 3,4; 3,5; 4,4; 4,5; 5,5) to create a total of 75 triplets.

In the kiviat charts, each domain spoke has an adjacent domain on either side but in the bar charts and line graphs, two domains (environment and leisure) have only one adjacent domain. To ensure that the target domain on each trial had an adjacent domain on either side, therefore, if the target value was 1, 2, 4, or 5, the target

domain was selected randomly from housing, education, and transport, as these had two adjacent values in the bar and line graphs. If the target value was 3, however, the target domain was randomly selected from all five domains. The values of the two remaining domains not adjacent to the target domain were randomly allocated a value of between 1 and 5. Responses to the target value of 3 were not to be included in the analysis as they were expected to be rapid and accurate for all graph conditions as this value was marked on both the graph and the scale.

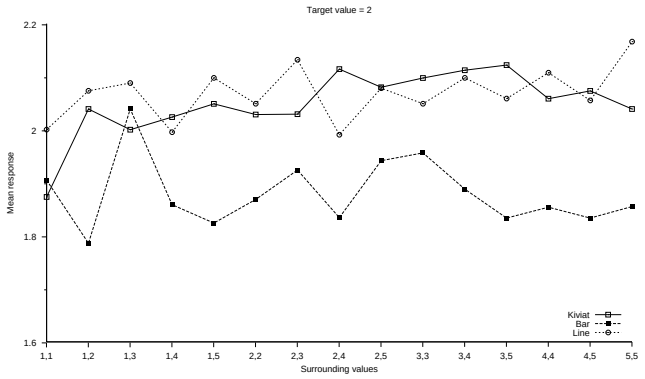


Figure 6: Mean response to target 2, all three diagrams.

Procedure Participants were randomly allocated to one of the diagram conditions. Before starting the task, participants were shown an example of the diagram they were to use and given as much time as they required to become familiar with it. The format of the example was the same for each diagram and was modelled closely on the format used in the Home Office document. When the participant had finished studying the example, the experimenter then explained the diagram further, explaining the subject matter, highlighting salient points and making sure that they were completely familiar with it. Participants were then told that on each trial of the experiment their task was to judge how much better or worse than average the performance of a particular authority was on a given domain. Participants were also shown how to enter their judgement by clicking the mouse on the scale below the diagram. It was stressed to participants that they should attempt to respond as rapidly but also as accurately as possible.

On each trial of the experiment, the target domain was first presented in the centre of the screen for 1500 ms, after which it was removed from the screen and replaced by a diagram. As soon as the participant had clicked the mouse cursor on one of the scale buttons the diagram was removed from the screen and, after a pause of 500 ms, the next target domain was presented for a new trial. Response times were recorded from the onset of the diagram to the mouse click on a scale button. Participants saw all 75 triplets twice—a total of 150 trials—in random order and were given the opportunity to take a brief, self-terminated break after 50 and 100 trials.

Results

Participants' responses were coded to re ect the underlying scale of the diagrams. A response click on the button to the extreme left of the scale was given the value 0 and each successive button was incremented by 0.2 to end at a final value of 6 at the extreme right of the scale.

An initial examination of the data revealed the existence of several outlying values that were not associated with a specific participant or condition but were sufficiently large to distort the mean for a specific cell. To reduce the in uence of these outliers, the 42 values in each cell were standardised and those cases at the extreme end of the distribution (i.e. with a z score greater than 3.29, $p < .001$, two-tailed test) were discarded (Tabachneck & Fidell, 2001). From the original set of 9450 data points, this procedure resulted in the removal of 165 cases (1.75%) of the response data and 128 (1.35%) cases of the RT data. Data from the target value = 3 condition were not included in the analysis because, as predicted, responses were almost entirely accurate for all of the diagrams.

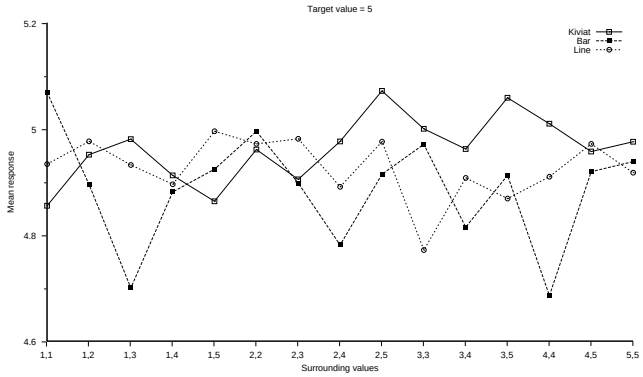


Figure 7: Mean response to target 5, all three diagrams.

An analysis of variance (ANOVA) was carried out on the response data. In the ANOVA, Mauchly's test of sphericity was significant for the target value (Mauchly's $W = .23$, $df = 5$, $p < 0.01$) and adjacent values (Mauchly's $W = .25$, $df = 104$, $p < 0.01$) so the more conservative Greenhouse-Geisser test was used in the analysis. The ANOVA showed that there was a significant main effect of the diagram used, $F(2, 59) = 15.48$, $p < .001$, the target value $F(1.66,97.66) = 4847.11$, $p < .001$ and the adjacent values, $F(9.87,582.49) = 3.96$, $p < .001$, together with significant interactions between the adjacent value and diagram, $F(19.75,582.49) = 2.84$, $p < .001$, and between target value and adjacent value, $F(16.07,974.94) = 2.51$, $p < .01$. The ANOVA also revealed a three-way interaction between diagram, target value and adjacent value, $F(32.13,974.94) = 1.56$, $p < .05$. Although response times were also recorded, due to lack of space, the data are not reported here. The main difference observed was that responses were slower in the kiviat condition than in the other two, probably due to the greater degree of unfamiliarity.

The complex effects revealed by the ANOVA are most

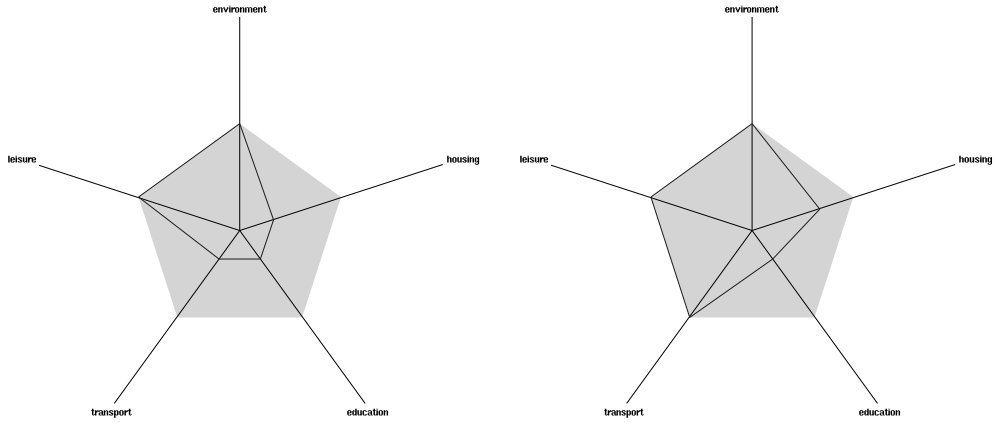


Figure 8: Kiviat charts illustrating a target value of 1 (education) with adjacent values 1,1 (left) and 2,3 (right).

clearly illustrated in Figures 5–7 which present the mean response for target values 1, 2, and 5 respectively as a function of the adjacent values. In these graphs, each labelled tick mark on the y-axis represents a button on the scale. The graphs show a large degree of variation in responses to individual target values both for individual diagrams and between the different diagram types.

One of the most striking differences in responses between the diagrams is shown in Figures 5 and 6, both of which show that participants viewing the line graphs consistently perceive target values of 1 and 2 to be closer to the average than bar chart users, despite the fact that the values are represented at exactly the same locations in the coordinate system in the two diagrams. This marked difference between the graphs is not present for target values of 4 and 5, however, and this may provide a possible explanation. Unlike points in a line graph, bars are attached to and proceed from the x-axis and so form a concrete object. When comparing the distance between the top of a bar with the mean line, therefore, participants' attention is drawn to the length of the bar in comparison to the height of the mean line, (rather than to the distance between them), which may serve to accentuate the perceived difference. In contrast, participants using the line graph may simply attempt the more accurate procedure of judging the distance between the points on the plotted and mean lines.

According to this account, the lack of such a major difference when the target values are 4 and 5 is because users of the bar chart are still judging the length of the bar but, this time are comparing it to the mean line below it. This is very similar to the procedure carried out by the line graph users in that both are judging the same distance. At the moment this is a plausible hypothesis that remains to be tested in a further eye movement study.

Figure 5 also illustrates the wide range of responses that users of the kiviat chart gave to the same target value. It is clear that the perception of the difference between the target value 1 and the mean value 3 is affected by values of the adjacent domains. To provide a clearer demonstration of this effect, two of the diagrams resulting in widely differing ratings are shown in Figure 8. In both diagrams, the target value of 1 is represented on the education domain. The chart on the left has adjacent values of 1,1, for which the mean rating is 0.67, whereas that on the right has adjacent values of 2,3, given a mean rating of 1.0. A t-test shows the difference between these ratings to be significant $t(60.11) = 3.10$, $p < .001$. Figure 5 shows that participants viewing the kiviat charts perceive the target value of 1 to be much closer to the average when surrounded by the values 2 and 3 than for any other combination but that the surrounding values of 1 and 1 were perceived as relatively far from the average. Although a precise explanation for this is still being considered, it is clear that the shape produced by the lines connecting the target value and the adjacent values has a distorting effect on viewers' perception of distance. Figure 5 shows that this effect is not a simple linear function in which greater adjacent values result in a larger perceived target value.

The distortion of perceived distance in the bar chart condition is perhaps best illustrated in Figure 7 which reveals a wide range of responses to the target value 5. As with the kiviat charts, the pattern of results does not conform to a simple analysis. At least one example, however, may be best explained by reference to the parallel lines illusion. The two bar charts in Figure 9 were given ratings at the extreme ends of the range. Both represent a target value of 5 on the education domain and the chart on the left has adjacent values of 1,1, for which the mean rating was 5.07, whereas that on the right has adjacent values of 4,4, given a mean rating of 4.69. A t-test shows the difference between these ratings to be significant $t(76.02) = 3.88$, $p < .001$. The target domain in the chart on the left of Figure 9 was perceived as being further away from the average line than that in right-hand chart. This can be explained in

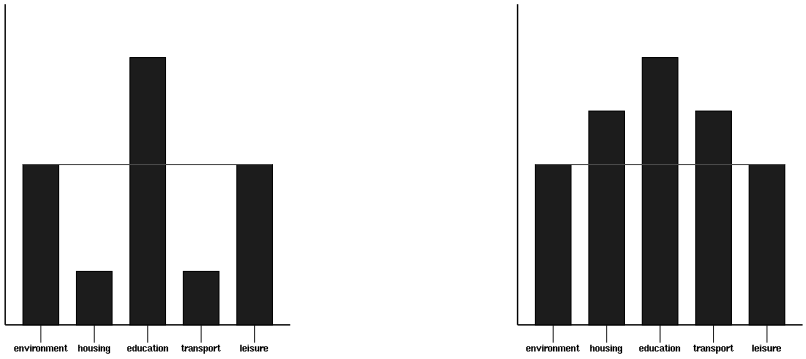


Figure 9: Bar charts illustrating a target value of 5 (education) with adjacent values 1,1 (left) and 4,4 (right).

terms of the parallel lines illusion as it re ects the phenomena of contrast and assimilation described earlier. The target domain in the left-hand chart is seen as being larger because it contrasts with two relatively small adjacent values. The target domain in the right-hand chart, however, was perceived as being smaller because viewers perceived the lengths of the target and adjacent bars to be closer than they actually are (assimilation). It is also interesting to note in Figure 7 that this pattern of responses is not found in the line graph condition.

Discussion

The results of this experiment provide concrete evidence of distortions in perceptual judgements of distance in two graphical representations of a type currently being employed to present public data in the UK. Specifically, the three examples clearly illustrate that simple comparative judgements between two points on a dimension can be significantly affected by the values of adjacent variables. Further work is required before firm conclusions can be drawn but from this initial analysis it seems that, as currently designed, the kiviat and bar charts in the UK government's police monitoring documents may be susceptible to the distorting effects highlighted here.

The use of anchor points, typically tick marks on axes, is generally seen as a way of facilitating the accurate reading of locations relative to a scale. Whether the incorporation of such anchor points into these diagrams reduces the distortions in distance judgements observed in this experiment is to be tested in a future study.

References

Deręgowski, J. B. (1980). *Illusions, Patterns, and Pictures*. London: Academic Press.

Jordan, K., & Schiano, D. J. (1986). Serial Processing and the parallel-lines illusion: Length contrast through relative spatial separation of contours. *Perception & Psychophysics*, *40*, 384–390.

Peebles, D., & Cheng P. C.-H. (2001). Graph-based reasoning: From task analysis to cognitive explanation, *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, *23*, 762–767.

Peebles, D., & Cheng P. C.-H. (2002). Extending task analytic models of graph-based reasoning: A cognitive model of problem solving with Cartesian graphs in ACT-R/PM, *Cognitive Systems Research*, *3*, 77–86.

Peebles, D., & Cheng, P. C.-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, *45*, 28–46.

Peebles, D., Cheng P. C.-H., & Shadbolt, N. (1999). Multiple processes in graph-based reasoning, *Proceedings of the Twenty- rst Annual Conference of the Cognitive Science Society*, *21*, 531–536.

Police Standards Unit. (2003). *Police Performance Monitoring 2001/02*. London: Home Office.

Police Standards Unit. (2004). *Police Performance Monitoring 2002/03*. London: Home Office.

Poulton, E. C. (1985). Geometric illusions in reading graphs. *Perception & Psychophysics*, *37*, 543–548.

Schiano, D. J. (1986). Relative size and spatial separation: Effects on the parallel-lines illusion. *Perceptional & Motor Skills*, *63*, 1151–1155.

Schiffman, H. R. (1995). *Sensation and Perception: An Integrated Approach* (Fourth Ed.). New York: John Wiley & Sons.

Tabachneck, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*, (Fourth edition). Needham Heights, MA: Allyn and Bacon.

Zacks, J., Levy, E., Tversky, B., & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied*. *4*, 119–138.

How do I Know how much I don't Know?
A cognitive approach about Uncertainty and Ignorance

Giovanni Pezzulo (pezzulo@ip.rm.cnr.it)
Istituto di Scienze e Tecnologie della Cognizione - CNR, viale Marx 15 – 00137 Roma, Italy
and Università degli Studi di Roma "La Sapienza" – piazzale Aldo Moro, 9 – 00185 Roma, Italy

Emiliano Lorini (e.lorini@istc.cnr.it)
Università degli Studi di Siena – via Banchi di Sotto, 5 - 3100 Siena, Italy

Gianguglielmo Calvi (calvi@noze.it)
Istituto di Scienze e Tecnologie della Cognizione - CNR, viale Marx 15 – 00137 Roma, Italy

Abstract

We propose a general framework for reasoning and deciding in uncertain scenarios, with possibly infinite source of information (open world). This involves representing ignorance, uncertainty and contradiction; we present and analyze those concepts, integrating them in the notion of *lack of confidence* or *perplexity*. We introduce and quantify the *strength of the beliefs* of an agent and investigate how he can do explicit *epistemic actions* in order to supply information lacks. Next we introduce a simple distributed game (RBG) and we use it as a testbed for comparing the performance of agents using the (classical) "expected utility maximization" and the "perplexity minimization" strategies.

Introduction

In an "open world" uncertainty and ignorance are difficult categories to deal with; how much can I be certain of a belief of mines? how much information there is that I have not considered and I should?

The first aim of the present work is to provide an analysis of epistemic dimensions such as *strength of belief*, *uncertainty*, *contradiction* and *ignorance* (or ambiguity). A special focus will be given to the third dimension. In Economical literature the notion of ignorance has been extensively investigated (Shackle, 1972) and ways to quantify it have been proposed (Shafer, 1976). In those approaches "lack of information" has been shown to affect the decision process and ambiguity aversion in subject has been identified; see Camerer & Weber (1992) for a review of the literature on decisions under ambiguity. We will argue in the following analysis that Ignorance is a subjective evaluation of actual lack of information on the basis of cognitive *evidential models*. The agent has a model (script) of his sources that allows him to evaluate that a certain type and a certain number of sources can provide *sufficient information* for reducing ignorance close to zero. In this way the strength of the belief and the (perceived) ignorance are two different measures, the second belonging to the meta-level. The second aim of the present work is to investigate the decision dynamics in an open world, with conflicting beliefs and multiple sources of information. We will formalize the process that leads the agent to acquire new information from the world (from witnesses) and that leads the agents to be "ready" to decide. We will claim that this process involves strength of beliefs that are relevant for deciding, as well as uncertainty and ignorance. The results of the current work are suitable e.g. for MAS environments, where an agent has to take decisions in open worlds.

The Red-or-Blue Card Game (RBG)

We introduce a simple distributed game that is suitable for Multi-agent system simulations as well as for human experiments: the agents (players) have to bid on the color of a card (red or blue) and they have many sources of information (their perception and potentially infinite witnesses); the game can last an indefinite number of turns. The bidding game is the following: a card is shown (very quickly) to the player; it can be either red or blue and the player has to bid on the right color (he starts with 1000 *Credits*). We assume that he cannot be totally sure of his own perception (e.g. it is shown very quickly, or the lights in the room are low), but he is able to provide a degree of certainty about the color. Before bidding he can ask for help to a (potentially) infinite number of witnesses that have observed the scene and provide the answer "red" or "blue" (without degrees of certainty); those new information can lead the agent to confirm or revise his beliefs. Asking a witness has not a cost in Credits but it costs 1 *Time*. Credits and Times can be aggregated in different ways.

When he decides that he is "ready", he can bid from 0 to 10 Credits on the color he wants. The true card color is shown: if he was right, he gains two times the bid; otherwise he loses the bid. The game lasts an indefinite number of turns; between the turns, depending on the result, the agent can revise the *reliability* he attributes to his sources: his perception and the witnesses (depending for example on the number of correct answers they provided) as well as his SCAI. Besides, his perception and the witnesses have *true reliability* values that determine the average correctness of their answers. True values are not known by the player; at the first game round they are initialized and they do not change during the game. At the end of the game the agent will collect a certain amount of Credits; a set of reliability values for his sources; a SCAI and he will have spent a

certain amount of Time. Using this game as a testbed, we compare different kinds of agents having many possible heuristics in order to collect information and Credits[1].

Theoretical Foundations

We present first an *evidential* notion of ignorance: ignorance is determined by the lack of belief sources. In our approach, following the cognitive approach of Castelfranchi (1995), the *strength of a belief* (i.e. how much I rely on one of my beliefs) depends on the *reliability* of its sources (i.e. the beliefs it is grounded on). Sources include: direct experience (such as perceptive evidences); information provided by other agents; reasoning (about other beliefs) and categorization (reasoning about classes and similarities). Since in an open world in principle there are infinite sources to take into account, the agent can never conclude that his own ignorance is zero. We propose in the present model a solution to the problem of ignorance quantification by identifying *Classes of Ignorance Acceptance* that reduce ignorance to finite values.

Uncertainty and contradiction have the same status of the notion of ignorance: they are meta-cognitive notions, i.e. agent's evaluations about his own "epistemic state". We are interested in this paper in investigating how those different epistemic states affect the decision process, and especially how they affect the way the agent decides to execute a pragmatic action (e.g. bet) or an epistemic one (e.g. query). For example, if the agent feels to be too much ignorant or uncertain he can decide to query, to bid a little amount, or not to bid at all. Here we describe the epistemic dimensions.

Ignorance

Intuitively ignorance depends on how much information I have with respect to how much it exists; in an open world there is a potentially infinite number of witnesses that have not been questioned; so if we calculate ignorance in this way the agent has always the maximum degree of ignorance. *The agent does not know how many witnesses he can consider at most* or better *he does not know how he can reduce his ignorance close to zero.* A qualitative and cognitive analysis is here required. Here we shift the issue to an evidential and subjective level[2]. We introduce the notion of *Structure of Classes of Acceptable Ignorance* (SCAI).

Classes of Acceptable Ignorance

Each agent has a SCAI that includes several *Classes of Acceptable Ignorance* (CAI) that include one or more sources (e.g. Witnesses), each having its reliability value. For instance, CAI_1 = (witness 1, witness 2, witness 3) could be one of those classes. Classes of acceptable ignorance can be intersected and unified (see Fig. 1): they have the normal properties of sets in set theory.

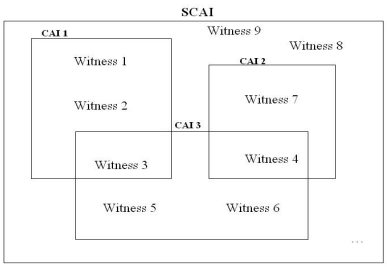


Fig. 1 Structure of Classes of Acceptable Ignorance (SCAI).

The agent knows that testing all witnesses in a given class is enough for making the ignorance acceptably close to value zero. Imagine for example that the agent wants to know if tomorrow will rain or will be sunny. He has several classes of acceptable ignorance. For instance he can believe that by acquiring information about tomorrow's weather from source 1 = "New York Times" and source 2 = "CNN" is enough for making ignorance acceptable. Moreover, the following points are crucial for understanding how the relation between SCAI and agents works.

A. The agent has explicit models (meta-level) of Classes of Acceptable Ignorance as shown in Fig.1. There are witnesses who are included in classes of acceptance but also witnesses who are not included in any class.

B. The agent can also make "queries" to witnesses who are not in the SCAI. Indeed the number of classes is finite even if, according to the agent, the set of witnesses he can make a query is indefinite. The agent can make a query to whatever witness even if this witness is not included in the structure of ignorance. In principle the agent can ask witness_10000 and he will always get an answer. Witnesses that are not in the structure do not have a value of reliability. A default value is assigned to them (through feature value assignment) whenever a query is made to them. After the query the witness belongs to the SCAI as a witness who is not included in any CAI (for example witness 8 in Fig.1).

C. The value of Ignorance is calculated at a meta-level whereas the value of reliability of a witness is calculated at a base-level.

[1]Since it is an "open world" (there are an infinite number of witnesses and an indefinite number of turns) it is not possible to perform full search. More, it is not possible to perform a long-term maximization because the agents don't know when the game will end (this condition is called "shadow of the future").

[2] Our notion of ignorance is very close to the notion of ambiguity identified in some recent economical and psychological literature where is stressed that decision making is affected by the decision maker's evaluation of his or her actual available information and competence to make judgments in specific domains (Heat & Tversky, 1991). Instead, our notion of ignorance is quite far from *Sample Space Ignorance* in Support Theory (Tversky & Koehler, 1994) where it is claimed that people do not follow the extensional logic of conventional probability theory. In Support Theory an agent can actually "ignore" actual information in the sense that he is not explicitly evaluating that evidences concerning a certain event e1 are also evidences concerning another event e2. Indeed it has been shown that unpacking (making information available for explicit evaluation) a compound event into disjoint components tends to increase the perceived likelihood of that event. An immediate implication is that unpacking an hypothesis and/or repacking its complement will increase the judged likelihood of that hypothesis.

Quantifying Ignorance through SCAIs

Class-Ignorance is given for each class at a certain point of a query sequence ($q_1,\ldots, q_n$) and is defined as the total number of witnesses in the class *minus* the number of tested witnesses in that class, weighted for the inverse of the total number of witnesses in the class**.**

Absolute-Ignorance is defined as the minimal value of Class-Ignorance among all CAIs.

Class-Ignorance (Class n, q_i) =
(n.wit. (Class n, q_i) - n.queried.wit. (Class n, q_i) $_{\text{Agent x, qi}}$)/ n.wit. (Class n, q_i) $_{\text{Agent x, qi}}$
Absolute-Ignorance (q_i) =
Min $_{\text{Class x}}$ (Class-Ignorance (Class x, q_i) $_{\text{Agent x, qi}}$)

We have already pointed out that after a query is made to a witness who does not belong to SCAI, the witness will be included in SCAI as a witness who is not included in any class (such as witness 8 and 9 in Fig.1). The measure of Absolute-Ignorance is not fixed: it depends on the single agent categorization and classes organization. That measure varies through learning, as new witnesses are added.

Uncertainty

Uncertainty is a measure of the difference between the value of strength of the belief "the card is red" and the value of strength of the belief "the card is blue". When the difference is 0 the value of uncertainty is maximal, when the difference is 1 the value uncertainty is minimum. This dimension takes into account the difficulty of deciding when the two strengths of beliefs are too close.

Contradiction

Contradiction is a (logical) inconsistency in a belief set; for example I can not believe consistently that (in the previous example) the ball in an urn is both red and black. In a normal (statistical) analysis there is contradiction if the sum of the two strengths of beliefs is more than 1. In the evidential approach the threshold of perceived contradiction (α) can be fixed at different values depending on cognitive biases (e.g. more or less contradiction tolerant).

If(Strength.belief(CardRed, q_i) +
Strength. Belief(CardBlue, q_i)) $\leq \alpha$ then
Perceived.contradiction(q_i) = 0

If(Strength.belief(CardRed,q_i) +
Strength.Belief(CardBlue,q_i)) $> \alpha$ then
Perceived.contradiction(q_i) =
(Strength.belief(CardRed,q_i) +
Strength.Belief(CardBlue,q_i)) – α

Perplexity

Ignorance, Uncertainty and Contradiction are three meta-level epistemic information that an agent can take into account in order to "decide if he is ready to decide". In order to model this kind of decisions we propose to integrate ignorance, uncertainty and contradiction in a single measure called **Perplexity** (i.e. lack of confidence). In calculating Perplexity, the three dimensions can be aggregated in different ways, depending on some more cognitive biases (e.g. Agents that are biased to consider ignorance, or contradiction, or uncertainty). The basic heuristic is summing them (and normalizing).

Value of Information: Epistemic Actions

An Epistemic Action (EpA) is *any action aimed at acquiring knowledge from the world*; any act of active perception, monitoring, checking, testing, ascertaining, verifying, experimenting, exploring, enquiring, give a look to, etc. (Castelfranchi & Lorini, 1998). The notion of epistemic action has been extensively considered both in psychology and in economics. The centrality of this notion comes from the fact that epistemic actions have a role in different cognitive functions. In the present model an Epistemic Action is always towards a witness (i.e. making a query). Epistemic Actions are directed either to reduce perplexity (or one of its dimensions) given a certain "perplexity aversion" threshold of the agent (first function); or to acquire new information in order to make a better decision (second function).

In both cases a value is assigned to epistemic actions. The first value is a measure of the capacity of a given witness of reducing perplexity: we call it *informativeness*. The second value is called *value of information* and has been extensively investigated in economical literature in the sense of "how much the agent is disposed to pay for obtaining that information?" In that approach a possible way to calculate the value of information is given with respect to utility functions. These two notions can lead to different decision strategies; in order to compare them, we have designed the simulative testbed "Red-or-Blue Card Game" (see above).

Considering the Sources

Strength of beliefs depends on its sources (perception; more or less reliable witnesses). Those sources are not all equal: in order to represent their relative contribute, we aggregate them using Fuzzy Cognitive Maps (Kosko, 1986). In order to represent the fact that there are diverging sources (and they aggregate in a different way with respect to converging ones) our FCMs have two "competing" branches for representing the competing beliefs "the card is red" and "the card is blue". FCMs are additive fuzzy systems with feedback, having nodes and edges. The weight of the nodes represents the strength of a belief (e.g. "I am pretty sure that the card is red"); the edges are weighted and they represent the impact of a belief over another. The FCM that we use can be seen as divided into two branches, each aggregating the values either for "red" or "blue". These nodes receive input from intermediate nodes ("perception for red" and "witnesses for red" the first; "perception for blue" and "witnesses for blue" the second); these edges are weighted by two fixed factors κ, λ representing the relative impact of perception and witnesses. The nodes "perception for red" and "perception for blue" assume either the value 0 or 1 depending on the perceptual input; their edges have the value of *perception reliability* (according to the agent). The "witnesses for red" and "witnesses for blue" nodes receive as input the information of the queried witnesses (either 0 or

1); the edges between each witness and "witnesses for <color>" have the value of the *witnesses' reliability*. There are also negative-weighted edges between the "red" and "blue" nodes, as well as for each source. In this way the contribute of diverging sources is modeled, because each positive evidence in a branch counts also as a negative one in the other branch. So, starting from the input values (the contributes of perception and the queried witnesses) the FCM calculates the final values for the strength of belief in "the card is red" and "the card is blue". There is not fixed "sum 1" between the two final values, so it is possible to model contradictory beliefs (that the agent can reduce performing epistemic actions). The FCM structure is the same for all the agents, but at each step it can be updated (e.g. modifying the impact of the edges, i.e. reliability values).

Player Agents and Decision Strategies

Here we describe three classes of decision strategies, implemented into three Agents. *Normative Agent* and *Satisficing Agent* do not use the notions of ignorance, uncertainty and contradiction. *Perplexity Reducing Agent* uses them in order to select the witness to be tested.

Normative Agent

A normative agent decides either to bet a certain amount of credits on a given option (either "the card is red" or "the card is blue") or to make a query to a specific witness as follows (this agent is not affected by perplexity).

The agent calculates the value of information obtainable from a given witness for all witnesses in the Structure of Classes of Ignorance Acceptance. *The value of information obtainable from witness z* is determined as: the average of the max value of expected utility given the information "the card is red" given by wit. z (which impacts on the agent's beliefs) and the max value of expected utility given the information "the card is blue" given by wit. z *minus* the max value of expected utility given the actual information. Afterwards the agent is able to decide. If the max value of information obtainable from witnesses is more than 0 then the agent decides to make a query to the witness who maximizes that value; otherwise he decides to bet a quantity y of credits on "the card is x" that maximizes his actual expected utility. We have not included the costs of making a query in the utility function (we assume only the cost in Time).

Potential-Chosen-Bet (q_i) = BET.yONx
such that
$\text{Max}_{x,y}$ (Strength.belief (x, q_i) credits (y, q_i))

where x is either "the card is blue" or "the card is red" and y is whatever sub-amount of the total amount of credits at a given point in the query sequence ($q_1,\ldots, q_n$). This agent has a very time consuming policy (minimizing the lack of information) and is not well suited for real time situations. Another agent can be introduced that limits Time spent.

Value-Information (wit.z, q_i) =
($\text{Max}_{x,y}$ ((Strength.belief (x, q_{i+1}) ← speech (wit.z, CardRed, q_{i+1})) credits(y, q_i)) + $\text{Max}_{x,y}$ ((Strength.belief (x, q_{i+1}) ← speech (wit.z, CardBlue, q_{i+1})) credits(y, q_i)))/2 – $\text{Max}_{x,y}$ (Strength.belief (x, q_i) credits (y, q_i))

Effective-Choice (q_i) =
1. If $\text{Max}_{\text{wit. z}}$ (Value-Information (wit.z, q_i)) > 0 then Effective-choice (q_i) = QUERY.wit. z such that $\text{Max}_{\text{wit. z}}$ (Value-Information (wit.z, q_i))
2. If $\text{Max}_{\text{wit. z}}$ (Value-Information (wit.z, q_i)) ≤ 0 then Effective-choice (q_i) = BET.yONx such that $\text{Max}_{x,y}$ (Strength.belief (x, q_i) credits (y, q_i))

Satisficing Agent

The Satisficing Agent makes sequential search through the witnesses in his SCAI. He starts with a given threshold γ for expected utility. At each step, he randomly calculates either the expected utility value associated with BET.yONx or the expected utility value associated with BET.yONx after that a given witness will be questioned. This value is the average of the expected utility value associated with BET.yONx in case the witness will say "Red Card" and the expected utility value associated with BET.yONx in case the witness will say "Blue Card". The first option during the sequential search that overcomes threshold γ is chosen. If no suitable option is found after *n* (fixed value) steps, the agent lowers the threshold of a certain value Δδ. With respect to the Normative Agent, the Satisficing Agent makes less queries and it is better suited for open worlds (Simon, 1990).

Perplexity Reducing Agent

The Perplexity Reducing agent has the goal to reduce the level of perplexity below a given threshold δ before betting. Since the only way to reduce perplexity is through queries, the agent starts choosing the witness to test**:** he makes a sequential search on witnesses and takes the first witness whose information is able to reduce perplexity under the threshold. If not suitable witness is found, the agent reduces the value of the threshold of a certain value Δδ and restarts with the same strategy. The expected capacity of a witness of reducing (or augmenting) perplexity represents the expected informative contribute of the epistemic act of querying him. This value is called *expected informativeness* and it is calculated as the actual value of perplexity *minus* the average of the value of perplexity after that witness z says "the card is red" and the value of perplexity after that witness z says "the card is blue"[3].

Expected-Informativeness (wit.z, q_i) =
(Subj.unconfidence(q_i)) – ((Subj.unconfidence(q_{i+1}) ← speech (wit.z, CardRed, q_{i+1}))+ (Subj.unconfidence(q_{i+1}) ← speech (wit.z, CardBlue, q_{i+1})))/2

Expected informativeness is quite different from the *value of information* as defined for the Normative Agent. The difference between those two definitions indicates two different theoretical perspectives: while the Normative Agent maximizes utility values (for bidding) the Perplexity Reducing Agent uses a cognitive theory of sources in order

[3] It follows from the definition that there could be negative values of expected informativeness.

to consider the contribute of the witnesses in the cognitive dimensions of uncertainty, ignorance and contradiction. The Perplexity Reducing Agent is implicitly biased to make queries to witnesses that are in the CAIs, since by definition they lower the value of absolute ignorance more than witnesses that are not in any CAIs. The Perplexity Reducing Agent should be combined with the two others agents (Normative or Satisficing). Once the level of perplexity is under the threshold, he could decide either which color and how much to bid or decide to make a query to another witness using his optimization methods. However, in order to simplify our experiments we did not allow perplexity reducing agents to carry on making queries to witnesses once the degree of perplexity was reduced under the threshold δ. This simplification is plausible for maintaining completely distinct the 2 different functions of epistemic actions: the function of *perplexity reduction* and the function of *"increase" of expected utility*.

Learning During the Game

The RBG game has many turns, so it is possible to learn between them. In the epistemic perspective, it is interesting to model how agents revise information about sources of beliefs.

Updating Reliability Values

All the agents have a representation of the witnesses reliability and are able to update these values depending on past interactions. Since reliability updating strategies are outside the scope of this paper, we used a linear statistical heuristic for all players: witnesses' reliability is lowered if they furnished a wrong advice, augmented otherwise, of a fixed amount $\Delta\phi$.

Updating Classes of Acceptable Ignorance

The Perplexity Reducing Agent is also able to change its SCAI adding or removing the witnesses in the Classes of Acceptable Ignorance. At the beginning of the game the SCAI is set randomly (e.g. the one shown in Fig.1) and it can be updated after each turn extending or contracting its CAIs. Imagine that the agent has queried in sequence w1, w2, w3, w4, w8 before deciding. Imagine he has verified that after the second test the value of perplexity has not changed so much (i.e. less than a threshold α). Since w1 and w2 belong to the same Class1, Class1 can be *contracted* eliminating w2 (that resulted not very informative). Imagine also he has verified that after the fifth test the value of perplexity has changed quite a lot (over a threshold β). Since w4 and w8 do not belong to the same Class, the class of w4 can be *extended* adding w8, that proved to be so informative. We do not describe here the full algorithm for CAIs contraction and extension[4]. We want only to present verbally its structure.

1. Given a previous sequence of queries $(q_1,\ldots, q_n)$, if during that sequence there were two queries q_i and q_{i+1} for *witness A* ← q_i and *witness B* ← q_{i+1} and witness A and witness B belong to the same CAI *x* and the degree of perplexity did not vary so much (in absolute value given threshold α) from q_i to q_{i+1} then the *witness B* is taken out from CAI *x*.
2. Given a previous sequence of queries $(q_1,\ldots, q_n)$, if during that sequence there were two queries q_i and q_{i+1} for *witness A* ← q_i and *witness B* ← q_{i+1} and witness A belongs to CAI *x* whereas witness B belongs to the SCAI but not to CAI *x*, and the degree of perplexity varied a lot (in absolute value given threshold β) from q_i to q_{i+1} then *witness B* is inserted into CAI *x*.

Experimental Setting and Variables

Here we show the comparison between three players: Normative (N), Satisficing (S), Perplexity Reducing (E). There are also two baselines: Random Bidder (B1) that chooses at random to test or to bid (and how much); and Perceptive Bidder (B2) that bids only according to his perceptive input.

The three independent variables we use are: *perception reliability (PR)*; *average witnesses reliability (AWR)*; *witnesses' convergence (WC)*. The first one describes how reliable in absolute is the perception of the agent; the second one indicates how reliable are in average the witnesses answers. They reflect also the "difficulty" of the task. The third one describes how convergent are the answers of the witnesses; this influences the final uncertainty value. We have built three scenarios: *good perception* (where PR is higher than AWR); *good witnesses* (the inverse); *high uncertainty* (where WC is set to a low value, and PR and AWR have the same value)[5].

Results and Discussion

In the following tables we present the preliminary results of our experiments (for Credits and Time) of the three Scenarios (250 simulations, 100 bid turns)[6]. As an indirect

[4] The variable ϕ for reliability updating, as well as thresholds α e β in classes of Acceptable Ignorance updating depend from cognitive biases towards belief revision. It is relevant to notice that for relatively high values of α and relatively low values of β and ϕ the agent is relatively *closed-minded* and conservative (he is less biased to revise the structure of classes of acceptance and the reliability values). But for relatively low values of α and relatively high values of β and ϕ the agent is relatively *open-minded*. This distinction is very close to the typology of cognitive epistemic styles in (Sorrentino et al., 1986).

[5] We use many thresholds and variables in our model: *Close Mind agents* vs. *Open Mind agents* in SCAI revision strategies (thresholds α and β); *strong* vs. *weak* need for low degree of perplexity (threshold δ); *degree of satisfaction* in expected utility (threshold γ); different way to weight different kinds of sources (bias towards perception or witnesses). In order to eliminate their effects we have randomly varied them through the experiments (three dimensions for each variable on average).

[6] The simulations were performed using the cognitive architecture AKIRA, developed at ISTC-CNR (http://www.akira-project.org/).

measure of "algorithm performance", we introduced also Hypothesis Time: it measures how many witnesses an Agent has considered (but not questioned) before deciding.

Table 1: *Good Perception Scenario*.

Agent	Credits	Time	H. Time
B1	**981**	**102**	**0**
B2	**1202**	**0**	**0**
N	1641	**6112**	**112453**
S	**1388**	**987**	**13936**
E	1622	409	10681

Table 2: *Good Witnesses Scenario*.

Agent	Credits	Time	H. Time
B1	**1009**	**101**	**0**
B2	**799**	**0**	**0**
N	1306	**9207**	**144582**
S	**1102**	**997**	**19103**
E	1298	603	13190

Table 3: *High Uncertainty Scenario*.

Agent	Credits	Time	H. Time
B1	**1007**	**99**	**0**
B2	**999**	**0**	**0**
N	**1803**	**8834**	**137866**
S	1551	**1156**	**21033**
E	1563	673	15943

In the first and second Scenarios the Perplexity Reducing Agent performs very well with respect both to gained Credits and temporal measures (Time and Hypothesis Time): it performs at the same level of Normative agent with respect the final amount of credits but his temporal measures are much better. The comparison with the Satisficing agent is even better. Not surprisingly, in the third Scenario he needs to query more witnesses and it is not able to perform as the Normative. Results in **bold** are significant with respect to the Perplexity Reducing Agent. These results show that Perplexity Reducing Agents are very suited in open world conditions where search of new information is in general very costly.

Moreover, a qualitative analysis allows to get a nice result about SCAIs updating: the final SCAIs are in average populated with small CAIs of very reliable witnesses: the average reliability changes from 0.5 to 0.7 in the three scenarios and the number of witnesses remains less to 20 in all simulations. The fact that final CAIs are small and include reliable witnesses is in accordance with the way we learn about belief sources. The more you know an environment, the less you need to question. Moreover, you prefer to question very reliable sources.

Conclusions and Future Work

We have proposed a theoretical foundation of some cognitive categories such as ignorance, uncertainty and contradiction that are generally difficult to quantify in an open world. We have introduced a MAS game (RBG) as a simulation setting in order to compare many agents that take or do not take into account epistemic dimensions. Our preliminary results show that perplexity reduction is a good heuristic for dealing with open world scenarios, and the Structure of Classes of Acceptable Ignorance can be used in order to quantify ignorance and reasoning about it. It would be interesting to test mixed decision strategies (e.g. considering the perplexity in the utility function; or using the Perplexity Reducing Agent as a filter). Another interesting direction is comparing simulation data with data from human experiments; actually the RBG game is being used as an experimental setting in order to collect such data.

Acknowledgments

We are indebted to Cristiano Castelfranchi for many insightful discussions about Ignorance and Uncertainty.

References

Camerer, C., Weber, M. (1992). Recent Developments in Modelling Preferences: Uncertainty and Ambiguity. *Journal of Risk and Uncertainty*, vol. 5, pp. 325-370.

Castelfranchi, C. (1995). Representation and integration of multiple knowledge sources: issues and questions. In Cantoni, Di Gesu', Setti e Tegolo (Eds.), *Human & Machine Perception: Information Fusion*, Plenum Press.

Castelfranchi, C, Falcone, R. & Pezzulo, G. (2003). Trust in information sources as a source for trust: a fuzzy approach. *AAMAS 2003*: 89-96.

Castelfranchi, C., Lorini, E. (2003). Cognitive Anatomy and Functions of Expectations. *IJCAI '03 Workshop on Cognitive modeling of agents and multi-agent interaction,* Acapulco, Mexico.

Ellsberg, D. (1961). Risk, Ambiguity and the Savage axioms. *Quarterly Journal of Economics*, 75, pp. 643-669.

Heath, C., Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, 4, pp. 5-28.

Kirsh, D., Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, pp. 513-549.

Kosko, B. (1986). Fuzzy Cognitive Maps. *International Journal Man-Machine Studies*, vol. 24, pp. 65-75.

Shackle, G. L. S. (1972). *Epistemics and Economics*. Cambridge: Cambridge University Press.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

Simon, H. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, pp.1-19.

Sorrentino, R. M., Short, J. C. (1986). Uncertainty orientation, motivation, and cognition. In R. M. Sorrentino, E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behaviou*r, vol. 1, Guilford Press, New York, pp. 379-403.

Tversky, A., Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* pp. 547-567.

Reinforcement Learning of Dimensional Attention for Categorization

Joshua L. Phillips & David C. Noelle
({JOSHUA.L.PHILLIPS,DAVID.NOELLE}@VANDERBILT.EDU)
Vanderbilt University
Nashville, TN 37235 USA

Abstract

The ability to selectively focus attention on stimulus dimensions appears to play an important role in human category learning. This insight is embodied by learned dimensional attention weights in the ALCOVE model (Kruschke, 1992). The success of this psychological model suggests its use as a foundation for efforts to understand the neural basis of category learning. One obstacle to such an effort is ALCOVE's use of the biologically implausible backpropagation of error algorithm to adapt dimensional attention weights. This obstacle may be overcome by replacing this attention mechanism with one grounded in the reinforcement learning processes of the brain's dopamine system. In this paper, such a biologically-based mechanism for dimensional attention is proposed, and the fit of this mechanism to human performance is shown to be comparable to that of ALCOVE.

Introduction

Human category learning performance cannot be easily explained without recourse to a mechanism for selective dimensional attention (Shepard et al., 1961). Dimensional attention is the cognitive process which emphasizes task relevant stimulus dimensions while deemphasizing others. Thus, contemporary formal models of categorization, such as the Generalized Context Model (GCM) (Nosofsky, 1984), have incorporated adaptable dimensional attention parameters. By adjusting these parameters in a category-specific fashion, the GCM has repeatedly provided excellent fits to human data reflecting the frequency (or probability) with which each stimulus is recognized as an instance of a target category. When the GCM is applied to experimental results, dimensional attention parameters are freely varied to optimize the model fit. This means that, while the GCM provides a powerful account of learned categorization performance, it offers no explanation for how dimensional attention is adjusted over the course of learning.

This shortcoming of the GCM has been addressed by a connectionist model called ALCOVE (Kruschke, 1992). ALCOVE incorporates the GCM's formalization of category knowledge, but it also provides a precise algorithm for modifying the attentional "weight" assigned to each stimulus dimension, based on feedback provided to learners on their categorization judgments. In a typical category learning experiment, learners are presented with stimulus objects, one at a time, and are asked to make classification judgments for each. Immediately following each judgment, feedback is provided, typically informing the learner of the correct category label for the preceding stimulus. Once learning is complete, categorization judgments on transfer stimuli, for which no feedback is provided, can provide a window into the structure of the learned category knowledge. The ALCOVE model uses the feedback provided during training to calculate an "error signal", which is simply the difference between the category assignment made by the model and the specified "true" category. A variant of the backpropagation of error learning algorithm (Rumelhart et al., 1986) is used to communicate this error signal to an early stage of stimulus encoding, and this backpropagated error signal is used to adjust ALCOVE's *dimensional attention weights*. Like the GCM, ALCOVE provides good fits to human performance data on learned categories. Unlike the GCM, ALCOVE provides a detailed account of how dimensional attention is shaped by experience.

ALCOVE has been proposed as a model of *psychological* processes, with virtually no aspiration to explain the neural basis of human category learning. Despite this fact, the empirical successes of ALCOVE and its connectionist formalization make the model a tempting candidate for a coarse characterization of associated brain mechanisms. Perhaps ALCOVE can be refined, with each of its proposed psychological mechanisms mapped onto a corresponding detailed account of the underlying neural machinery. One feature of ALCOVE that stands in the way of such a theoretical reduction is its use of the backpropagation of error algorithm in order to learn dimensional attention weights. This powerful learning algorithm has long been criticized for its lack of biological plausibility (Crick, 1989), suggesting that the brain cannot be adapting dimensional attention based on such a gradient-based technique (c.f., O'Reilly (1996)).

As a first step toward a biological model of category learning, we replaced the backpropagation-based dimensional attention mechanism used by ALCOVE with a reinforcement learning mechanism intended to reflect the role of the brain's dopamine (DA) system in learning. This role for dopamine has been formalized by other researchers in terms of an algorithm called *temporal difference (TD) learning* (Sutton, 1988; Montague et al., 1996). Versions of ALCOVE which adapt dimensional attention weights using the biologically supported TD

learning method, instead of the more computationally powerful but biologically implausible backpropagation method, were found to fit human performance data about as well as the original ALCOVE. Thus, this work offers a more biologically realistic model of the adaptation of dimensional attention without sacrificing accuracy in accounting for human categorization behavior. Also, the ability to capture human performance with the highly stochastic TD learning method suggests that cognitive mechanisms for adapting dimensional attention may not need to be particularly precise.

Background

ALCOVE Architecture

The ALCOVE (Kruschke, 1992) model of category learning is a feedforward connectionist model that involves three layers of processing units (see Figure 1(a)). The input layer consists of a set of units that each correspond to a single dimension in the stimulus psychological space. Explaining the structure of this perceptual representation is outside of ALCOVE's scope. When fitting ALCOVE to human data, multidimensional scaling (MDS) techniques are typically applied to collected stimulus similarity ratings in order to discern the psychological space used by human learners (Shepard, 1962a; Shepard, 1962b). Each input unit has its own dimensional attention weight, α_i. These weights are non-negative scalar values that modulate the amount of attention paid to the corresponding stimulus dimension. Higher α_i values magnify the differences between stimuli along the given dimension, making them easier to discriminate based on that dimension. As learning progresses, these weights are adjusted via the backpropagation of error algorithm.

The hidden layer in ALCOVE contains a set of units that are arranged in psychological space, one for each training *exemplar*. The activation level of each hidden unit is determined by the following equation:

$$a_j^{hid} = exp\left[-c\left(\sum_i \alpha_i \left|h_{ji} - a_i^{in}\right|^r\right)^{r/q}\right]$$

where a_j^{hid} is the activation of hidden unit j, c is the specificity of the hidden units, α_i is the attention weight for input unit i, h_{ji} is the preferred stimulus input for hidden unit j along stimulus dimension i, a_i^{in} is the activation value of input unit i, r is the psychological distance metric, and q is the similarity gradient. Hidden unit activity is at a maximum when the inputs match the preferred stimulus of the unit (i.e., a_i^{in} matches h_{ji}). This activation fades exponentially as the stimulus becomes more distant from the preferred exemplar in psychological space, with the c, r, and q parameters controlling exactly how activation decreases with psychological distance.

Finally the output layer contains a set of units receiving activation from the hidden layer via *association weights*. Each output unit corresponds to a category label that might be assigned to a stimulus. These units are standard linear units, with their activation levels, a_k^{out}, computed as the sum of exemplar unit activation levels, a_j^{hid}, weighted by the corresponding association weights, w_{kj}. Output unit activations are mapped onto response probabilities using an exponential Luce choice rule:

$$P(K) = exp\left(\phi a_K^{out}\right) / \sum_k exp\left(\phi a_k^{out}\right)$$

where $P(K)$ is the probability of selecting category K for the current stimulus, and ϕ is a gain term. These response probabilities may be used to compare network responses with human performance data.

After the presentation of each stimulus and the consequent outputs are produced, the output unit corresponding to the correct response is presented with a target activation level of $+1$, and other units are presented with targets of -1. An error signal consisting of the difference between a_k^{out} and these targets is used to adjust weight values (though output units that "overshoot" their target values are assigned zero error). The association weights are then adjusted using this error signal directly (i.e., using the delta rule), but the selective attention weights are adjusted based on a backpropagated error signal. The resulting weight update equations are:

$$\Delta w_{kj}^{out} = \lambda_w \left(t_k - a_k^{out}\right) a_j^{hid}$$

$$\Delta\alpha_i = \lambda_\alpha \sum_j \left[\sum_k \left(t_k - a_k^{out}\right) w_{kj}\right] a_j^{hid} c \left|h_{ji} - a_i^{in}\right|$$

where Δw_{kj}^{out} is the adjustment value for the association weight from hidden unit j to output unit k, $\Delta\alpha_i$ is the adjustment value for the attention weight for input unit i, λ_w and λ_α are the learning rate parameters for the association weights and attention weights, respectively, and t_k is the target value for output unit k.

Temporal Difference Learning

Electrophysiological studies of the dopamine neurons of the basal ganglia have suggested that the firing rates of these cells code for *changes in expected future reward* (Shultz et al., 1997). This is particularly interesting because a measure of change in expected reward is the key variable of a reinforcement learning method called *temporal difference (TD) learning* (Sutton, 1988). This has led a number of researchers to develop TD learning models of the role played by the midbrain dopamine system in learning (Barto, 1994; Montague et al., 1996).

In the TD framework, a continuous reward value (r) is delivered on each time step (t), with positive reward being desirable. A neural system called the *adaptive critic* learns to predict expected future reward (V), given features of the current situation. When future rewards are exponentially discounted by a factor, γ (between 0 and 1), with immediate rewards being valued more than temporally distant ones, the change in expected future reward between two consecutive time steps is given by:

$$\delta(t) = r(t) + \gamma V(t) - V(t-1)$$

This δ value is called the *temporal difference (TD) error*. The global TD error value can be used to drive learning in the adaptive critic, improving predictions of future reward, and it can also be used to adapt connection weights in neural networks which select actions, pushing those choices toward actions that regularly lead to reward. Models of this kind have been used to explain motor sequence learning in the striatum (Barto, 1994) as well as other forms of learning. We propose that this form of reinforcement learning may also be used to learn dimensional attention weights that lead to correct categorization responses and, thus, reward.

Modeling Approach

Applications of TD learning typically focus on choosing an action from a discrete set. There is currently no clear understanding of how to apply these methods to domains in which a continuous output is needed. Dimensional attention weights are continuous parameters, however, so some modification to standard TD learning is needed to apply this technique to the adaptation of dimensional attention. We have devised two novel connectionist architectures to accomplish this. Our strategy encodes attentional weight vectors (with one α_i weight per dimension) across a single layer of standard connectionist processing units, called the *attention map* layer. Each unit in this layer possesses a fixed preferred attentional weight vector, and activation of a unit encourages the use of that unit's preferred dimensional attention weights. The activation level of each unit is largely determined by its individual bias weight, and the TD learning method is used to adapt these bias weights so as to optimize reward.

At the start of each trial, each of these attention map units is activated, to some degree, by its bias weight. The units then compete to determine the set of dimensional attention parameters to be used by ALCOVE, and the result of this competition is a set of such attention weights. ALCOVE then processes the current stimulus in its usual fashion, producing a categorization judgment. ALCOVE's association weights are then modified in the usual way, using the delta rule, but the dimensional attention weights are handled differently. If ALCOVE confidently chooses the correct category, it is rewarded. Otherwise, it is not. The TD error, δ, is calculated based on this reward signal, and this error is used to modify the bias weights of all active attention map units.

Two different architectures for the attention map layer were investigated. The first of these used *conjunctive coding*, resulting in a *localist* representation of dimensional attention. Under this scheme, the preferred attentional weight vectors of processing units were distributed evenly throughout the weight vector space. Thus, each unit corresponded to a position in attention weight vector space, and the positions of all of the units in the attention map layer formed a uniform grid in this space. On each trial, a simple winner-take-all competition determined the one unit whose preferred weight vector would specify the distribution of attention for that trial. Learning occurred only for the winning unit, using the following weight update equation for its single bias weight:

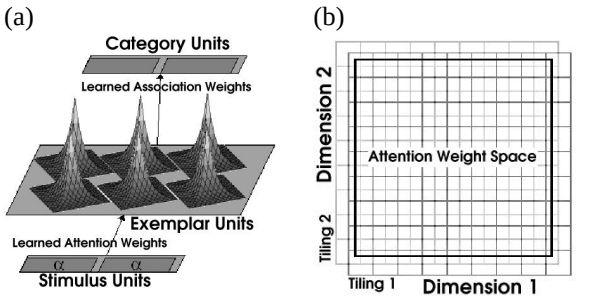


Figure 1: (a) ALCOVE Network Architecture. (b) Tile Coding Of The Attention Map Layer — A single unit is centered in each tile.

$$\Delta w_i \;=\; \lambda_r\,(r - a_i)\; f'(net_i)$$

where λ_r is the attention map learning rate, r is the reward for the current trial, a_i is the activation value of the winning attention map unit, and $f'(net_i)$ is the derivative of the unit's activation function (which was the standard logistic sigmoid). Note that this is the standard method for updating weights based on TD error, under the condition of absorbing reward (i.e., we don't predict reward past the end of the trial). In this case, a_i acts as our reward prediction ($V(t-1)$), and we do not predict beyond this trial, so $V(t) = 0$ and $\delta = r - a_i$. A reward value (r) of $+1$ was delivered to the network on trials in which ALCOVE selected the correct category label and produced a confident response (i.e., all output units within 0.5 of their targets). A reward of 0 was delivered, otherwise.[1]

Our second attention map architecture used *tile coding*, resulting in a *distributed* representation of dimensional attention. In this case, the attention map layer was partitioned into disjoint *tilings*, where each tiling contained a set of units with preferred dimensional attention weight vectors that uniformly spanned the full weight space. The preferred weight vectors of the units in the various tilings were not identical, however, because each tiling was "offset" from the others, as shown in Figure 1(b). To precisely represent a position in the attention weight space, one unit in each tiling is activated, with the overlap in the *tiles* surrounding the positions of these units determining the dimensional attention weights to be used. This kind of distributed representation was originally used in the Cerebellar Model Articulation Controller (CMAC) (Albus, 1975), and improved generalization in TD learning systems has been

[1] An obvious alternative reward schedule involves stochastically making category judgments based on $P(K)$ and rewarding any correct judgment. While we are currently investigating this approach, it is likely that it will produce behavior that deviates substantially from that of standard ALCOVE. Dimensional attention weights do not change much in ALCOVE until the network starts to make strong responses. This is part of the "three-stage learning" profile that ALCOVE exhibits. Our reward schedule encourages this pattern of learning.

found to result from their use (Sutton, 1996). Previous models have used such representational schemes to encode network inputs, but here they have been used in a novel way to select dimensional attention weights. As in the conjunctive coding architecture, each attention map unit is activated by a bias weight, and a competition ensues between units. In the tile coding scheme, the most active unit across all tilings restricts activity in the other tilings to those that are close to the winning unit (i.e., units whose tiles overlap with that of the winning unit). This competitive process is recursively applied to tilings that do not contain the winning unit until one unit is active in each tiling, and the tiles corresponding to these units all overlap. The attention weight vector at the center of this overlapping region is then used by ALCOVE to process the current stimulus. Once feedback is provided, reward is calculated as in the conjunctive coding case, and TD learning is used to adjust the bias weights of all of the winning units in the attention map layer.

In standard ALCOVE, the initial attention weights are often set to be all equal and to sum to one. This effectively emphasizes all dimensions equally at the start of training. We selected initial bias weights in the attention map layer so as to form a similar initial bias in our models. The unit in the attention map whose preferred attention weight vector matched ALCOVE's standard initial attention weights was given a maximum bias weight (0.05), and the bias weights assigned to other units fell off in a Gaussian fashion as the distance from this peak increased (in attention weight space), bottoming out at -0.05. A small amount of uniformly sampled noise was then injected into each bias weight, and the result was clipped to the $[-0.05, 0.05]$ range. The variance of the Gaussian and the range of the injected noise were free parameters of the model.

Results

In order to assess the ability of our reinforcement-based dimensional attention mechanism to account for human performance, we applied our models to several previously reported category learning studies. The performance of our modified version of ALCOVE was compared to that of the standard version of ALCOVE and to the performance of the GCM. In all cases, the values of dimensional attention weights were bound between zero and one. (This was only a new upper bound for ALCOVE, which standardly forces these weights to be nonnegative.) In all of the learning models, weights were updated after every simulated trial.

Dimensional Attention & Learning Difficulty

Shepard et al. (1961) examined the effect of category structure on the relative speed with which a category is learned. Stimuli were composed of three easily separable binary dimensions, for a total of eight possible stimuli. Six category structures were examined, ordered approximately by increasing number of relevant dimensions. Thus, the Type 1 category structure requires attention to only one binary dimension to solve the task, the Type 2 structure requires that only two of the dimensions be attended, while Types 3, 4, 5, and 6 all require attention to all three, in order of increasing dimensional significance. The speed with which humans learn these categories matches this ordering of tasks, but models that lack a dimensional attention mechanism fail to learn Type 2 categories faster than some of the more difficult categories. Kruschke (1992) showed that ALCOVE, with its adaptive dimensional attention mechanism, learned Type 2 tasks at a relative rate comparable to human learners. We have replicated these simulations (using bounded attention weights and learning after every trial), and the results are shown in Figure 2.

We applied our reinforcement learning version of ALCOVE to these six categorization tasks. Since stimuli had three dimensions, the attention weight space was three-dimensional. The conjunctive coding model used a $15 \times 15 \times 15$ unit topology in its attention map layer (3375 units total), while the tile coding model used five tilings of $9 \times 9 \times 9$ units each (3645 units total). The results of these simulations are shown in Figure 2. Note that our models learn Type 2 categories faster than the higher numbered types, just as ALCOVE does. Model parameter values were manually selected to produce performance that matched the category learning times exhibited by ALCOVE. These results demonstrate that TD learning can adapt dimensional attention weights so as to speed category learning.

Categorization of Continuous Separable Stimuli

In order to demonstrate the ability of our models to quantitatively fit human performance on categorization tasks involving stimuli with continuous and separable dimensions, we applied these models to an experiment conducted by Nosofsky (1986) . The stimuli in this experiment consisted of semicircles that varied in size and contained a radial line oriented at different angles. These stimuli were to be categorized as members of one of two categories, and four different category structures were explored (see Figure 3). The frequency with which each of the sixteen possible stimuli were placed in a target category was measured after training, and the GCM was fit to these response probabilities.

We fit both standard ALCOVE and our reinforcement learning models to this data, as well. Since the stimulus space was two-dimensional, our models used a two-dimensional attention map layer. In the conjunctive coding case, a 15×15 unit topology was used (225 units total), and the tile coding model used 9 tilings of 5×5 units each (225 units total). While both schemes used the same number of units, the tile coding model discretized the space with a much greater resolution. Stimuli were presented to the models using the MDS code found by Nosofsky. Free parameters of the models were fit to Nosofsky's Subject 1 data for each category structure separately. A simple hill-climbing optimization algorithm on sum-squared error was used.

The quality of the resulting fits are summarized in Table 1. While the original ALCOVE model provided the

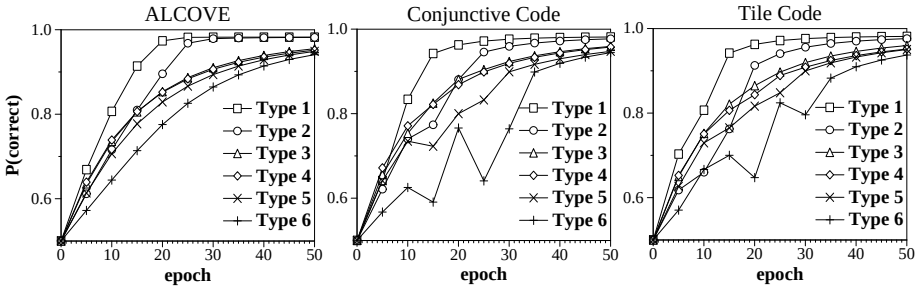


Figure 2: Model Learning Curves For Shepard's 6 Tasks — One epoch involves one trial with each distinct stimulus.

	Category Structure			
Model	*1*	*2*	*3*	*4*
GCM	99.93%	94.73%	84.52%	98.31%
ALCOVE	99.65%	96.45%	86.62%	98.61%
Conj. Code	99.51%	95.84%	86.01%	97.72%
Tile Code	99.54%	95.62%	83.55%	97.72%

Table 1: Model Fits To Nosofsky (1986) – Percent Variance Accounted For

	Category Structure					
Model	*1*	*2*	*3*	*4*	*5*	*6*
GCM	99.10%	98.30%	97.20%	99.80%	98.20%	99.20%
ALCOVE	98.56%	99.29%	93.53%	99.79%	98.34%	98.71%
Conj. Code	98.44%	98.16%	92.51%	99.57%	97.94%	98.53%
Tile Code	98.25%	97.27%	91.15%	99.15%	97.80%	97.30%

Table 2: Model Fits To Nosofsky (1987) – Percent Variance Accounted For

best overall fits, our models matched the data almost as well, and all general trends in the ALCOVE and GCM fits are present in our models. This suggests that our mechanisms for learning dimensional attention can quantitatively capture human performance on learning tasks that require selective attention to separable dimensions.

Categorization of Continuous Integral Stimuli

Integral stimulus dimensions often entail a difficulty in focusing attention on individual dimensions. Despite this fact, Nosofsky (1987) showed that models equipped with a dimensional attention mechanism fit human categorization performance on such stimuli slightly better than models that lacked such a mechanism. This study involved 12 different color chips which varied in saturation and brightness. Six different category structures were used, and these are shown in Figure 3. The frequency with which each of the 12 stimuli were placed in a target category was measured after training, and, once again, the GCM was fit to these response probabilities.

We applied both the original ALCOVE and our reinforcement learning versions to this human data. The same attention map layer sizes as used in the previous simulations were used here, and, as before, MDS representations of the stimuli were presented to the models. A summary of the model fits is shown in Table 2.

The GCM provides the best fits to the data in this study. It seems that the ALCOVE model and our models had trouble learning Category Structure 3. This is a difficult category structure which benefits little from selective attention to specific dimensions. Note, however, that the fits of our reinforcement learning models are close to the standard ALCOVE fits, and our models continue to exhibit the same trends in learning as ALCOVE.

Discussion

Our results show that established computational models of the brain's dopamine system can provide an adequate replacement for the biologically implausible backpropagation of error method for adapting dimensional attention during category learning. The new models were able to learn useful dimensional attention weights from their less-informative global reinforcement signal. This suggests that cognitive mechanisms for allocating dimensional attention may not be as precise as those posited by the original ALCOVE model.

One noteworthy feature of our reinforcement learning models was their tendency to exhibit fluctuations in performance over training, rather than smooth and monotonic learning as displayed by the original ALCOVE model. If each network model is to mirror the performance of an individual learner, these performance fluctuations may reflect stochasticity commonly observed in individual behavior. Also, if performance is averaged across multiple "simulated individuals", smooth learning curves, like those generated by ALCOVE, are produced.

Our models encoded dimensional attention weights in a fairly conjunctive fashion, with individual units in the attention map layer specifying levels of attention for all of the dimensions. This is needed because the appropriateness of attention to one dimension depends on how attention is allocated to the other dimensions. Such a conjunctive encoding requires very large attention map layers, however, and this may limit the scalability of this approach. In order to address this issue, we are currently exploring more compact distributed representations for dimensional attention weight vectors.

Eventually we hope to modify ALCOVE to make use of additional biologically plausible mechanisms of neural computation. This work represents the first step in this process, identifying a biologically realistic method for governing dimensional attention.

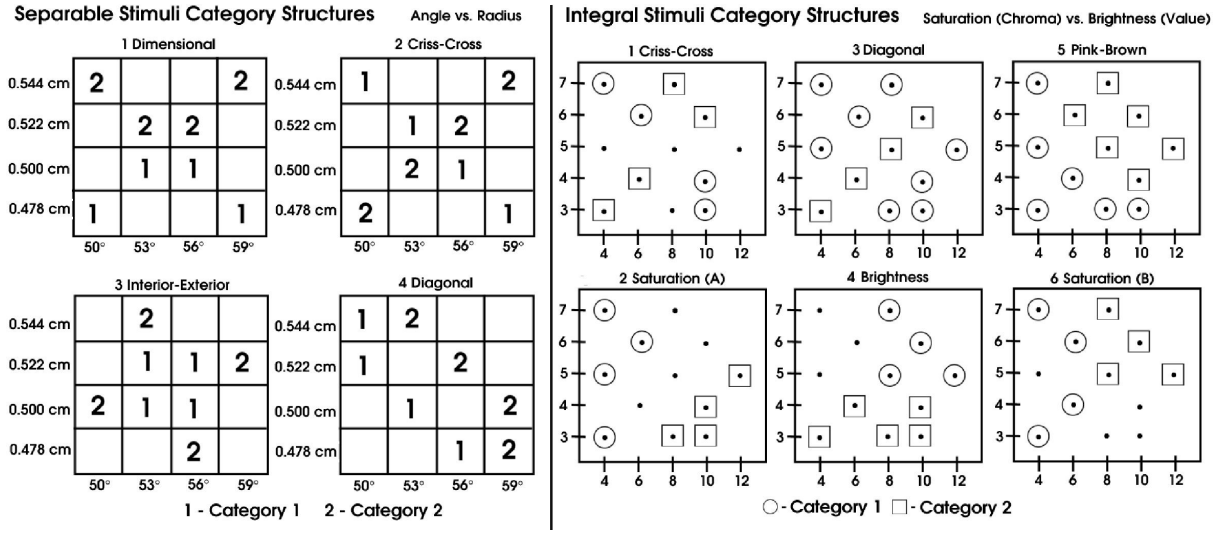


Figure 3: Category Structures Used In Nosofsky (1986) and Nosofsky (1987)

Acknowledgments

The authors extend their thanks for helpful comments on this work to Tom Palmeri, the members of the Vanderbilt University Computational Cognitive Neuroscience Laboratory, and three anonymous reviewers.

References

Albus, J. S. (1975). A new approach to manipulator control: The cerebellar model articulation controller CMAC. *Journal of Dynamic Systems, Meaasurement, and Control*, 97(3):220–227.

Barto, A. G. (1994). Adaptive critics and the basal ganglia. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 215–232. MIT Press.

Crick, F. M. C. (1989). The recent excitement about neural networks. *Nature*, 337:129–132.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16:1936–1947.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(1):104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *JEP: General*, 115(1):39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1):87–108.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, 8(5):895–938.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27:125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27:219–246.

Shepard, R. N., Howland, C. L., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75(13 Whole No. 517).

Shultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44.

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, 8:1038–1044.

The Time-Course and Cost of Telicity Inferences

Andrea S. Proctor (a-proctor@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Michael Walsh Dickey (m-dickey@northwestern.edu)
Department of Communication Sciences and Disorders, Northwestern University
2240 Campus Drive, Evanston, IL 60208 USA

Lance J. Rips (rips@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Abstract

Recent evidence suggests that perceivers have consistent intuitions regarding the boundedness properties of objects and events (Solomon, Proctor, & Rips, in preparation). This paper presents a self-paced reading study examining the speed and accuracy with which readers draw such telicity inferences during on-line language comprehension. Participants read sentences containing either a consumption verb ("consume") or an observation verb ("monitor") followed by either a mass or a count object ("ice water" vs. "ice cube"). Each sentence ended with an adverbial phrase that was either consistent or inconsistent with the telicity of the preceding event description ("in" or "for" adverbials), along with a comprehension question. Reading-time results suggest that comprehenders are slow to draw telicity inferences, even when the type of verb unambiguously determines the telicity of the sentence. However, responses to post-sentential comprehension questions suggest that verb and noun information together have a surprisingly robust influence on comprehenders' telicity inferences, even in the face of supposedly unambiguous adverbial information. Together, these results suggest that comprehenders make use of all relevant information in making telicity inferences, but that they do so much more slowly than strongly incremental models of natural language understanding would predict (e.g., Marslen-Wilson & Tyler, 1980).

Introduction

We readily distinguish two types of physical entities in the world—individuated objects and substances. We refer to these two types of entities with different types of nouns: substances are often referred to with mass nouns (e.g., tea) and objects with count nouns (e.g., cat). These distinctions seem to rest on the boundaries of the physical entities: count nouns typically refer to objects with well-defined boundaries, such as *mouse* or *iceberg*, while mass nouns typically refer to substances[1] without clear boundaries, such as *mud* or *water*.

The domain of events can be divided up similarly. Events can be classified according to whether or not they have an endpoint, or temporal boundary. Actions described by atelic verbs have no inherent endpoint or boundary; these actions have the potential to go on without end (e.g., singing). Verb phrases describing atelic or unbounded events go naturally with *for* adverbials, which describe the duration of an event, and less naturally with *in* adverbials, which presuppose the endpoint of an event. "She sang for an hour" sounds much more natural than "She sang in an hour." Actions described by telic verbs such as *delivering*, on the other hand, have an inherent endpoint; once an object has arrived at its destination, delivering has reached its end, and cannot logically continue. Verb phrases describing telic or bounded events go naturally with *in* adverbials and less naturally with *for* adverbials: "She delivered the package in an hour" sounds much more natural than "She delivered the package for an hour." These distinctions among verbs or verb phrases are commonly referred to as lexical aspect (e.g., Vendler, 1967; Dowty, 1979).

The current experiment explores how these lexical aspect distinctions are computed during sentence comprehension. It also examines when perceivers draw inferences about the boundedness of events, just as they must draw inferences about the boundaries of physical entities (see Solomon, Proctor, & Rips, in preparation). There has been some previous work exploring the cost of modifying or retracting such inferences once they have been drawn: Piñango, Zurif, and Jackendoff (1999) and Todorova, Straub, Badecker, and Frank (2000) demonstrate that encountering information (such as a *for* adverbial) that forces an event to be construed as atelic causes processing difficulty if previous information had suggested it was telic. The current experiment uses this effect to explore when telicity inferences are drawn on-line.

[1] It should be noted, however, that this distinction between mass and count nouns does not strictly coincide with the substance/object distinction. Jackendoff (1991) notes that, in addition to substances, aggregates of individuated objects can act like mass nouns. For instance, the terms *cattle* and *change* behave like mass terms, even though each refers to individuated objects (cows and coins, respectively) and not to an unindividuated substance.

Parallels Between Noun and Verb Contrasts

Several authors have noted that strong parallels exist between mass nouns and atelic events, and between count nouns and telic events (Bach, 1986; Langacker, 1987; Vendler, 1967). Just as masses have no intrinsic physical bound, atelic events, such as *running*, *painting*, or *watching*, have no intrinsic temporal bound. In contrast, just as counted objects have inherent physical boundaries, telic events, such as *delivering*, *drowning*, or *walking a mile*, all have an intrinsic temporal bound. One cannot continue delivering after a package has reached its destination, drowning after one is dead, or walking a mile after that distance has been crossed. Once the endpoint is reached, the action is completed. The present experiment is concerned with one subtype each of atelic and telic verbs: activities and accomplishments, respectively.

Another key parallel between actions and objects involves the extent to which a part of an object or action can be considered to be in the same category as its whole. For both masses and activities, a subpart (down to some lower limit) of the whole is qualitatively equivalent to the whole—any part of chocolate sauce is still chocolate sauce, just as any part of eating is eating. This subpart or subinterval property (Bennett & Partee, 1978) does not hold for counted objects and accomplishments, however—any part of an aluminum boat is not, itself, an aluminum boat, nor is any part of lighting a fire (e.g., crumpling up newspaper) itself lighting a fire.

Interactions Between Noun and Verb Boundaries

The physical boundaries of objects influence the temporal boundaries of events affecting them. Several authors have noted that the telicity or boundedness of an event often depends on whether the verb describing it takes a mass or count noun as its object (Pustejovsky, 1991, 1995; Verkuyl, 1993). When a consumption verb, such as *eat*, takes a count noun as its object, readers should infer that the VP is an accomplishment—the depletion of the object must end when the object's boundary is reached. When such a verb takes a mass noun as its object, however, the VP is an activity—since the substance is unbounded, the depletion could potentially go on indefinitely. These telicity shifts only hold for a subset of verbs such as consumption and creation verbs, which describe events that cannot easily be repeated and that entail an irreversible effect on their objects (Krifka, 1998). Verbs that do not entail an irreversible effect on their objects should not demonstrate such an aspectual shift. For example, whether a mass or count noun appears as the object of a verb of observation (e.g., watching) should not have an effect on the VP's telicity—the action of watching a mug should be just as unbounded as the action of watching soup.

To date, there is limited evidence regarding how perceivers draw these telicity inferences during comprehension. Solomon and her colleagues (in preparation) provide evidence from off-line reasoning tasks suggesting that readers are sensitive to the boundedness of different events, and that their reasoning about event boundedness parallels their reasoning about the boundedness of physical objects. Further, readers make inferences about lexical aspect on-line, re-interpreting their default assumptions regarding the boundedness of an event in order to bring it in line with the temporal characteristics of the context in which it appears (Piñango et al., 1999; Todorova et al., 2000). For example, "jump" might be interpreted to be an iterative action when it appears in the sentence "He jumped all day."

Studies examining aspectual coercion have demonstrated that sentential aspect is sensitive to parts of the sentence other than the verb. However, they do not directly examine how the boundedness of a verb's object or the verb itself is capable of influencing telicity inferences. If perceivers draw inferences about a sentence's temporal profile incrementally, they should show early sensitivity to the difference between observation verbs (whose boundedness does not depend on the properties of the following object) and consumption verbs (whose boundedness does depend on the object). Similarly, if perceivers are actively and predictively computing telicity based on verb and object information, they should show evidence of an interaction of verb and object information at the position of the object: perceivers' comprehension of a sentence should be affected by whether an object is a mass or count noun only if the verb preceding it is a consumption verb, not if it is an observation verb. Strongly incremental views of language understanding (e.g., Marslen-Wilson & Tyler, 1980; Clark, 1996) predict such early inferencing.

Overview

Is it the case that when we read a sentence in which an (unbounded) activity verb, describing the consumption of its object, takes a (bounded) count noun as its object, we then interpret the sentence as if the verb phrase were bounded? If, for instance, we read "Carol ingested Henderson Foods' rice *cake* merrily for ten minutes" will we interpret that action as being more temporally bounded than we would if we had read that she had ingested Henderson Foods' rice *cereal*? When do we make these inferences? Do we begin to draw inferences regarding telicity as soon as we encounter a verb? Do we compute a sentence's aspect as soon as we have both the verb and noun information? Or do we hold off making inferences until all potentially informative information is available, until late in the sentence? Previous results suggest that drawing telicity inferences can be cognitively costly (see Piñango, et al., 1999; Todorova, et al., 2000); the present experiment uses this finding to address these questions.

The Experiment

We presented participants with sentences describing characters either consuming or observing a mass or a counted object (see Table 1 for examples). Sentences were divided into five segments, and participants read through

Table 1: Sample Set of Sentences

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Adverb	Verb	Noun
a	Leslie consumed	Polar Purity's	ice water	with zeal	for eight minutes.	For	Telic	Mass
b	Leslie consumed	Polar Purity's	ice cube	with zeal	for eight minutes.	For	Telic	Count
c	Leslie monitored	Polar Purity's	ice water	with zeal	for eight minutes.	For	Atelic	Mass
d	Leslie monitored	Polar Purity's	ice cube	with zeal	for eight minutes.	For	Atelic	Count
e	Leslie consumed	Polar Purity's	ice water	with zeal	in eight minutes.	In	Telic	Mass
f	Leslie consumed	Polar Purity's	ice cube	with zeal	in eight minutes.	In	Telic	Count
g	Leslie monitored	Polar Purity's	ice water	with zeal	in eight minutes.	In	Atelic	Mass
h	Leslie monitored	Polar Purity's	ice cube	with zeal	in eight minutes.	In	Atelic	Count

these sequentially at their own pace while their reading times were recorded. The first segment contained either a consumption or an observation verb, while the third segment contained either a mass or a count noun. A fourth segment consisted of a manner adverbial, which served as a wrap-up segment. The final segment specified an interval of time, preceded by either a *for*-adverbial (e.g., for eight minutes) or *in*-adverbial (e.g., in eight minutes). Recall that these adverbials typically appear with VPs describing activities and accomplishments, respectively, and rarely with the opposite type of event.

If incremental views of language understanding are correct that readers will begin to draw inferences about the temporal profile of a described event as soon as they encounter relevant information, we would expect to find an effect of verb at the first segment. Specifically, we would expect to find that reading times for segments containing observation verbs (e.g., "Leslie monitored") would be longer than those for segments containing consumption verbs (e.g., "Leslie ingested"). Observation verbs license immediate telicity inferences (since the telicity of such events is independent of noun information later in the sentence), but the telicity of consumption verbs depend on noun information. Such a cost could conceivably continue through the second and third segments.

Furthermore, if inferences about an event's aspect are occurring early on-line (as in the incremental view) and are associated with a processing cost, then we would expect to find slowed reading times at the third and/or fourth segment (near the time of the object noun information) of sentences containing consumption verbs. In the case of Leslie, for example, if there were immediate processing costs associated with integrating noun and verb boundary information, we would expect participants to be slower to read that she had consumed "ice cube" than "ice water." Such a cost could conceivably carry over to the following segment, in which case we would expect slower reading of "with zeal" among readers who had read the count version relative to those who had read the mass version.

If, on the other hand, participants delay drawing inferences about the telicity of described events until all relevant information is available (until after verb and noun information has been encountered), we would not expect to see an effect of verb at the first segment, nor would we expect an interaction between noun and verb later on.

In the absence of slowing at segments three or four, we could still determine that inferences about aspect were being made based on a combination of a consumption verb with a mass or count noun if we were to find slowing due to mismatching of grammatical information at segment five. If participants are making inferences about the aspect of the sentences (e.g., inferring that eating a chocolate bar is an accomplishment) and are then presented with a final adverb (e.g., for ten minutes) that contradicts this inference, their reading times at that final segment may be slowed.

As an additional test of readers' inferences about the aspect, we presented participants with a follow-up question immediately after they had read through each sentence. For example, after reading about Leslie consuming ice water, participants were presented with the question, "After four of those eight minutes, had Leslie actually ingested Polar Purity's ice water?" The questions were modeled after an inferential test (Dowty, 1979) that distinguishes between activities and accomplishments by assessing whether the subinterval property applies to the action. If a participant believes that the action was an activity, he or she should be willing to ascribe the subinterval property to it. We thus expected that participants would be more likely to respond "yes" to the follow-up question after they had read about Leslie consuming ice water than if they had read about Leslie consuming Polar Purity's ice cube, since the consumption of a bounded object should lead participants to interpret the action as telic. Observation verb sentences should have the subinterval property ascribed to them regardless of whether they contain a mass or count noun, as the action has no effect on the object. We expected the pattern for sentences ending with the in-adverb to be qualitatively similar to that for sentences ending with the for-adverb. However, overall likelihood of an atelic response should be lower for *in*, given the strong association of in-adverbials with telic actions.

Methods

Procedure In this study, participants completed a self-paced reading task. They read through sentences, presented on a computer screen while we recorded their reading times. We instructed participants to read through each segment at their normal pace, and to progress through the segments by pressing the spacebar. The segments appeared sequentially

and disappeared from view once they had been read; participants could not return to a previously-viewed segment.

A follow-up question appeared immediately after participants had read the final segment of the sentence. The follow-up question appeared with two possible responses ("yes" and "no" for the experimental sentences), one appearing on the left, and the other on the right. For each list, participants saw "yes" on the left for half of the experimental sentences, and "no" on the left for the other half. We instructed participants to press one key (d) if they felt that the response on the left was correct and another (k) if they felt that response on the right was the better choice.

Each participant saw the sentences and their associated follow-up questions in a different random order.

Materials We constructed 80 sentences, all describing a character performing some action on an object or substance over a specified interval. For each of these sentences, we varied the type of verb (observation verbs vs. consumption verb), the boundedness of the noun (mass vs. count noun), and the final adverb (for X minutes vs. in X minutes), yielding a set of eight variations on each of the 80 base sentences. Each sentence contained five segments (an example set of segmented sentences appears in Table 1). The 640 experimental sentences were separated into eight lists, such that each of the eight variations of any given base sentence was assigned to a different list, and each list contained ten of each of the variation types. Thus, participants saw only one version of each base sentence, but saw an equal number of each of the verb/noun/adverb combinations. Each list also contained 32 additional sentences, unrelated to the experimental sentences, which served as filler items (e.g., "At the break, | Emily | had already | finished the memo, | to her boss's relief.")

Follow-up questions were constructed for all the sentences. The follow-ups asked whether the character had *actually* completed the specified action halfway through the mentioned interval (an example set of follow-up questions is presented in Table 2). If, during the course of reading a sentence, the participant inferred that the described event was atelic, then the subinterval property should apply, and the expected response would be "yes". If, on the other hand, the participant inferred that the described event was telic, he or she should respond "no".

Table 2: Sample Set of Follow-up Questions

a, e	After four of those eight minutes, had Leslie actually consumed Polar Purity's ice water?
b, f	After four of those eight minutes, had Leslie actually consumed Polar Purity's ice cube?
c, g	After four of those eight minutes, had Leslie actually monitored Polar Purity's ice water?
d, h	After four of those eight minutes, had Leslie actually monitored Polar Purity's ice cube?

Participants Forty-eight undergraduate students enrolled at Northwestern University participated in this experiment. Participation was part of a course requirement in an introductory psychology course. All participants were native English speakers.

Results

An examination of the reading time data suggests that readers did not begin generating telicity inferences as soon as they encountered the relevant verb or verb+object information. Reading times were, however, longer for sentences in which the final adverbials were inconsistent with the telicity of the preceding verb+object combination, indicating that the verb+object information was used in generating inferences about the telicity of these described events. Responses to follow-up questions suggest that these inferences were surprisingly robust: even in cases where the sentence-final adverbial conflicted with the telicity of the preceding verb phrase, participants showed some evidence of having stuck with their original telicity inference.

Analyses on reading times and responses to follow-up questions were computed separately using participants and items as random factors.

Reading-Time Analyses Mean reading times for each segment are presented in Figure 1. Reading times that were over 10 seconds or under 100 milliseconds were excluded from the analysis. These responses consisted of less than 2% of the data.

Segment-by-segment reading time analyses revealed a significant effect of verb type at segment one ($F_P(1,47) = 7.21$; $F_I(1,79) = 4.11$, $p < .05$ for both); however, this effect was in the direction opposite to that expected under a strongly incremental view—observation verbs were read more quickly than consumption verbs. There was no evidence of additional inferencing work going on in the atelic conditions.[2]

Also speaking against the predictions of an incremental account is the finding that there were no significant differences at segment three (the count/mass noun segment) or segment four (the manner adverbial segment), indicating there is no immediate processing cost associated with drawing telicity inferences based on the integration of noun and verb information.

Despite the lack of immediate processing, an examination of the reading times for the final segment (see Figure 2) provides some evidence that participants were combining the verb and noun information to generate inferences about the telicity of the events. Analyses of segment five reading times reveal evidence of processing costs when participants encounter grammatical information conflicting with telicity inferences. A 2x2x2 repeated measures

[2] This main effect may have been the result of a confound with word frequency—the mean frequency (Kučera & Francis, 1967) of the observation verbs was significantly higher than that of the consumption verbs ($t(59) = 3.64$, $p < .01$).

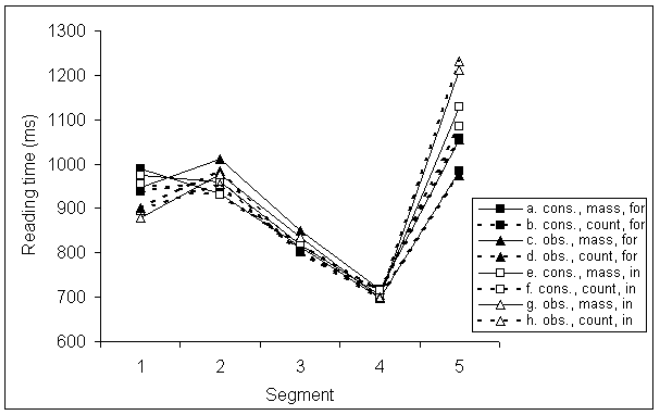


Figure 1: Mean Reading Times

ANOVA revealed a three-way interaction between noun, verb and final adverb ($F_P(1,47) = 4.29$, $F_I(1,79) = 4.46$, both $p < .05$). This interaction reflects both the relatively long segment five reading times for sentences whose final adverbs conflict with the telicity information contained in earlier segments (types e, g, and h). This difference was confirmed with planned contrasts ($F_P(1,47) = 11.23$, $F_I(1,79) = 33.01$, $p < .01$ for both).

Looking at the for-adverbial conditions alone, there was a two-way interaction of verb and noun type with slower reading times for sentences with consumption verbs and count nouns than sentences with consumption verbs and mass nouns, though it was only marginally significant in the items analysis ($F_P(1,47) = 9.72$, $p < .01$; $F_I(1,79) = 3.71$, $p < .06$). This suggests that participants had successfully drawn telicity inferences based on verb and noun information earlier in the sentence.

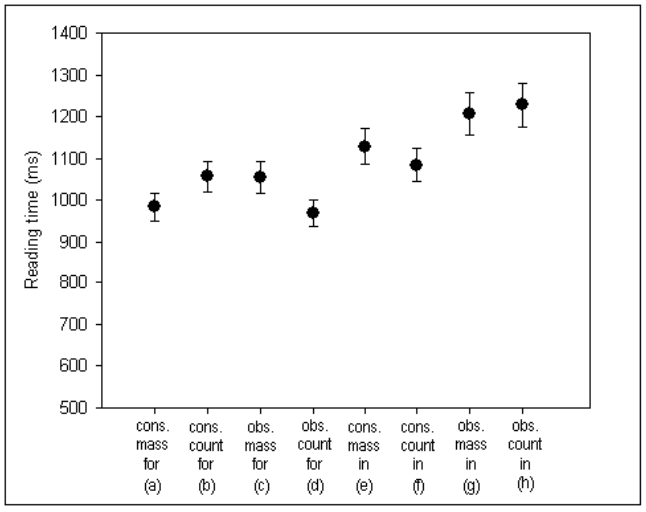


Figure 2: Reading Times for Segment 5

Follow-up Question Data Responses to the follow-up questions indicate that the participants were sensitive to the experimental manipulations. Repeated measures ANOVAs revealed main effects for final adverb ($F_P(1,47) = 66.84$; $F_I(1,79) = 757.12$, both $p < .001$), verb ($F_P(1,47) = 64.08$; $F_I(1,79) = 131.67$, both $p < .001$) and, though only marginal in the items analysis, for noun ($F_P(1,47) = 4.27$, $p < .05$; $F_I(1,79) = 3.14$, $p = .08$). This pattern of responses indicated that, as expected, participants were more likely respond "yes" (indicating an atelic interpretation) to the questions following sentences that contained *for* adverbs, atelic verbs, and mass nouns than they were for sentences containing *in* adverbs, telic verbs, and count nouns, respectively (see Figure 3). There were no significant interactions. Planned t-tests comparing responses to questions following mass and count versions of the critical consumption/for sentences revealed that participants were more likely to attribute the subinterval property to an event if it involved the consumption of a mass than a counted object, though the difference was only marginal in the items analysis ($t_P(47) = 2.23$, $p < .05$; $t_I(79) = 1.97$, $p = .05$), indicating that they were more likely to treat a consumption/mass action as an activity than a consumption/count action. This pattern occurred despite the presence of the for-adverbial, which should force an atelic interpretation.

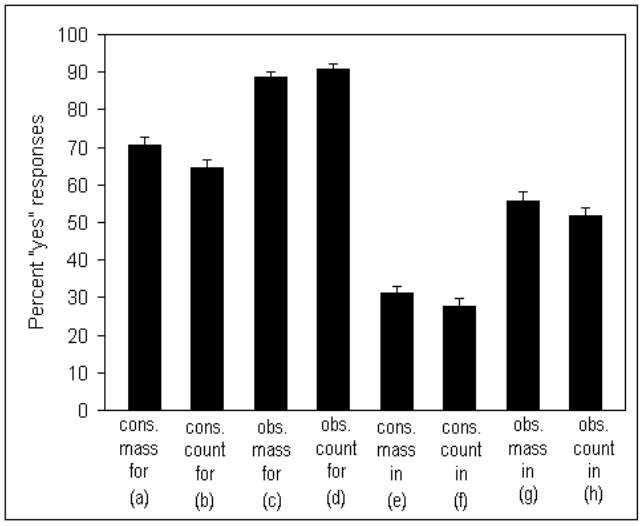


Figure 3: Follow-up Question Responses

General Discussion

An incremental view of language understanding would predict that participants should begin drawing telicity inferences as soon as those inferences are licensed by the text. In the current study, such a view predicts that participants should show slowed reading times at the first segment (the verb segment) for observation verbs (relative to consumption verbs) since such verbs license immediate inferences about the telicity of the events they describe, whereas consumption verbs do not. The present study found little support for this prediction—the main effect of verb found at the first segment was in the opposite direction of that predicted.

An incremental view further predicts an interaction between noun and verb at the third and fourth segments (the noun and manner adverbial segments). If drawing inferences about the telicity of a VP based on verb and noun boundaries were costly, we would expect to see a slowing in reading times as soon as conflicting information appeared (e.g., when a count noun followed a consumption verb). This prediction also failed—there was no such interaction. Nevertheless, participants were using the combined verb and object information to make inferences about the boundedness of events. Both reading time differences at segment five and responses to follow-up questions provide support for these inferences. In the former case, reading times increased when a final adverbial was inconsistent with the verb+object combination. In the latter case, participants were more likely to agree with the subinterval property for the consumption of a mass than for the consumption of a counted object.

Conclusions

Earlier work shows that there is a processing cost for drawing inferences about the telicity of events (Piñango et al., 1999; Todorova et al., 2000). The present study investigated the time course of such inference-drawing. Two strong possibilities presented themselves at the outset: participants could either make inferences early, as the relevant information was presented to them (as in an incremental account) or, alternatively, they could hold off making telicity inferences until late in sentence processing, when all information was available (minimally, until after verb and noun information was available; Pustejovsky, 1991, 1995; Verkuyl, 1993). The reading time results from the present study support the second alternative—there is no indication that participants made rapid use of either the verb or verb+object information to draw boundedness inferences. Instead, it seems that all the costly inferencing work was carried out at the final segment. This finding is consistent with other work: for example, Todorova and her colleagues (2000) find no cost for combining a telic verb (such as *send*) with a bare plural noun (such as *letters*), even though the bare plural forces an atelic interpretation for the verb phrase. The absence of such a cost is surprising under strongly incremental views of natural language interpretation (Marslen-Wilson & Tyler, 1980; Altmann & Kamide, 1999). This pattern is also consistent with the possibility that drawing (or delaying) inferences regarding telicity may be a relatively cost-free process, much as delaying choosing among different metonymic or metaphoric uses of a polysemous noun (such as *newspaper*) (Rayner & Frazier, 1989). In the domain of telicity, participants are willing to hold off on doing costly inferences until they are forced to make an interpretation.

Acknowledgments

This work was supported by NSF grant SES-9907414. We would like to thank Sam Day for his assistance with programming, and three anonymous reviewers for detailed and useful comments on the paper.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247-264.

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy, 9*, 5-16.

Bennett, M., & Partee, B. (1978). *Toward the logic of tense and aspect in English*. Bloomington, Ind: Indiana Linguistics Club.

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Dowty, D. (1979). *Word meaning and Montague grammar*. Boston: D. Reidel Publishing Company.

Jackendoff, R. (1991). Parts and boundaries. *Cognition, 41*, 9-45.

Krifka, M. (1998). The origins of telicity. In S. Rothstein (Ed.), *Events and Grammar*. Boston: Kluwer Academic Publishers.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Langacker, R. W. (1987). Nouns and verbs. *Language, 63*, 53-94.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1-71.

Piñango, M. M., Zurif, E. B., & Jackendoff, R. (1999). Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research, 28*, 395-414.

Pustejovsky, J. (1991). The syntax of event structure. *Cognition, 41*, 47-81.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: The MIT Press.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 556-573.

Solomon, K. O., Proctor, A. S., & Rips, L. J. (in preparation). Concept boundaries: Analogies between objects and events.

Todorova, M., Straub, K., Badecker, W., & Frank, R. (2000). Aspectual coercion and the online computation of sentential aspect. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, (pp. 3-8).

Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, NY: Cornell University Press.

Verkuyl, H. J., (1993). *A theory of aspectuality*. Cambridge, England: Cambridge University Press.

Artefacts as Mediators of Distributed Social Cognition: A Case Study

Jana Rambusch (a00janra@ida.his.se)
University of Skövde, School of Humanities and Informatics
Box 408, 54128 Skövde, Sweden

Tarja Susi (tarja@ida.his.se)
University of Skövde, School of Humanities and Informatics
Box 408, 54128 Skövde, Sweden

Tom Ziemke (tom@ida.his.se)
University of Skövde, School of Humanities and Informatics
Box 408, 54128 Skövde, Sweden

Abstract

Traditionally, cognition has been regarded as the outcome of internal cognitive processes manipulating mental representations. More recently, however, it has become clear that cognition cannot be separated from the social and material environment in which people live and act, and that in many cases cognition is distributed among individuals and environmental properties. One important aspect has turned out to be artefacts and their use, and there is growing interest in understanding how tool use affects cognition. However, even with this increased awareness of the role of artefacts, the focus has mainly been on the cognitive processes and representations of individuals, while the social role of artefacts has received less attention. An ethnographically inspired field study**,** observing a hospital's children admission unit, was conducted to investigate the way individual and collaborative work are affected by the use of artefacts within a given social context. The results indicate that the use of artefacts is closely coupled to the social environment, that to some degree social interactions are transformed into more indirect, individual processes, and that artefacts are crucial for high-level processes such as memory and coordination.

Introduction

Most work in cognitive science has for a long time been based on a general consensus that cognition is best described and analysed in terms of *internal*, often symbolic, representations and computational processes manipulating them. Thus, cognition has been considered to take place largely within the individual mind, with a focus on mental representations and processes, while the environment largely has been reduced to inputs and outputs (e.g., Pylyshyn, 1990). However, since the mid-1980s there has been a growing awareness that individuals are socially and culturally situated and that the environment needs to be considered in order to understand cognition (Clancey, 1997; Clark, 1997; Hendriks-Jansen, 1996; Hutchins, 1995; Suchman, 1987). Humans are, for instance, very proficient in using environmental properties as cognitive aids (Clark, 1997; Kirsh, 1995, 1996), and there is growing interest in finding out how artefacts/tools affect cognition (e.g., Preston, 1998). The terms artefact, tool, and tool use are not particularly well defined, even though there are numerous definitions, resulting from differing focuses in different areas (see, e.g., Gibson & Ingold, 1993; Neuman & Bekerman, 2000; Preston, 1998). In this paper, artefact and tool are used, in accordance with much of the literature, more or less interchangeably.

Despite an increasing interest in cognition and artefacts (see, e.g., Norman, 1993), there is so far a limited understanding of the way artefacts affect the individual within a social context. The present paper aims to contribute to the understanding of the way people are affected by artefacts, and the role artefacts can have in a certain social context. A field study was conducted in a Swedish hospital, at the children's admission, where artefacts (as it turned out) constitute an important part of work tasks. The results indicate that artefacts play an important role in the social context, in a manner different from their role when considered in relation to an individual.

The next section elaborates in some more detail work on situated, distributed and social cognition that constitutes the background for the work of the present paper. Then methods, analysis, and results are described, followed by a discussion.

Situated, Distributed, and Social Cognition

Situated cognition has become an influential approach in many different areas, such as artificial intelligence (e.g., Brooks, 1999), cognitive anthropology (Hutchins, 1995), cognitive psychology (e.g., Barsalou, 1999), and developmental psychology (e.g., Thelen & Smith, 1994). Although there is not yet any universally accepted notion or definition of 'situatedness' (cf. Clancey, 1997; Wilson, 2003; Ziemke, 2002), generally speaking there is an agreement that cognition is a continuous process (e.g., perception-action-loops) with changing boundaries, and that cognition is more than what takes place within the individual mind (e.g., Clark & Chalmers, 1998; Clark, 1999; Susi et al., 2003). The context in which human activities take place is equally important. Hence, there is a growing interest in understanding the role of scaffolds or 'wideware', i.e., external structures such as artefacts, in cognition (e.g., Clark, A., 1997, 1999, 2003; Hutchins, 1995). In itself an

artefact may not be much, but coupled with human cognitive abilities artefacts can become powerful tools, and it has been argued that they extend cognitive abilities such that 'thinking' cannot be reduced to internal cognitive processing (Chalmers & Clark, 1998). Hence, the use of scaffolds or (cognitive) artefacts amplifies cognition. However, it has been pointed out that artefacts actually do not amplify cognition as such (Cole & Griffin, 1980; Hutchins, 1995). Even though a tool may appear to amplify cognition, it is really a coordination of different cognitive processes, which can be aided by using appropriate tools, but no cognitive ability, or process, has been amplified. Other considerations on the topic of artefacts and cognition concern, for instance, that tool use extends the body and a person's body schema (Bateson, 1972; Berti & Frassinettis, 2000; Maravita et al. 2001; Maravita & Iriki, 2004). We are also spatial beings, and, subsequently, all actions are taken in relation to the environment (Kirsh, 1995). People continuously organise and reorganise, for instance, their work environments to reduce the cognitive effort needed.

For a long time tool use and technology were degraded to 'by-products' of cognitive evolution (Saito, 1996), but with increased knowledge it has become clear that artefacts and their use have a considerable effect on cognitive processes. This issue received much attention already by Vygotsky and his followers (see, e.g., Gal'perin, 1969; Haenen, 1996; Vygotsky, 1978, 1981). However, in order to understand this relation further research is needed concerning questions such as what makes an object become a tool, and the development of tool use behaviour (Preston, 1998).

The fact that artefacts, for a long time, have received relatively little attention in cognitive science is somewhat surprising. Nowadays they are commonly described as the "the other major form of cognitive mediation between individual and world" (Preston, 1998, p. 514), besides language, but obviously language has always played a much more central role in cognitive science. Artefacts also played a crucial role, as controlling behaviour from the outside, in the cultural-historical school in psychology, in particular the work of Vygotsky (1978, 1981), but with the advent of cognitive science attention, for the abovementioned reasons, for a long time shifted towards internal, individual processes and representations.

Recent work on *distributed cognition* (Hutchins, 1995) has in some sense rediscovered the integral role that artefacts play in both individual and collaborative cognitive processes that are distributed over people and the material resources they use. This view takes an interest in the way information is represented, transformed, and propagated in the material and social environment. That way, cognitive processes can be described in terms of functional relationships between brains, other people, and external objects. The role of artefacts as mediators of social cognition, however, is far from being fully understood. The present paper considers cognition from a situated and distributed perspective, i.e., it views high-level cognition as resulting from a close interplay between brain, body, and the social and material environment in which humans live and act.

Like most of cognitive science, research in *social cognition* has traditionally focused on individual cognitive processes involving social information, such as attention, perception, and memory, and the internal representations they generate or manipulate (e.g., Augoustinos & Walker, 1995; Fiske & Taylor, 1984; Gilbert, Fiske & Lindzey, 1998). Hence, few studies have taken a situated perspective (Semin & Smith, 2002). However, people are also *socially* situated, which means, one the hand, social interaction between individuals, and, on the other hand, that cognition is situated within a wider social and cultural context (Lave, 1988; Wertsch, 1993; Semin & Smith, 2002). Hence, besides interactional aspects, cognition is also affected by (cultural) artefacts (Levine & Resnick, 1993). Due to the fact that they constitute part of a culture's intellectual history, their use actually turns even seemingly individual activities into a social process as artefacts are affected by social aspects (Resnick, 1993). As pointed out by Hutchins (1995, p. xiv), "human cognition is not just influenced by culture and society, but ... is in a very fundamental sense a cultural and social process".

Case study

Most research on cognition and artefacts has to a large extent focused on the individual (e.g., Norman, 1991, 1993) while contextual and environmental aspects largely have been disregarded**,** with some notable exceptions (e.g. Hutchins, 1995). Subsequently there is a limited understanding as to how artefacts affect the individual within a social context. Further research is needed that considers both social interactions and tool use to gain further understanding of the relation between artefacts, individual, and social cognition. Accordingly, the underlying question for the present study was "*how does tool use affect individual cognitive processes within a social context*"? The term 'artefacts' in this case refers to objects that are significant for everyday work tasks, while 'cognitive processes' here refer to high-level processes such as attention, memory, and coordination.

Method and Setting

In order to investigate the above questions, a field study was conducted at the children's admission in a Swedish hospital. Such work places are indeed highly *social* work places: well functioning daily work requires well-organised cooperation between the members of the staff, as well as between different wards. The work tasks at the children's admission are individual in the sense that most work is carried out individually. For example, parents and children arriving at the admission would usually first meet an administrator who registers their arrival, then a nurse who, for example, might draw a blood sample, and eventually a doctor. At the same time though, all these individual activities are, of course, socially situated and have a strong coupling to the social and environmental context.

In the children's admission, we chose (a part of) the central office (which functions as a communication,

coordination, and administrative centre), as the setting for the study. The office has three units: a reception, an administrative unit, and a unit where all incoming phone calls are handled. The study was limited to the *administrative unit* (see Figure 1), since it is a central place for much of the daily activities, and nurses frequently visit this unit. There are always three to four nurses working at the same time, and the study focused mainly on their work. However, other people also visit the administrative unit during the day, e.g., doctors who come by to collect patient records. The main part of the nurses' work consists of taking care of patients that have an appointment, as well as urgent cases that appear during the day. They also handle administrative tasks, phone counselling, and patient-related tasks, e.g. drawing blood samples. Each member of the staff is responsible for certain tasks, but must also be aware of the others' tasks and responsibilities in order to coordinate their work. On an overall level, the daily routine consists of registering patients on their arrival, and getting each patient's medical record from the archive. Patients that are expected during the day are listed on a patient list. A nurse carries out an initial examination (weight, etc.), and if the patient has a doctor's appointment, he or she is shown into a consultation room. In cases where some kind of a sample needs to be drawn or collected, the doctor notifies a nurse.

The study was inspired by *cognitive ethnography* (Hollan et al. (2000), with observations (moderate participation; DeWalt & DeWalt, 2002), video recordings, and interviews. The combination of such techniques provides a means for gaining insight into the interactions between people and their use of artefacts, and subsequently cognitive processes. The staff was informed that the office was the subject of a study, and the time when it was to take place. Initial interviews were conducted to gain information about the staff, their work tasks, and the setup of the office. Observations were made during two days, and when necessary, questions were asked during the observation. During the observations the office was also videotaped. When all material had been analysed, another interview took place to verify that, e.g., work tasks had been correctly understood.

Analysis and Results

The videotaped material (three hours in total) and notes taken during the observations were analysed from the perspective of the staff's work tasks and activities, and the function of artefacts used in relation to the identified activities.

A highly social work setting, such as the observed children's admission, requires well functioning cooperation, interaction, communication, shared knowledge about routines, others' tasks, etc. In this particular setting the staff uses various artefacts with varying functions, which requires an additional interpersonal understanding of their different functions. In the present study a number of artefacts turned out to be crucial with respect to processes such as coordination of the ongoing work. However, due to space limitations, only a few artefacts are discussed here in some detail.

Most activities in the administrative unit take place around a small table, on which various items are kept and placed (cf. Figure 1). The structure of the office unit also provides structure to the work tasks since the artefacts draw attention to what is going on and what needs to be done.

Patient's record One of the most important artefacts in this particular setting is the patient's medical record. Basically, it is a folder containing a collection of documents with information about a patient. All the documents are sorted in a specific manner in order to reduce the effort of finding the right information (a document out of its proper placement causes disturbances in the work-flow), and when a patient visits the children's admission new information is added. The patient record has several functions, besides the obvious one of storing information about patients. Clearly, no single person could keep all the information in the head, and there is no need to either, as the patient record provides an external memory (on some rare occasion a patient's medical record has been displaced, which caused serious problems). As different people handle the patient record, more information is added to it, and its contents (the representations) become transformed. Commonly, patients do not meet the same nurses and/or doctors each time they come to the children's admission, but information concerning the patient is transferred between staff members through the patient record, which functions as a 'communicator' between different people. That way the contents of a patient's record transform intrapersonal knowledge to interpersonal knowledge shared by several people. The patient records also contribute to an overall coordination of work processes since, depending on where a patient record is placed, it causes different people to take different actions. Placed in the reception's tray, it triggers a nurse to take the patient record into the administrative unit**,** while placed in a tray labelled with a doctor's name it informs the doctor that a patient is waiting, etc.

The patient list Another highly crucial artefact is the *patient list* placed on the wall above the small table (cf. Figure 1) (another list is in the reception, but that was not included in the study). The list actually consists of nine smaller lists (together referred to as the patient list), each corresponding to a consultation room (each doctor uses the same consultation room throughout the day). Thus the list tells, not only who is coming when, but also in which room the patient will be received and by whom. The list also contains information about each patient and the measures to be taken. The list is computer-generated, and during the day the nurses make additional markings by hand. The markings consist of symbols that are typically understood only by the staff. For instance, a certain symbol next to a name on the list means that the patient is waiting in the consultation room (which the list concerns), another symbol means that a doctor has been delayed, and yet another one means that the patient has left the admission. All markings are made in red so that they are highly visible and easy to perceive. During the observation it became clear that the nurses had no difficulties in understanding the added markings, and hence the list provides a means for communication and

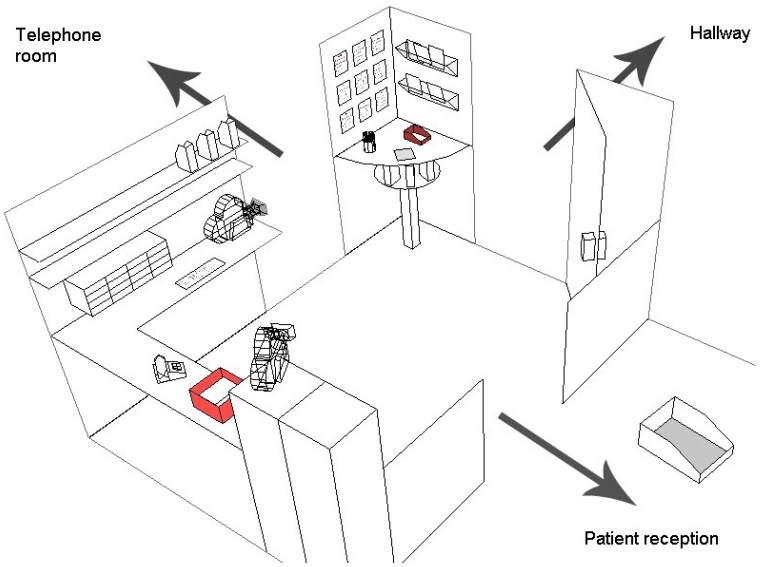


Figure 1. The observed administrative unit. In the upper corner is the small table with a tray (follow-up instructions), and with labelled trays on the shelf below. On the left wall above the table is the patient list(s). On the bench to the lower left is the (orange) tray for filled in sample documents. The cameras indicate the two placements of the (one) camera during the observation.

coordination between people, even when they do not interact directly.

Another function of the list is that it provides a (shared) external memory, providing everyone with the same necessary information. No one needs to memorise the information, as it is 'there' all the time, visible to those who need it, as they need it. As symbols are added to the list, the information becomes transformed and it is propagated from one individual to another when needed. The list also provides an overview and a visualisation of the consultation rooms, and each doctor only needs to pay attention to that part of the list that is related to their room, which in turn delimits the amount of information that needs to be attended to.

Paper trays There are a number of trays[1] in the office, each with its own function(s) and assigned meaning. As patients are registered on their arrival, their medical records are withdrawn and placed in a paper tray in the reception. A patient's record in the tray is a signal to the nurses that a patient has arrived. The patient records are brought by the nurses to the administrative unit, and, eventually, they are placed in trays labelled with the name of the doctor that the patient is going to see. Usually there are four to five doctors working at the same time, and their (labelled) trays are placed on a shelf under the small table (cf. Figure 1). Trays belonging to doctors who are not on duty are placed somewhere else (top shelf to the left of the small table, cf. Figure 1). During the day each doctor collects the patient records that are placed in their tray. When a sample needs to be drawn or collected, the doctor notifies it by leaving follow-up instructions in a blue tray (to the right on the small table in Figure 1). When a nurse has performed the procedure, a filled-in document concerning the sample is placed in an orange tray (on the bench to the lower left in Figure 1). Thus, the spatial arrangement of the trays (and other artefacts) contributes greatly to structuring the ongoing work.

Besides containing documents, the trays have several other functions. For instance, rather than having to keep each ongoing process of the work place in memory, the trays, and their contents, provide information about what is going on and matters that need to be taken care of, thereby providing an external memory. The trays also serve as a means of indirect communication between individuals. The nurses, for instance, do not have to tell a doctor, in person, that a patient is waiting. Instead that information is mediated through the contents of the labelled tray. That way each person can attend his or her own individual work tasks, while at the same time, on an overall level, the indirect communication contributes to a well-functioning operation. The trays also limit the amount of information that needs to considered. A doctor, for instance, only needs to pay attention to *one* labelled tray (or *one* part of the patient list).

Other artefacts There are also a number of other artefacts, equally important but taken for granted to the extent that they become 'invisible' (Gauvain, 2001). One such artefact that should be mentioned here is the small table (Figure 1), which plays a crucial role in the organisation of the daily work. Such a common artefact might seem trivial to discuss, but in this case it is an important part of the spatial arrangement that contributes to the overall structure of work tasks (cf. Kirsh, 1995). People know that it is the place where things that are used or needed often, are kept, and that it is a place for important information. Thus it provides, e.g.,

[1] The items containing documents that are discussed in this paragraph are not all paper trays in the real sense of the word, but are here, for simplicity sake, collected under the label of 'trays'.

an external memory, and people often take a glance at the table before leaving the room, to see if they left anything there as they entered the room (as is often done), or if there is something they need to take care of. The table is also used as a message board, where people leave notes for others.

Discussion

The question guiding our study was "how does tool use affect individual cognitive processes within a social context?". One perspective of interest in the present analysis is the ecological perspective (Gibson, 1986), since an artefact's function is closely related to its appearance (affordance). However, due to space limitations that discussion has been left out in this paper.

Obviously, a two-day study has its limitations, and can only provide a rough understanding of the complexities of a work setting, artefacts, the relations between individuals and social contexts, etc. Nevertheless, despite possible shortcomings, this study is an important step in what we consider an important research direction. Even limited studies can provide valuable insights, and this particular study illustrates some aspects of the relation between artefacts, individuals, and social context.

On a general level, the study shows the importance of the artefacts used in this particular setting, and the way they contribute to a well-functioning operation where much of the activities are coordinated in an implicit manner. It can also be argued that just how powerful artefacts are becomes evident only when considered within the social context in which they are used. Some artefacts, such as a patient list, only make sense to and are understood by those who use them (cf. Levine & Moreland, 1993). As it turned out, it is not only the way an artefact is used that is important for cognition, but also *where* it is used (cf. Kirsh, 1995). The environmental structure and spatial arrangements partly determine the function and the meaning of an artefact. For instance, the labelled trays have different meanings depending on where (on which shelf) they are placed, and patient records trigger different activities depending on where they are placed (cf. also Clark, H., 2003). Artefacts also provide a scaffold for different cognitive processes depending on *who* the user is, i.e., the user's role in the overall social arrangement. Artefacts play an important role as organisers: the state of an artefact (e.g., a tray that is empty, or not) helps the individuals to organise their work, and on the social level they contribute to coordination, cooperation and structure. Some artefacts make information available and visible, and contribute to the propagation of information between people and artefacts.

As discussed by Hollan et al. (2000), the activities of a group cannot be fully understood from the individual's perspective, rather the functional relationships of people and artefacts need to be considered. Thus, individual actions cannot be explained without considering what others are doing, the interpersonal codes, knowledge, and shared understanding of the functions of all the artefacts they use (e.g., Leont'ev, 1978; Thompson & Fine, 1999). Another important aspect is the way artefacts transform individual processes into social processes, and vice versa. For instance, when a nurse adds a marking to the patient list the information becomes part of a social activity and individual knowledge becomes shared knowledge (in a sense propagated 'on demand' only). Likewise, social or shared activities may become individual activities, e.g., when the nurse attends to information added by others.

To summarise, the artefacts analysed in this study function as mediators of distributed social cognition, i.e., they constitute or facilitate shared memory, coordination, communication, and sharing of information. Many artefacts have the same, or similar, functions¸ which however vary depending on *who* is using them, *where* (spatially) they are used, their functional coupling to other artefacts, and the social context. As illustrated in this paper, artefacts in many cases transform social interactions into individual processes, but at the same time they also mediate the indirect interaction of these processes¸ and thus maintain their social nature (cf. Susi & Ziemke, 2001).

This study has illustrated that, in order to understand artefacts and their role in individual and social cognitive processes, we need to consider artefacts, individuals, the social context, and their functional interrelations. Much of the work addressing these issues, including this paper, has been presented in the form of examples, anecdotal evidence, case studies, etc. Future work will have to further address the development of a more systematic, principled understanding of the role that artefacts play in distributed social cognitive processes.

Acknowledgements

The authors would like to thank the employees of the children's admission at the hospital in Skövde (Kärnsjukhuset) for participating in the case study.

References

Augoustinos, M. & Walker, I. (1995). *Social cognition: An integrated introduction*. London: SAGE Publications.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-609.

Bateson, G. (1972). *Steps to an ecology of mind*. Northwale, New Jersey: Jason Aronson Inc.

Berti, A. & Frassinetti, F. (2000). When far becomes near: remapping of space by tool use. *Journal of Cognitive Neuroscience*, 12(3), 415-420.

Brooks, R. (1999). *Cambrian Intelligence*. Cambridge, MA: MIT Press.

Clancey, W. J. (1997). *Situated Cognition*. New York: Cambridge University Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A. (1999). Where brain, body, and world collide. *Cognitive Systems Research*, 1, 5-17.

Clark, A. (2003). *Natural born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford: Oxford University Press.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 56, 10-23.

Clark, H.H. (2003). Pointing and placing. In: S. Kita (Ed.), *Pointing - Where language, culture, and cognition meet*. Hillsdale, NJ: Lawrence Erlbaum.

DeWalt, K.M. & DeWalt, B.R. (2002). *Participant observation: A guide for field-workers*. Walnut Creek: Altamira Press.

Fiske, S.T. & Taylor, S.E. (1991). *Social cognition*. McGraw-Hill, New York:

Gal'perin, P.Y. (1969). Stages in the development of mental acts. In: M. Cole & I. Maltzman (Eds.), *A handbook of contemporary Soviet psychology*. New York: Basic Books, Inc,. Publishers.

Gauvain, M. (2001). Cultural tools, social interaction and the development of thinking. *Human Development*, 44, 126-144.

Gibson, J.J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibson, K.R. (1993). General introduction: Animal minds, human minds. In: K.R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution*. Cambridge, MA: Cambridge University Press.

Gibson K.R. & Ingold, T. (Eds.) (1993). *Tools, language and cognition in human evolution*. Cambridge, MA: Cambridge University Press.

Gilbert, D.T., Fiske, S.T. & Lindzey, G. (1998). *The handbook of social psychology*, Vol. 1 & 2. New York: McGraw-Hill.

Haenen, J. (1996). *Piotr Gal'perin: Psychologist in Vygotsky's footsteps*. New York: Nova Science Publishers, Inc.

Hendriks-Jansen, H. (1996). *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought*. Cambridge, MA: MIT Press.

Hollan, J., Hutchins, E. & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7(2), 174-196.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, 73, 31-68.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, 4(3/4), 415-452.

Lave, L.B. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge: Cambridge University Press.

Leont'ev, A.N. (1978). *Activity, Consciousness, and personality*. Englewood Cliffs, NJ: Prentice-Hall.

Levine, J.M. & Moreland, R.L. (1993). Culture and socialization in work groups. In: L.B. Resnick, J.M. Levine & S.D. Teasley (Eds.), *Perspectives on socially shared cognition*. Washington: American Psychological Association.

Levine, J. M., Resnick, L. B. & Higgins, E. T**.** (1993). Social foundations of cognition. *Annual Review of Psychology*, 44, 585-612.

Maravita, A. & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences*, 8(2), 79-86.

Maravita, A., Husain, M., Clarke, K. & Driver, J. (2001). Reaching with a tool extends visual-tactile interactions into far space: Evidence from cross-modal extinction. *Neuropsychologia*, 39, 580-585.

Neuman, Y. & Bekerman, Z. (2000). Where a blind man ends: Five comments on context, artifacts and the boundaries of the mind. *Systems Research and Behavioral Science*, 17, 315-319.

Norman, D. (1991). Cognitive artifacts. In: J.M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface* (reprinted edition, 1993). Cambridge: Cambridge University Press.

Norman, D. (1993). *Things that make us smart: Defending human attributes in the age of the machine.* Reading, MA: Addison-Wesley.

Preston, B. (1998). Cognition and tool use. *Mind & Language*, 13(4), 513-547.

Pylyshyn, Z.W. (1990). Computation and cognition. In: J.L. Garfield (Ed.), *Foundations of cognitive science*. New York: Paragon House.

Resnick, L. B. (1993). Shared cognition: Thinking as social practice. In: L.B. Resnick, J.M. Levine & S.D. Teasley (Eds.), *Perspectives on socially shared cognition*. Washington: American Psychological Association.

Saito, A. (1996). Social origins of cognition: Bartlett, evolutionary perspective and embodied mind approach. *Journal for the Theory of Social Behavior*, 26(4).

Semin, G.R. & Smith, E.R. (2002). Interfaces of social psychology with situated and embodied cognition. *Cognitive Systems Research*, 3(3), 385-396.

Suchman, L.A. (1987). *Plans and situated actions: The problem of human machine communication*. Cambridge, MA: Cambridge University Press.

Susi, T., Lindblom, J. & Ziemke, T. (2003). Beyond the bounds of cognition. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum

Susi, T. & Ziemke, T. (2001). Social Cognition, Artefacts, and Stigmergy. *Cognitive Systems Research*, 2(4), 273-290.

Thelen, E. & Smith, L. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Vygotsky, L.S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, MA: Harvard University Press. Original work published 1932.

Vygotsky, L.S. (1981). The instrumental method in psychology. In: J.V. Wertsch (Ed.), *The concept of activity in Soviet psychology*. Armonk, NY: M.E. Sharpe, Inc. Original work published 1960.

Wertsch, J.V. (1993). A sociocultural approach. In: L.B. Resnick, J.M. Levine & S.D. Teasley (Eds.), *Perspectives on socially shared cognition*. Washington: American Psychological Association.

Wilson, M. (2003). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9(4), 625-636.

Wynn, T. (1993). Layers of thinking in tool behavior. In: K.R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution*. Cambridge, MA: Cambridge University Press.

Ziemke, T. (2002). Introduction to the special issue on situated and embodied cognition. *Cognitive Systems Research*, 3(3), 271-274.

Making Graphical Inferences: A Hierarchical Framework

Raj M. Ratwani (rratwani@gmu.edu)
George Mason University

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
Naval Research Laboratory

Abstract

A hierarchical framework suggesting how graph readers go beyond explicitly represented data to make inferences is presented. According to our hierarchical framework, graph readers use read-offs, integration and pattern extrapolation to make inferences. Verbal protocol data demonstrates high-level differences in the way inferences are made and eye track data examines these processes at the perceptual level.

Introduction

Imagine a scientist examining Figure 1 in order to infer which county in California is going to be hit next by the flu epidemic. How would the scientist go about making this prediction?

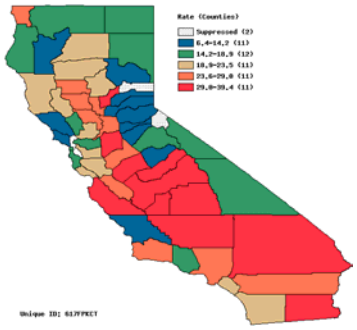


Figure 1. Cases of the flu in California.

Making inferences from graphs is considered one of the more complex skills graph readers should possess. According to the National Council of Teachers of Mathematics (NCTM) the simplest type of question involves the extraction or comparison of a few explicitly represented data points (read-offs) (*NCTM: Standards for Mathematics*, 2003). A more difficult question is an integration question where multiple data points need to be extracted and integrated by some mental operation. The most difficult type of question requires the graph reader to make inferences from the graph. Because the information is not explicitly represented in the data, the graph reader is forced to extrapolate from the current data to make a prediction (Trickett, Ratwani, & Trafton, under review).

How do graph readers go beyond the explicitly represented data to make inferences from graphs? Less is known about how inferences are made from graphical representations, despite the importance of having this skill. Most of the classical theories of graph comprehension (Kosslyn, 1989; Lohse, 1993; Pinker, 1990) do not go into detail about how integration and inferences are made, but instead focus on read-offs from fairly simple graph types. One of the reasons that current theories of graph comprehension do not have much to say about making inferences from graphs is the paucity of data. There are, in fact, very few empirical papers that have systematically investigated how graph readers make inferences from graphs.

We propose a hierarchical framework of graph comprehension for how these different types of questions (read-off, integration, inference) are answered. The most basic type of information extraction is the read-off of explicitly represented data. The more difficult integration of information requires the use of read-off's and spatial transformations (Trafton, Marshall, Mintz, & Trickett, 2002; Trickett & Trafton, in press). For example, in order to integrate information in choropleth graphs (see figures 1 and 2), graph readers read-off specific data points and use spatial transformations by forming clusters of proximate same colored counties and then reason with and compare those clusters (Ratwani, Trafton, & Boehm-Davis, 2003). Finally, in order to make inferences from graphs, we believe graph readers use the same processes used to integrate information (read-off's and spatial transformations) and in addition use pattern extrapolation and mental models (Trafton et al., 2002).

Going beyond the limits of the current data in order to make an inference requires the use of extrapolation (Bott & Heit, 2004); when graph reader's go beyond the limits of visible data, pattern extrapolation may be used. Pattern extrapolation is a process by which graph readers examine known data points and then, based on the pattern of these data points, make an inference.

While the hierarchical framework suggests what cognitive processes graph readers will use when extracting different types of information from graphs, graph readers are likely to use the simplest process to extract the information they desire. For example, when integrating information, if possible, graph readers will use mostly read-offs because read-offs are a simple way of extracting information from graphs and require very little cognitive effort in comparison to spatial transformations or mental model building. Similarly, if a graph reader needs to integrate information from a graph, they are not likely to need to build mental models and extrapolate patterns.

In this paper, we examine which processes are used to make inferences from graphs. We focus on inferences because read-offs are quite well understood (Kosslyn, 1989; Lohse, 1993; Pinker, 1990), there is a preliminary framework for integration of graphical information (Ratwani, Trafton, & Boehm-Davis, 2003), but there are no theories that can adequately describe how inferences are made. Previous research examining graphical inferences has focused on the use of mental models, spatial transformations

and the role of domain knowledge (Hegarty, Shimozawa, & Canham, under review; Trafton et al., 2000). We are interested in how all of these processes are combined in order to make inferences from graphs. In this paper, we focus on read-offs, integration and pattern extrapolation because it is relatively straightforward to identify read-offs, integration, and pattern extrapolation and much more difficult to identify spatial transformations or mental model building.

If the hierarchical framework of graph comprehension is correct, graph readers will make inferences by reading-off explicit information, using pattern extrapolation and integrating information. Experiment 1 serves to explore higher-level thinking about how graph readers make inferences from graphs by using the protocol analysis methodology. Experiment 2 further investigates the processes used to make inferences by using an eye tracker to examine graph readers' eye movements.

Experiment 1

The first experiment was designed to explore the types of processes graph readers use to make inferences from multiple choropleth graphs. Choropleth graphs depicting population densities were selected for use in this experiment; these graphs use different colors, shades of gray, or patterns to represent different quantities. Choropleth graphs were chosen for multiple reasons. First, they are more complex than the graph types used in more traditional studies of graph comprehension and better reflect how graphs are used in the real-world. Second, these particular graphs do not require a great deal of domain knowledge and can be presented to undergraduates without much training. Finally, choropleth graphs represent a class of graphs that are commonly used by scientists in such domains as meteorology, geology and oceanography.

Method

Participants

Three George Mason University undergraduate psychology students served as participants for course credit. Informed consent was received from all participants.

Materials

Twenty sets of choropleth graphs were created; each set consisted of three conceptually related graphs. The graphs in each set displayed the population densities of fifty fictitious counties. The first graph in each set displayed the population from the year 1990, the second graph displayed 1995 and the third graph displayed the population from the year 2000 (See Figure 2 for an example). Only one county in each set of graphs was labeled with a county name (referred to as the target county) in order to reduce search time. Previous studies found that graph readers spent a great deal of time searching for the county of interest when every county was labeled (Ratwani et al 2003). One inference question was asked of each set of graphs: What will the population of the target county be in the year 2005?

Design

Five sets of graphs showed a clear decrease in the overall population densities from 1990 to 2000 while the population of the target county did not change in any of these graphs. These counties surrounding the target county had a powerful contextual indication that the population was decreasing. Five sets of graphs showed a clear increase in the overall population densities from 1990 to 2000 while the population of the target county did not change. These counties surrounding the target county had a powerful contextual indication that the population was increasing. Ten of the sets of graphs served as fillers and were removed from all analyses; the populations were jumbled and had increasing, decreasing or no clear pattern to the population movement across the graphs. The purpose of these sets was to randomize the patterns of increasing and decreasing population in the ten sets of interest. The order in which the twenty sets of graphs were presented was randomized for each participant. The increasing and decreasing sets were combined in the analyses below.

Procedure

All participants first read the question for the set of graphs they were about to view and then examined the graphs. For example, the participant would read the question, "What will the population of county x be in the year 2005?" After reading the question the participant would then view the graphs from the three time periods and make their prediction. This process continued for each of the twenty sets of graphs.

The participants could view each of the graphs for as long as they desired, and the participants were permitted to look back to any of the graphs within a particular set as needed. Each graph was presented on a single sheet of paper. After answering each question, the participant went on to the next set of graphs. Each participant provided a talk-aloud protocol (Ericsson & Simon, 1993) as they examined the graphs and answered the questions. The participants' verbal protocols and the graphs they were examining were videotaped.

Coding Scheme

Transcriptions of the verbal protocols were made prior to data analysis. The first step was to segment the protocols into individual utterances. Utterances were defined as a single thought and utterances that were not germane to the task at hand were coded as "off task" and eliminated from further analysis. Each remaining utterance was then coded according to our hierarchical framework. The utterances were coded as either being a **target read-off** (extracting information regarding the target county of interest only) or **integrative** (extracting general trend information from the graph). All the answers were, by definition, inferences. There were no non-target read-off's in any of the utterances made by the participants. Table 1 shows examples of each utterance type.

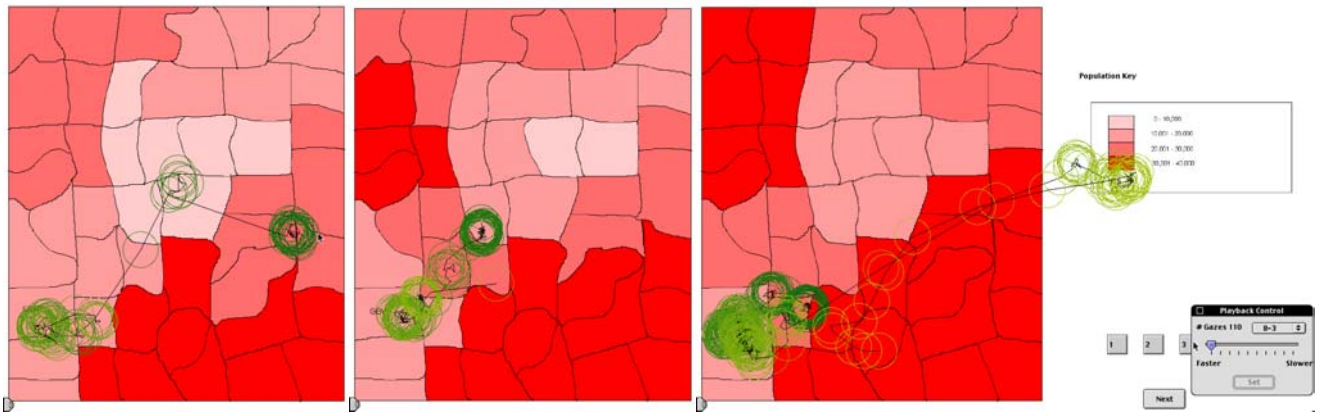


Figure 2. Graphs depicting population growth from 1990, 1995, and 2000 (left to right).

Code	Example
Target read-off	In 1990 Stow County was 20,000 to 30,000
Integrative	All the other areas are increasing in population

Table 1: Examples of each extraction type.

Results and Discussion

Our main goal was to explore how graph readers made inferences from graphs. We first examined the types of extractions made by each graph reader and then compared these extractions to the answer given to the inference questions. The participants gave a numerical answer indicating that the population of the target county would either change or not change. Of the three participants, one participant made change responses the majority of the time, one participant made non-change responses the majority of the time and one participant had mixed responses. Thus, graph readers were not always using the same strategy to make these inferences. When participants made a change response, their inference was in the direction consistent with the surrounding counties. For example, when the participant made a change response and inferred that the population of the target county would grow in the future, the surrounding counties were also growing.

As figure 3 suggests, when graph readers said the population of the target county would not change in the future all of their extractions were target read-offs. When graph readers said that the target county would change in the future, the graph readers made some target read-offs but made a significantly greater number of integrative extractions, $\chi^2(1) = 4.9$, $p < .05$. In addition, when a non-change response was given, graph readers made a significantly greater number of target read-offs than when a change response was made, $\chi^2(1) = 8.02$, $p < .01$.

The verbal protocol data indicates that when graph readers performed mostly target read-off's they made a non-change response to the inference questions. That is, despite the fact that the powerful overall context of the three graphs suggested that the population was increasing (for example), the graph reader inferred that the population of the target county would not change in the future. However, when graph readers looked beyond the target and used the global context of the graphs the graph readers used context to infer that the growth would continue to the target county and that the target county would change in the future.

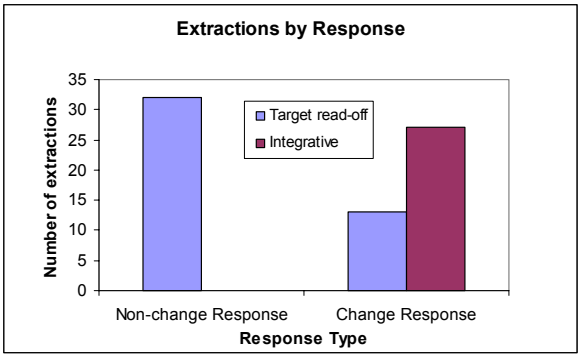


Figure 3: Number of extractions by response.

These data suggest there are differences in the way people think about making inferences. Based on the type of extractions the graph reader made, their response could be categorized as either inferring a change or not inferring a change in the future population. There appear to be two general ways in which graph readers made inferences from these graphs. One way was to focus only on the target county in each of the three time periods and, based on how the population changed in the target county, an inference was made as to the future population. For example, if the target county did not change population in any of the three time periods then it would not change in the future. This no change response appears to be based solely on pattern extrapolation of the target county. Alternatively, when making change responses, graph readers appeared to be making read-off's for pattern extrapolation and taking into

consideration the contextual influences of the graph. Some aspects of the global context of the graph were being integrated with the population of the target county in order to make the inference. When participants made change responses, their verbal protocol data is consistent with an interpretation of them creating a dynamic mental model: participants were imagining the growth in the counties extending to nearby counties, eventually "hitting" the target county. While we are not directly measuring mental model formation in this paper, we are interested in what information is needed to form those mental models.

Based on our hierarchical framework, we would expect graph readers to make inferences by both reading-off information for pattern extrapolation and integrating information. It appeared that when graph readers made a non-change response, they extracted target information only, noticed it did not change, and extrapolated that it would not change in the future. They did not seem to explicitly extract information from nearby counties. Graph readers who made a change response appeared to be using read-offs, pattern extrapolation and integration as our hierarchical framework suggests.

Experiment 1 showed that people had different strategies when answering inference questions: a change strategy and a non-change strategy. It could be that at the perceptual level these strategies are identical. For example, it could be that participants who made a no-change answer did, in fact, look all over the graph but decided to simply ignore that information, or assume that the target county was the most important determinant of future change. Additionally, the protocol data did not show how or what types of information was extracted by change-response participants. Experiment 2 investigates these issues.

Experiment 2

How, then, did participants make inferences from these graphs? By performing a small task analysis, it is obvious that when information needs to be integrated, it can be integrated in at least two ways: within a specific graph and between related graphs. If a participant integrates information within a specific graph, the participant would presumably examine nearby counties to see how their population was different from the target county. If a participant integrates information between related graphs, the participant would probably examine graphs that had changed over time.

Experiment 2 will explore three main issues. First, do participants who answer change and non-change have different perceptual strategies? Second, what types of integration do change participants engage in (within graphs, between graphs, or both)? Third, what is the proportion of read-offs and integration used in order to answer inference questions and how do those proportions relate to the answers that participants gave?

Method

Participants

Thirteen George Mason University undergraduate psychology students served as participants for course credit. Informed consent was received from all participants.

Materials

The same sets of graphs used in the first experiment were used in the second experiment. In this experiment the materials (graphs and questions) were displayed on a computer screen. Eye track data was collected using an LC Technologies Eye gaze System eye tracker operating at 60Hz (16.7 samples/second).

Design

The design was the same as Experiment 1.

Procedure

The procedure was very similar to that used in Experiment 1; however, the use of the computer and eye tracker did necessitate some changes. The participants were seated at a comfortable distance from the monitor and used a chin rest. Participants first were calibrated on the eye tracker. Participants were then shown the question at the top of a blank screen and read the question out loud. Previous studies (Ratwani et al., 2003) have shown that the process of collecting eye track data was not hindered by the participant talking. After reading the question the participant proceeded to the first graph. The interface allowed the participants to progress from graph to graph within a set with a button-click. The participants were instructed to say their answer out loud when they made their inference. After answering the question, the participant could progress to the next question and set of graphs.

Coding Scheme

A gaze was defined from each sample being no more that 10 pixels in Euclidian distance from the center of gravity of the previous point for at least 100 milliseconds. Frequencies were created by counting the number of gazes to different areas of the graph. The areas of the graph that were coded were: the legend, the title of the graph, and the main part of the graph itself.

Participant's gazes to the main part of the graph were coded to examine how much reading-off and how much integrating the participants were doing. Gazes to the target county were coded to examine how often participants were making read-offs. Gazes to locations other than the target were coded in order to examine whether graph readers were integrating information within the graph. In order to capture how far away from the target county participants were gazing the location of the gaze relative to the target county was coded for. For example, if a participant gazed at a county that was three counties away from the target this distance was coded.

Integration of information between graphs was coded by examining areas of change relative to the previously viewed

graphs. If a gaze to the graph was to a county where the population value changed from a previously viewed graph this was coded for. For example, if the participant gazed to a county in 1995 that had changed in population relative to the map from 1990 then this was coded as a change gaze. Thus, the first map viewed by each participant in every set did not have any change gazes.

Results and Discussion

Experiment 2 was designed to examine what processes occurred at the perceptual level when graph readers made inferences from graphs. Specifically, we wanted to further investigate the process differences when graph readers made a change response as compared to when they made a non-change response.

The responses made by graph readers were mixed, but mostly (76%) change responses were made. The raw frequencies of gazes were normalized by dividing the frequency of gazes by the number of responses in either the non-change or change category.

There were no significant differences in the number of target gazes when participants made a non-change response as compared to a change response as figure 4 suggests, $\chi^2(1) = .34$, $p = .56$. Participants appeared to be reading off the same amount of target information regardless of what type of response they made. Thus, participants were reading off information nearly equally when making inferences.

However, participants who made change responses did more integration within the graph than participants who did not make change responses. As figure 4 suggests, those who made change responses on average made a greater number of gazes to counties other than the target, $\chi^2(1) = 4.52$, $p < .05$.

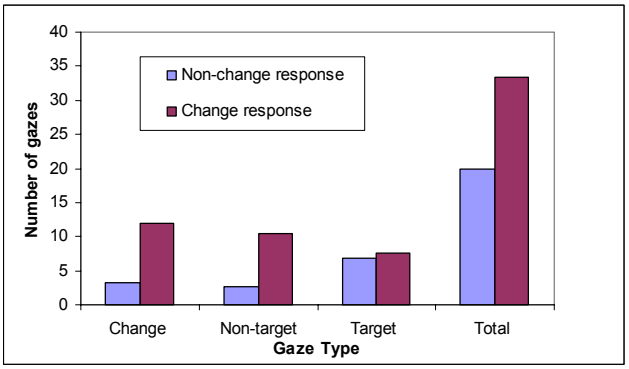
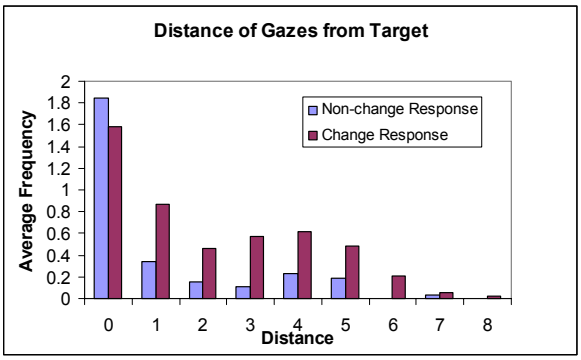


Figure 4. Number of coded gazes.

How far away from the target did participants look? In order to examine this issue, we created histograms showing the location of the counties that were gazed at based on the participants response. Figures 5 shows the frequency of gazes participants made to the target and to counties other that the target. The x-axis shows how far away the county gazed at was from the target. Zero represents the target and one through eight represent how far away the county gazed at was from the target. The patterns in these histograms are significantly different, $\chi^2(8) = 22.82$, $p < .01$, suggesting that when graph readers made a change response, they frequently looked at the target and counties away from the target, whereas graph readers who made a non-change response focused primarily on the target. Consistent with this interpretation, the proportion of change gazes to non-targets (68%) was far greater than the proportion of non-change gazes to non-targets (37%), $\chi^2(1) = 9.2$, $p < .005$. Graph readers who made a change response were frequently looking at counties as far as 6 away from the target county.

Did graph readers integrate information between graphs when they were making inferences? Integration between graphs was examined by looking at the number of gazes to areas of change from one graph relative to another. As figure 4 suggests, participants who made change responses made a significantly greater number of gazes to areas of change as compared to participants who made non-change responses, $\chi^2(1) = 4.88$, $p < .05$. This suggests participants who made change responses were integrating information between graphs by comparing the areas that changed in population.



Figures 5. Histogram of distance by response.

The process of integrating information between graphs is further supported be examining the number of times graph readers examined each of the three graphs. For example, some participants viewed each graph once; the sequence of graphs they looked at was 1→2→3. Whereas other graph readers examined each graph more than once and had a sequence such as 1→2→1→2→3→2→3. Participants who made change responses looked at the three graphs in each set more often in order to compare the counties that changed population between graphs, $\chi^2(2) = 5.24$, $p < .05$. Thus, graph readers who made change responses integrated information between graphs by paying attention to areas that changed in the graphs and looking at the graphs frequently in order to make these comparisons.

When participants made non-change responses, their eye movements suggest they are primarily examining the target county. These participants generally looked at each map only once. Furthermore, they appeared to be reading-off target county information in each graph, noticing the pattern does not change, and then using pattern extrapolation to infer that the pattern will not change in the future.

Graph readers making change responses appeared to read off target county information and also focused a great deal of attention on non-target counties. These gazes to non-target counties appeared to be a way to integrate the information from the other counties with the information about the target county. These graph readers also integrated information between graphs by paying attention to areas of change between the graphs. Finally, they compared areas of change to infer the future population by looking back and forth at the graphs in each set. This is suggestive evidence of the formation of dynamic mental models which may be used to understand how the contextual growth or decay is influencing the target county.

General Discussion

How do people make inferences from graphs? Most classic theories do not provide any mechanisms for making graphical inferences. These studies examined inferences at a high-level by focusing on graph reader's thought processes with the verbal protocol data and also at the perceptual level by examining graph reader's eye movements. These studies demonstrate that people certainly can make inferences, and that people make inferences in different ways. One way that people make inferences is to examine the specific object that will change over time (target county in our case). Depending on the type of change that is observed, a pattern is extracted and then extrapolated. In our studies, approximately a quarter of the answers conformed to this strategy. The remainder took context into account. That is, they observed the surrounding counties (especially the ones that changed from graph to graph) and presumably imagined the change affecting the target county.

Our hierarchical framework of graph comprehension is consistent with both views, though it is supported more strongly by the participants who made change answers. In order to make inferences, the hierarchical view framework suggests that people need to extract specific information from graphs (well described by most theories of graph comprehension), integrate information into a reasonable whole (in this case by combining information between and within graphs), use that information to extrapolate beyond the given data, mentally manipulate the graphical information by spatial transformations and build mental models. It is interesting that when non-change answers were made, only a subset of this framework was used: the evidence for integration in particular was quite weak. It seems that when non-change answers were given, participants simply took in the specific information for the target county, performed simple extrapolation, and then gave an answer. Using the surrounding counties was just not a priority for these participants.

Finally, how inferences are made from graphs is more complex than we have described here. Our hierarchical framework identifies the processes used to make inferences; however, further empirical data is needed to understand how read-offs, integration, spatial transformations, mental models and pattern extrapolation are combined in the process of making inferences. In addition the processes outlined by our hierarchal framework are likely to be dependent on many factors such as knowledge of the graphical display and domain knowledge (Hegarty et al., under review).

Acknowledgements

This research was supported in part by grant 55-7850-00 to the second author from the Office of Naval Research and by George Mason University. We thank Mike Schoelles for designing the interface of the eye tracker.

References

Bott, L., & Heit, E. (2004). Nonmonotonic Extrapolation in Function Learning. Journal of Experimental Psychology: Learning, Memory & Cognition, 30(1), 38-50.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. (2nd ed.). Cambridge, MA: MIT Press.

Hegarty, M., Shimozawa, N., & Canham, M. (under review). Inferences from Graphics: Do you need a Weatherman to Tell which Way the Wind Blows?

Kosslyn, S. M. (1989). Understanding charts and graphs. Applied Cognitive Psychology, 3, 185-225.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. Human Computer Interaction, 8, 353-388.

NCTM: Standards for Mathematics. (2003, 2003).

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), Artificial intelligence and the future of testing (pp. 73-126). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2003). Thinking graphically: Extracting local and global information. In R. Alterman & D. Kirsch (Eds.), 25th. Annual Meeting of the Cognitive Science Society. Boston, MA: Erlbaum.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: Extracting and generating information from complex visualizations. International Journal of Human Computer Studies, 53(5), 827-850.

Trafton, J. G., Marshall, S., Mintz, F., & Trickett, S. B. (2002). Extracting explicit and implicit information from complex visualizations. In M. Hegarty, B. Meyer, & H. Narayanan (Eds.), Diagrammatic Representation and Inference (pp. 206-220). Berlin Heidelberg: Springer-Verlag.

Trickett, S. B., & Trafton, J. G. (in press). Spatial Transformations in Graph Comprehension. Diagrammatic Representation and Inference, 3rd International Conference, Cambridge, U.K.

Trickett, S. B., Ratwani, R. M., & Trafton, J. G. (under review). Real World Graph Comprehension: High-Level Questions, Complex Graphs, and Spatial Cognition.

Implicit and Explicit Learning of a Covariation Across Visual Search Displays

Colleen A. Ray (caray@u.arizona.edu)
Department of Psychology, 1503 E University Blvd. Building 68
Tucson, AZ 85721 USA

Eyal M. Reingold (reingold@psych.utoronto.ca)
Department of Psychology, 100 St. George Street
Toronto, Ontario M5S 3G3 Canada

Abstract

The goal of this study was to extend prior reports of implicit learning in visual search (e.g., Chun & Jiang, 1999) by employing eye movement monitoring and reaction time measures to contrast implicit and explicit learning. Towards this end, participants' eye movements were monitored as they performed a visual search task in the 'change blindness' flicker paradigm. In each trial, participants were asked to detect a letter that differed in shape or color across otherwise identical alternating letter arrays. In a subset of trials, for some participants the background luminance covaried with target color (Color rule condition) and for other participants letter thickness covaried with target shape (Shape rule condition). In addition, half of the participants were told of the existence of a covariation (Informed group) and the other half were not notified of this regularity and in a post-experimental interview reported no awareness of this covariation (Uninformed group). In both groups, reaction time data indicated that visual search was facilitated for trials that contained the covariation, and eye movement data showed that participants guided eye movements to potential targets based on the covariation information. Further, Informed participants in the Color rule condition were able to use covariation information to a greater extent than those in the Shape rule condition. In contrast, no differential sensitivity across rule conditions was found for Uninformed participants. Implications to the study of implicit learning in visual search are discussed.

Introduction

Cognitive psychologists regularly differentiate cognitive processing as being either implicit or explicit in nature. However, there still exists debate in the field over whether these processes are indeed distinct (e.g., Shanks and St. John, 1994). One way to strongly support such a separation would involve demonstrating qualitative differences between these processes (e.g., Cheesman & Merikle, 1986; Dienes & Berry, 1997; Dixon, 1981; Neal & Hesketh, 1997; Reingold & Merikle, 1990; Shevrin & Dickman, 1980; Stadler & Frensch, 1994). As such, the present investigation attempts to make progress towards this goal by extending research demonstrating implicit learning in visual search (e.g., Chun & Jiang, 1999; Chun & Nakayama, 2000; Durgin, 1999; Flowers & Smith, 1998; Miller, 1991). For example, in a study by Chun and Jiang (1999), participants performed a visual search task where they were instructed to look for an object symmetric around the vertical axis. There were two display conditions: in the first condition each of the targets was paired with a distractor set and this pairing was preserved throughout the experiment, and in the second condition the pairings of distractor sets and targets was varied randomly across trials. Chun and Jiang (1999) reported that search efficiency in the consistent pairing condition was significantly better than in the randomized pairing condition. Importantly, an evaluation of participants' awareness of the experimental manipulation indicated that the benefit participants derived from consistent pairing reflected implicit rather than explicit learning. Based on these results Chun and Jiang argued that following implicit learning of the pairing of targets and distractor sets, distractor identity information (i.e., context information) could cue knowledge of target shape information, reducing search time.

In the present study we employed eye movement monitoring as an index of learning in addition to reaction time. Eye movements have been shown to be a tangible trace of perceptual and attentional processes, and represent an ideal indirect measure of processing which can be recorded unobtrusively, concurrent with direct discrimination measures of performance. This may allow for demonstrating dissociations between some aspects of eye movement behavior, such as saccadic selectivity, and overt task performance, such as reaction time. Employing such methodology, this research investigates whether we can learn to make a more efficient visual search based on a covariation that *across* search displays provides target identity information, as well as addressing how awareness of this information affects search performance. In addition, by employing multiple methodologies to empirically investigate implicit and explicit learning, we hope to make progress towards the goal of demonstrating qualitative differences between these processes.

To this end, we exploited a well-established finding in the eye movement and visual search literature that demonstrates a bias in the distribution of saccadic endpoints toward distractors that share stimulus dimensions and features with the target including color, shape, contrast polarity, and size (e.g., Findlay, 1997; Hooge & Erkelens, 1999; Pomplun, Reingold, & Shen, 2001, 2003; Pomplun, Reingold, Shen, & Williams, 2000; Scialfa & Joffe, 1998; Shen & Reingold, 1999; Shen,

Reingold, & Pomplun, 2000, 2003; Shen, Reingold, Pomplun, & Williams, 2003; Williams & Reingold, 2001). This finding is typically referred to as saccadic selectivity.

In the present study participants performed a visual search task in the 'change blindness' flicker paradigm (Rensink, O'Regan, & Clark, 1997). As shown in Figure 1, participants were asked to detect a letter (target) that differed in color (see Panel A) or shape (see Panel B) across otherwise identical alternating letter arrays. In a subset of trials, there was a covariation embedded in the task that reduced the target to half of the display items. Reaction times for the trials containing the covariation (Covariant trials), were compared to the reaction times obtained for the trials that did not contain covariation information (Random trials). In addition, half of the participants were told of the existence of a covariation (Informed group) and the other half were not notified of this regularity (Uninformed group). Learning of the covariation rule would be expressed by faster reaction times for the Covariant trials than for the Random trials. In addition, learning of the covariation rule would provide target identity information and may be manifested as a bias towards fixating on the letters that have the same shape or color as the target (i.e., saccadic selectivity). Finally, we were also interested in determining if there would be differential search performance across the Informed and Uninformed groups reflecting qualitative differences between implicit and explicit learning.

Methods

Participants

Forty-eight participants were paid $20 for their involvement in the two hour experiment: 24 participants searched for the letter that differed in shape and 24 participants searched for the letter that differed in color. Half of the participants in each group were informed of the covariation information (Informed group), while the other half were not (Uninformed group). Participants were tested individually, and all had normal or corrected-to-normal vision.

Apparatus

The eyetracker employed in this research was the SR Research Ltd. EyeLink system. This system has high spatial resolution (0.005°), and a sampling rate of 250 Hz (4 msec temporal resolution). The EyeLink headband has three cameras, allowing simultaneous tracking of both eyes and of head position for head-motion compensation. By default, only the participant's dominant eye was tracked in our study. The EyeLink system uses an Ethernet link between the eyetracker and display computers for real-time saccade and gaze position data transfer. In the present study the configurable acceleration and velocity thresholds were set to detect saccades of 0.5° or greater.

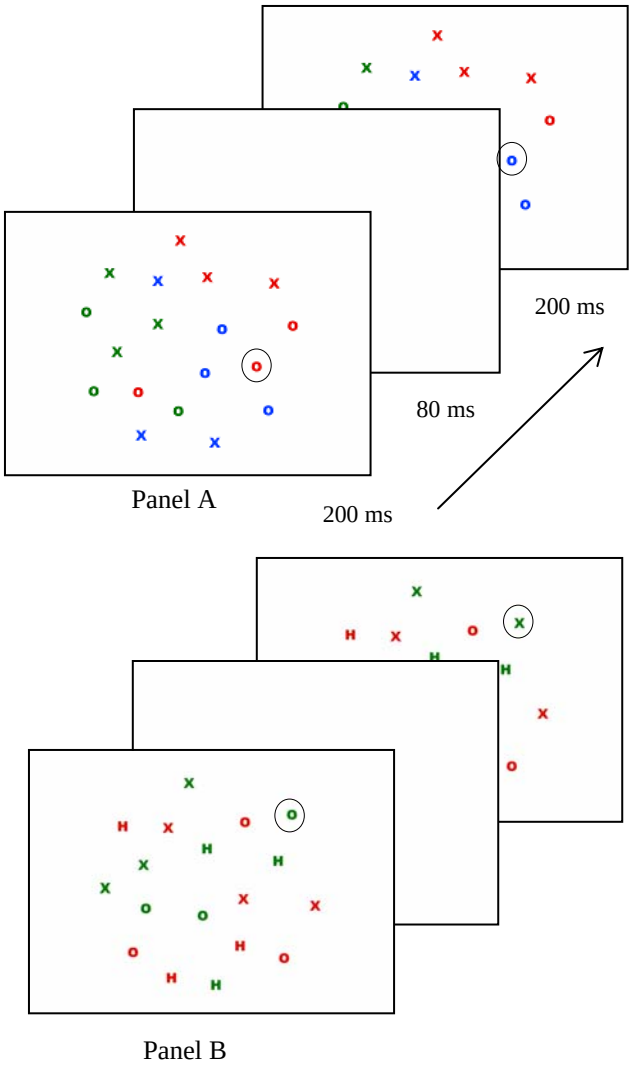


Figure 1: An illustration of the change blindness paradigm. Panel A shows a target changing color, and Panel B shows a target changing shape. The targets are circled for illustration purposes only.

Stimulus displays were presented on two monitors, one for the participant (a 17-inch Viewsonic 17PS) and one for the experimenter. The experimenter monitor was used to give feedback in real-time about the participant's computed gaze position. This feedback was given in the form of a cursor measuring 1° in diameter which was overlaid on the same image being viewed by the participant. This allowed the experimenter to evaluate system accuracy and to initiate a recalibration if necessary. In general, the average error in the computation of gaze position was less than 0.5° of visual angle.

Materials and Design

As shown in Figure 1, the change blindness flicker paradigm was used (Rensink et al., 1997). As can be seen, in each trial, the screen flickered between the

display and a blank screen (200ms display/80ms blank screen/200ms display). Displays were composed of 18 letters (1 target, 17 distractors) that were arrayed on three invisible concentric rings. On all trials there was an equal number of letters of each color and shape. The radii of the rings were 2.8, 5.6, and 8.3 degrees of visual angle (at a distance of 70 cm). The minimum distance between items was set at 3.5 degrees. Individual letters subtended 0.86 degrees both horizontally and vertically. Colors were matched in luminance and saturation (CIE coordinates: *x*, *y*, red: 0.578, 0.350, green: 0.327, 0.549, blue: 0.170, 0.117 and yellow: 0.446, 0.454). The location of the target and the configuration of the distractor items were randomized for every trial.

As the screen flickered, the participants' task was to detect the letter (target) that differed in color or shape. For trials where the target changed color, half of the displays contained an equal number of X's and O's (X/O displays), and half of the displays contained an equal number of C's and T's (C/T displays). In both cases, an equal number of red, green and blue letters was displayed, one of which was chosen at random to change color. For example, Figure 1, Panel A shows an X/O display where the target (which is circled for illustrative purposes) changes color from red to blue. In this trial the screen would alternate between the display with the red target letter, the blank display, and the display with the blue target letter, until the participant found the target. Alternatively, for trials where the target changed shape, half of the displays contained an equal number of green and red letters (green/red displays), and half of the displays contained an equal number of yellow and blue letters (yellow/blue displays). In both cases, an equal number of X's, O's, and H's was displayed, one of which was chosen at random to change shape. For example, Panel B shows a green/red display where the target changes shape from an O to an X. In this trial the screens would alternate between the given displays until the participant located the target changing shape.

In a subset of each of these types of trials, there was a covariation embedded in the task that reduced the possible target locations to half of the display items. Two covariation rules were used as illustrated in Table 1. In the Shape rule condition, the displays contained either thick letters or thin letters. Letter thickness predicted target shape in the Covariant trials, but not the Random trials. In the example in Table 1, X/O displays correspond to Covariant trials, in which thick letter displays predicted the target to be an X and thin letter displays predicted the target to be an O. In contrast, C/T displays correspond to the Random trials, in which letter thickness did not predict target shape.

In the Color rule condition, the displays contained either a light or a dark background. Background luminance predicted target color in the Covariant trials, but not the Random trials. In the example in Table 1,

Table 1: An illustration of the covariation rules for the Shape and Color conditions.

Rule Condition	Display type	Trials: Covariant	Trials: Random
Shape (Fig. 1 Panel A)	Thick Letter	X -> X	C -> C or T -> T
	Thin Letter	O -> O	C -> C or T -> T
Color (Fig. 1 Panel B)	Light Background	O -> X	O -> X or O -> X
	Dark Background	O -> X	O -> X or O -> X

green/red displays correspond to Covariant trials, in which a light background predicted the target to be green and a dark background predicted the target to be red. In contrast, yellow/blue displays correspond to the Random trials, in which background luminance did not predict target color.

Each participant performed 624 trials. The order of the stimulus displays was random under the restrictions that there be no more than 4 displays of a given display type or trial type in a row, and that each 48 trial block contained an equal number of trials of each display type and each trial type (Covariant, Random). Mapping of the display types to conditions was counterbalanced across participants.

Procedure

The experiment was run in a lighted room with a luminance of approximately 30 cd/m^2. Before beginning the task, the participants were informed that stimulus displays would consist of letters that would flash intermittently, and that their task was to search for and detect the single letter that was changing shape, for participants in the Color rule condition, or changing color, for participants in the Shape rule condition. Participants were asked to locate the target as quickly as possible and terminate the trial by maintaining fixation on the target. Participants then completed 48 practice trials, followed by 2 blocks of 288 experimental trials. Every 48 trials participants were given a rest break.

At the beginning of each trial, a fixation point was presented in the center of the computer screen in order to correct for drift in gaze position. A button press from the participant initiated the trial and it ended 700 ms after the participant fixated on the target (gaze-controlled response). The time between display onset and the participant's gaze-controlled response was recorded as the response time (RT). In addition, the participant's eye movements, monitored via the EyeLink tracking system, were recorded.

In order to compare explicit and implicit covariant rule learning, following completion of the practice block, half of the 24 participants in the each rule condition were told that there was a 100% covariation between the relevant features in one of the two display types (the Informed group). However, they were not told which display type contained the covariation. The other half of participants were not told anything about this relationship (the Uninformed group). At the end of the experiment a structured questionnaire was given to investigate awareness of the covariation.

Results

Reaction Time

For each participant and condition the mean RT was calculated, excluding those trials that were greater than 3 standard deviations from the mean. This resulted in less than 3% of trials being omitted. As Figure 2 reveals, RTs improved over the course of the experiment ($F\ (2, 88)$ = 101, p <.001) and were faster for Covariant than Random trials ($F\ (1, 44)$ = 31.6, p <.001). Importantly, participants improved more over time for Covariant trials than for Random trials, demonstrating a reaction time benefit for learning the covariation, $F\ (2, 88)$ = 3.87, p = .025. Moreover, this RT benefit of Covariant trials over Random trials was significant for both the Informed ($F\ (1, 22)$ = 96.9, p < .001) and Uninformed ($F\ (1, 22)$ = 9.40, p < .01) groups. There were no RT differences for the practice block across trial type ($t\ (1, 47)$ = .82, p > .4).

As can be seen upon further inspection of Figure 2, the instruction manipulation (Informed, Uninformed) affected the degree to which participants benefited from the covariation information. Specifically, there was a greater benefit for the Covariant trials over the Random trials for participants in the Informed group (mean difference = 1010 ms) as compared to participants in the Uninformed group (mean difference = 165 ms), $F\ (1, 44)$ = 53.1, p < .001. Further, Informed participants showed significantly greater RT advantages for Covariant trials than for Random trials in the Color rule condition (mean difference: 1520 ms, SE = 180) when compared to the Shape rule condition (mean difference: 500 ms, SE = 98.0), F (1, 22) = 24.7, p < .001). In contrast, the Uninformed participants showed no differential RT benefit across rule conditions (F (1, 22) < 1, mean difference: 155 ms, SE = 81.2, for the Color rule condition vs. mean difference: 175 ms, SE = 70.8, for the Shape rule condition). These differences manifested themselves as a trial type by instruction by rule condition interaction, $F\ (1, 44)$ = 20.1, p < .001.

Saccadic Selectivity

Saccadic selectivity was computed by assigning saccadic endpoints to the closest display item and calculating the percentage of saccades directed towards the 7 distractors out of the total of 17 distractors that shared shape with the target in the Shape rule condition or color with the target

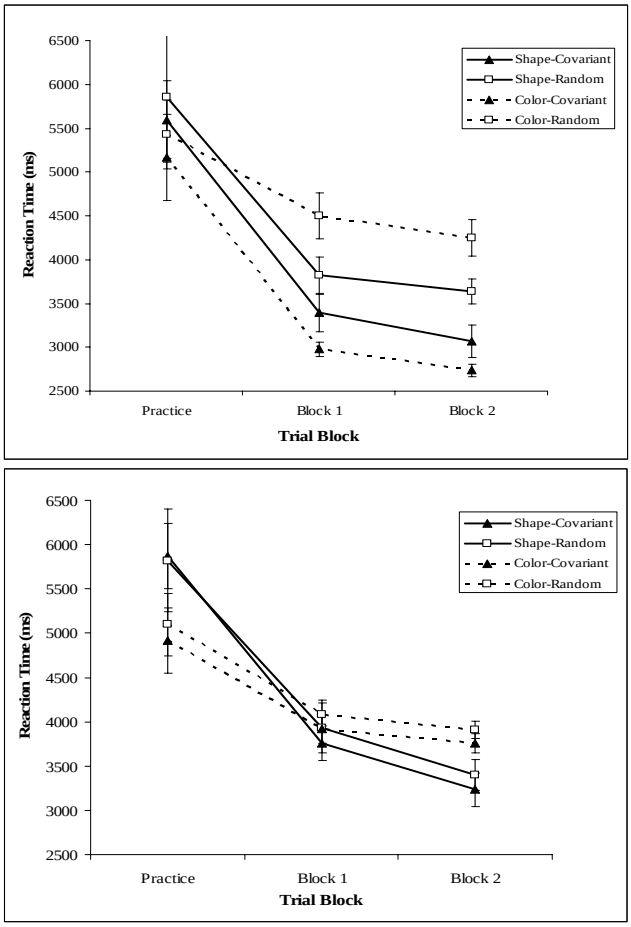


Figure 2: Reaction time by trial Block is shown for both the Covariant and Random trials by rule condition (Shape, Color). The top panel contains the results for the Informed group, and the bottom panel contains the results for the Uninformed group.

in the Color rule condition. Accordingly, chance performance on the saccadic selectivity measure (i.e., when all distractors have an equal probability of being fixated) was expected to be 47.06%. Consistent with this, saccadic selectivity was at chance (Mean = 47.06; all t's $(1, 47)$ < 1, p's > .5) for the Random trials over the course of the experiment. In contrast, as can be seen in Figure 3, while saccadic selectivity did not significantly differ from chance for Covariant trials for the practice block (Mean = 47.06; $t\ (1, 47)$ = 1.24, p > .2), participants' saccadic selectivity increased over time, demonstrating a behavioral consequence for learning of the covariation, $F\ (2, 88)$ = 81.3, p < .001. This bias towards fixating on the letters that had the same shape or color as the target occurred more strongly for the Informed participants ($F\ (1, 22)$ = 27.5, p < .001), than the Uninformed participants ($F\ (1, 22)$ = 10.3, p < .005). Further, Informed participants showed significantly greater saccadic selectivity in the Color rule condition (mean = 75.3%, SE = 1.57) as compared to the Shape rule condition (mean =

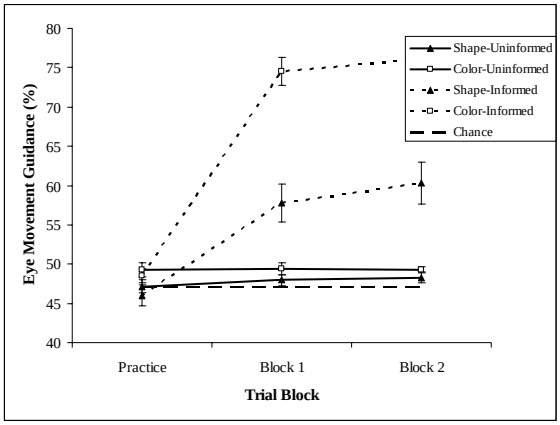


Figure 3: Percent eye-movement guidance by trial Block for the Covariant trials by group (Informed, Uninformed) and rule condition (Shape, Color). Chance guidance is shown by a dashed line at 47.08%. The practice block contains 48 trials, and Blocks 1 and 2 average 288 trials each.

59.1%, SE = 2.43), F (1, 22) = 31.5, p < .001). In contrast, the Uninformed participants showed no differential saccadic selectivity across rule conditions (F (1, 22) = 3.05, p = .095; mean = 49.3%, SE = .482 for the Color rule condition, mean = 48.2%, SE = .476 for the Shape rule condition). These differences manifested themselves as an instruction by rule condition interaction, $F\ (1, 44) = 25.71$, $p < .001$.

The effectiveness of the instructions was further evaluated by a retrospective report in the form of a structured questionnaire and intensive questioning. It was found that participants in the Uninformed group reported being unaware of the covariation, while those in the Informed group could explain the covariation, consistent with instructions. In fact, most Uninformed participants expressed disbelief when they were told of the covariation, reporting that they thought that letter thickness (for those in the Shape rule condition), or background luminance (for those in the Color rule condition), were irrelevant.

Discussion

In an attempt to make progress towards the goal of demonstrating qualitative differences between implicit and explicit processing, the present study employed eye movement monitoring as an index of learning, in addition to reaction time, to further explore implicit learning in the context of visual search. It was found that participants can enhance visual search efficiency by utilizing a covariation that provided target identity information, and that such learning occurs even when participants are uninformed and claim to be unaware of this information. Implicit learning was most strongly demonstrated by the response time measure. Specifically, for the Uninformed participants, while the RT measure showed savings for trials containing the covariant rule condition (165ms in both the Shape and Color rule conditions), only a slight saccadic selectivity (1% bias in the Shape rule condition and 2% bias in the Color rule condition) favoring relevant distractors was demonstrated. In contrast, participants who were informed of the covariation rule showed substantial learning of the covariation in the form of a reaction time advantage (500ms in the Shape rule condition and 1520ms in the Color rule condition) and saccadic selectively (12% bias in the Shape rule condition and 28% bias in the Color rule condition). Taken together, saccadic selectivity seems to be one mechanism by which Informed participants made their search more efficient, whereas the weak selectivity displayed by the Uninformed participants suggests that this was not the primary mechanism employed by this group. Future investigation into what other means the Uninformed participants were using to improve search efficiency is warranted. The different picture portrayed by the RT and saccadic selectivity results points to the importance of using multiple concurrent measures of performance in attempting to more conclusively document implicit learning and in attempting to better understand the mechanisms underlying variation in performance associated with the presence or absence of claimed awareness pertaining to the learned covariation.

The present finding of implicit learning in visual search is consistent with previous evidence using other visual search paradigms (e.g., Chun & Jiang, 1999; Chun & Nakayama, 2000; Durgin, 1999; Flowers & Smith, 1998; Miller, 1991). A benefit to our use of the change blindness paradigm was that it allowed us to create a methodology whereby no target-absent trials were needed and where the covariant rule involved a dimension that was seemingly irrelevant to the search task.

Finally, the rule condition (Color, Shape) differentially influenced performance in the Informed and Uninformed groups. Specifically, the Informed participants in the Color rule condition were able to use the covariation information to a greater extent than those in the Shape rule condition, as shown by both the RT and saccadic selectivity results (see Figure 2 and 3, respectively). In contrast, no such differential sensitivity across rule condition was found for the Uninformed participants for either of these measures. We believe that the demonstration of differential search performance across groups and conditions in the present study constitutes progress toward the crucial goal of establishing qualitative differences between implicit and explicit learning in the context of visual search.

Acknowledgments

This research was supported by an NSERC research grant to E.M.R. and an NSERC postgraduate scholarship to C.A.R.

References

Cheesman, J., & Merikle, P. M. (1986). Distinguishing conscious from unconscious perceptual processes. *Canadian Journal of Psychology, 40,* 343-367.

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science, 10*, 360-365.

Chun, M. M., & Nakayama, K. (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition, 7*, 65-81.

Dienes, Z., & Berry, D. (1997.) Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review, 4*, 3-23.

Dixon, N. F. (1981). *Preconscious processing.* Chichester: Wiley.

Durgin, F. H. (1999). Supporting the "Grand Illusion" of direct perception: Implicit learning in eye-movement control. In S. R. Hameroff, A. W Kaszniak and D. J. Chalmers (Eds.), *Toward a Science of Consciousness III: The Third Tucson Discussions and Debates* (pp. 179-188). Massachusetts Institute of Technology, U.S.A.

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research, 37*, 617-631.

Flowers, J. H., & Smith, K. L. (1998). What is learned about nontarget items in simple visual search? *Perception & Psychophysics, 60*, 696-704.

Hooge, I. T., & Erkelens, C. J. (1999). Peripheral vision and oculomotor control during visual search. *Vision Research, 39*, 1567-1575.

Miller, J. (1991). The flanker compatibility effect as a function of visual angle, attentional focus, visual transients, and perceptual load: A search for boundary conditions. *Perception & Psychophysics, 49*, 270-288.

Neal, A., & Hesketh, B. (1997). Episodic knowledge and implicit learning. *Psychonomic Bulletin & Review, 4*, 24-37.

Pomplun, M., Reingold, E. M., & Shen, J. (2001). The effects of peripheral and parafoveal cueing and masking on saccadic selectivity in a gaze-contingent window paradigm. *Vision Research, 41*, 2757-2769.

Pomplun, M., Reingold, E. M., & Shen, J. (2003). Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science, 27*, 299-312.

Pomplun, M., Reingold, E. M., Shen, J., & Williams, D. E. (2000). The area activation model of saccadic selectivity in visual search. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd annual conference of the cognitive science society* (pp. 375-380). Mahwah, NJ: Elrbaum.

Reingold, E. M., & Merikle, P. M. (1990). On the inter-relatedness of theory and measurement in the study of unconscious processes. *Mind & Language, 5*, 9-28.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368-373.

Scialfa, C. T., & Joffe, K. (1998). Response times and eye movements in feature and conjunction search as a function of target eccentricity. *Perception & Psychophysics, 60*, 1067-1082.

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367-447.

Shevrin, H., & Dickman, S. (1980). The psychological unconscious: A necessary assumption for all psychological theory? *American Psychologist, 35*, 421-434.

Shen, J., & Reingold, E. M. (1999). Saccadic selectivity during visual search: The effects of shape and stimulus familiarity. In M. Hahn and S. C. Stoness (Eds.), *Proceedings of the 21st annual conference of the cognitive science society* (pp. 649-652). Mahwah, NJ: Erlbaum.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. *Perception, 29*, 241-250.

Shen, J., Reingold, E. M., & Pomplun, M. (2003). Guidance of eye movements during conjunctive visual search: The distractor-ratio effect. *Canadian Journal of Experimental Psychology, 57*, 76-96.

Shen, J., Reingold, E. M., Pomplun, M., & Williams, D. E. (2003). Saccadic selectivity during visual search: The influence of central processing difficulty. In J. Hyönä, R. Radach & H. Deubel (Eds.), *The Mind's eyes: Cognitive and applied aspects of eye movement research* (pp. 65-88). Amsterdam: Elsevier.

Stadler, M. A., & Frensch, P. A. (1994). Whither learning whither memory? *Behavioural and Brain Sciences, 17*, 423-424.

Williams, D. E., & Reingold, (2001). Preattentive guidance of eye movements during triple conjunction search tasks: The effects of feature discriminability and saccadic amplitude. *Psychonomic Bulletin & Review, 8*, 476-488.

Structure Dependence in Language Acquisition: Uncovering the Statistical Richness of the Stimulus

Florencia Reali (fr34@cornell.edu) and Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology; Cornell University; Ithaca, NY 14853 USA

Abstract

The *poverty of stimulus argument* is one of the most controversial arguments in the study of language acquisition. Here we follow previous approaches challenging the assumption of impoverished primary linguistic data, focusing on the specific problem of auxiliary fronting in polar interrogatives. We develop a series of child-directed corpus analyses showing that there is indirect statistical information useful for correct auxiliary fronting in polar interrogatives, and that such information is sufficient for producing grammatical generalizations even in the absence of direct evidence. We further show that there are simple learning devices, such as neural networks, capable of exploiting such statistical cues, producing a bias to correct *aux*-questions when compared to their ungrammatical counterparts. The results suggest that the basic assumptions of the poverty of stimulus argument need to be reappraised.

Introduction

How do children learn aspects of their language for which there appears to be no evidence in the input? This question lies at the heart of the most enduring and controversial debates in cognitive science. Ever since Chomsky (1965), it has been argued that the information in the linguistic environment is too impoverished for a human learner to attain adult competence in language without the aide of innate linguistic knowledge. Although this *poverty of the stimulus argument* (Chomsky, 1980; Crain & Pietroski, 2001) has guided most research in linguistics, it has proved to be much more contentious within the broader context of cognitive science.

The poverty of stimulus argument rests on certain assumptions about the nature of the input to the child, the properties of computational learning mechanisms, and the learning abilities of young infants. A growing bulk of research in cognitive science has begun to call each of these three assumptions into question. Thus, whereas the traditional nativist perspective suggests that statistical information may be of little use for syntax acquisition (e.g., Chomsky, 1957), recent research indicates that distributional regularities may provide an important source of information for syntactic bootstrapping (e.g., Mintz, 2002; Redington, Chater and Finch, 1998)—especially when integrated with prosodic or phonological information (e.g., Christiansen & Dale, 2001; Morgan, Meier & Newport, 1987). And while the traditional approach only tends to consider learning in highly simplified forms, such as "move the first occurrence of *X* to *Y*", progress in statistical natural language processing and connectionist modeling has revealed much more complex learning abilities of potential relevance for language acquisition (e.g., Lewis & Elman, 2001). Finally, little attention has traditionally been paid to what young infants may be able to learn, and this may be problematic given that recent research has demonstrated that even before one year of age, infants are quite competent statistical learners (Saffran, Aslin & Newport, 1996—for reviews, see Gómez & Gerken, 2000; Saffran, 2003).

These research developments suggest the need for a reappraisal of the poverty of stimulus argument, centered on whether they together can answer the question of how a child may be able to learn aspects of linguistic structure for which innate knowledge was previously thought to be necessary. In this paper, we approach this question in the context of structure dependence in language acquisition, specifically in relation to auxiliary fronting in polar interrogatives. We first outline the poverty of stimulus debate as it has played out with respect to forming grammatical questions with auxiliary fronting. It has been argued that the input to the child does not provide enough information to differentiate between correct and incorrect auxiliary fronting in polar interrogatives (Chomsky in Piatelli-Palmarini, 1980). In contrast, we conduct a corpus analysis to show that there is sufficiently rich statistical information available in child-directed speech for generating correct *aux*-questions—even in the absence of any such constructions in the corpus. We additionally demonstrate how the same approach can be applied to explain results from studies of auxiliary fronting in 3- to 5-year-olds (Crain & Nakayama, 1987). Whereas, the corpus analyses indicate that there is rich statistical information available in the input, it does not show that there are learning mechanisms capable of utilizing such information. We therefore conduct a set of connectionist simulations to illustrate that neural networks are capable of using statistical information to distinguish between correct and incorrect *aux*-questions. In the conclusion, we discuss our results in the context of recent infant learning results.

The Poverty of Stimulus and Structure Dependence in Auxiliary Fronting.

Children only hear a finite number of sentences, yet they learn to speak and comprehend sentences drawn from a language that can contain an infinite number of sentences. The poverty of stimulus argument suggests that children do

not have enough data during the early stages of their life to learn the syntactic structure of their language. Thus, learning a language involves the correct generalization of grammatical structure when insufficient data is available to children. The possible weakness of the argument lies in the difficulty to assess the input, and in the imprecise and intuitive definition of 'insufficient data'.

One of the most used examples to support the poverty of stimulus argument concerns auxiliary fronting in polar interrogatives. Declaratives are turned into questions by fronting the correct auxiliary. Thus, for example, in the declarative form '*The man who is hungry is ordering dinner'* it is correct to front the main clause auxiliary as in 1, but fronting the subordinate clause auxiliary produces an ungrammatical sentence as in 2 (Chomsky, 1965).

1. *Is the man who is hungry ordering dinner?*
2. **Is the man who hungry is ordering dinner?*

Children can generate two types of rules: a structure-independent rule where the first *'is'* is moved; or the correct structure-dependent rule, where only the movement of the *'is'* from the main clause is allowed. Crucially, children do not appear to go through a period when they erroneously move the first *is* to the front of the sentence (e.g., Crain & Nakayama, 1987). It has moreover been asserted that a person might go through much of his or her life without ever having been exposed to the relevant evidence for inferring correct auxiliary fronting (Chomsky, in Piatelli-Palmarini, 1980).

The purported absence of evidence in the primary linguistic input regarding auxiliary fronting in polar interrogatives is not without debate. Intuitively, as suggested by Lewis & Elman (2001), it is perhaps unlikely that a child would reach kindergarten without being exposed to sentences such as 3-5.

3. *Is the boy who was playing with you still there?*
4. *Will those who are hungry raise their hand?*
5. *Where is the little girl full of smiles?*

These examples have an auxiliary verb within the subject NP, and thus the auxiliary that appears initially would not be the first auxiliary in the declarative, providing evidence for correct auxiliary fronting. Pullum & Scholz (2002) explored the presence of auxiliary fronting in polar interrogatives in the Wall Street Journal (WSJ). They found that at least five crucial examples occur in the first 500 interrogatives. These results suggest that the assumption of complete absence of evidence for correct auxiliary fronting is overstated. Nevertheless, it has been argued that the WSJ corpus is not a good approximation of the grammatical constructions that young children encounter and thus it cannot be considered representative of the primary linguistic data. Indeed, studies of the CHILDES corpus show that even though interrogatives constitute a large percentage of the corpus, relevant examples of auxiliary fronting in polar interrogatives represent less than 1% of them (Legate & Yang, 2002).

Although the direct evidence for auxiliary fronting in polar interrogatives may be too weak to be helpful in acquisition—as suggested by Legate & Yang (2002)—other more indirect sources of statistical information may provide sufficient basis for making the appropriate grammatical generalizations. Recent connectionist simulations provide preliminary data in this regard. Lewis & Elman (2001) trained simple recurrent networks (SRN; Elman, 1990) on data from an artificial grammar that generated questions of the form 'AUX NP ADJ?' and sequences of the form 'A_i NP B_i' (where A_i and B_i represent a variety of different material) but no relevant examples of polar interrogatives. The SRNs were better at making predictions for correct auxiliary fronting compared to those with incorrect auxiliary fronting. This indicates that even without direct exposure to relevant examples, the statistical structure of the input nonetheless provides useful information applicable to auxiliary fronting in polar interrogatives.

However, the SRNs in the Lewis & Elman simulation studies were exposed to an artificial grammar without the complexity and noisiness that characterizes actual child-directed speech. The question thus remains whether the indirect statistical regularities in an actual corpus of child-directed speech are strong enough to support grammatical generalizations over incorrect ones—even in the absence of direct examples of auxiliary fronting in polar interrogatives in the input. Next, in our first experiment, we conduct a corpus analysis to demonstrate that the indirect statistical information available in a corpus of child-directed speech is indeed sufficient for making the appropriate grammatical generalizations in questions involving auxiliary fronting.

Experiment 1: Measuring Indirect Statistical Information Relevant for Auxiliary Fronting

Even if children only hear a few relevant examples of polar interrogatives, they may nevertheless be able to rely on indirect statistical cues for learning the correct structure. In order to assess this hypothesis, we trained bigram and trigram models on the Bernstein-Ratner (1984) corpus of child-directed speech and then tested the likelihood of novel example sentences. The test sentences consisted of correct polar interrogatives (e.g. *Is the man who is hungry ordering dinner?*) and incorrect ones (e.g. *Is the man who hungry is ordering dinner?*)—neither of which were present in the training corpus. We reasoned that if indirect statistical information provides a possible cue for generalizing correctly to the grammatical *aux*-questions, then we should find a difference in the likelihood of these two alternative hypotheses.

Bigram/trigram models are simple statistical models that use the previous one/two word(s) to predict the next one. Given a string of words or a sentence it is possible to compute the associated cross-entropy for that string of words according to the bigram/trigram model trained on a particular corpus (from Chen & Goodman, 1996). Thus, given two alternative sentences we can compare the probability of each of them as indicated by their associated cross-entropy as computed in the context of a particular corpus. Specifically, we can compare the two alternative

generalizations of doing auxiliary fronting in polar interrogatives, comparing the cross-entropy associated with grammatical (e.g., *Is the man who is in the corner smoking?*) and ungrammatical forms (e.g., *Is the man who in the corner is smoking*). This will allow us to determine whether there may be sufficient indirect statistical information available in actual child-directed speech to decide between these two forms. Importantly, the Bernstein-Ratner corpus contains no examples of auxiliary fronting in polar interrogatives. Our hypothesis is therefore that the corpus nonetheless contains enough statistical information to decide between grammatical and ungrammatical forms.

Method

Models For the purpose of corpus analysis we used bigram and trigram models of language (see e.g., Jurafsky & Martin, 2000). The probability P(*s*) of a sentence was expressed as the product of the probabilities of the words (w_i) that compose the sentence, with each word probability conditional to the last *n*-1 words. Then, if s = $w_1 \ldots w_k$ we have:

$$P(s) = \Pi_i P(w_i|w^{i-1}_{i-n+1})$$

To estimate the probabilities of $P(w_i|w_{i-1})$ we used the *maximum likelihood* (ML) estimate for $P(w_i|w_{i-1})$ defined as (considering the bigram model):

$$P_{ML}(w_i|w_{i-1}) = P(w_{i-1}w_i) / P(w_{i-1}) = (c(w_{i-1}w_i)/Ns)/(c(w_{i-1})/Ns);$$

where Ns denote the total number of tokens and $c(\alpha)$ is the number of times the string α occurs in the corpus. Given that the corpus is quite small, we used the *interpolation smoothing technique* defined in Chen & Goodman (1996). The probability of a word (w_i) (or unigram model) is defined as:

$$P_{ML}(w_i) = c(w_i)/Ns;$$

The smoothing technique consists of the interpolation of the bigram model with the unigram model, and the trigram model with the bigram model. Thus, for the bigram model we have:

$$P_{interp}(w_i|w_{i-1}) = \lambda P_{ML}(w_i|w_{i-1}) + (1-\lambda)P_{ML}(w_i)$$

Accordingly for trigram models we have:

$$P_{interp}(w_i|w_{i-1}w_{i-2}) = \lambda P_{ML}(w_i|w_{i-1}w_{i-2}) + (1-\lambda)(\lambda P_{ML}(w_i|w_{i-1}) + (1-\lambda)P_{ML}(w_i)),$$

where λ is a value between 0 and 1 that determines the relative importance of each term in the equation. We used a standard $\lambda = 0.5$ so that all terms are equally weighted. We measure the likelihood of a given set of sentences using the measure of cross-entropy (Chen & Goodman, 1996). The cross-entropy of a set of sentences is defined as:

$$1/N_T \Sigma_i -\log_2 P(s_i) \quad \text{(where } s_i \text{ is the i}^{th} \text{ sentence).}$$

The cross-entropy value of a sentence is inversely correlated with the likelihood of it. Given a training corpus, and two sentences A and B we can compare the cross-entropy of both sentences and estimate which one is more probable according to the statistical information of the corpus. We used Perl programming in a Unix environment to implement the corpus analysis. This includes the simulation of bigram and trigram models and cross-entropy calculation and comparisons.

Materials We used the Bernstein-Ratner (1984) corpus of child-directed speech for our corpus analysis. It contains recorded speech from nine mothers speaking to their children over 4-5 months period when children were between the ages of 1 year and 1 month to 1 year and 9 months. This is a relatively small and very noisy corpus, mostly containing short sentences with simple grammatical structure. The following are some example sentences: *Oh you need some space; Where is my apple?; Oh. That's it*'.

Procedure We used the Bernstein-Ratner child-directed speech corpus as the training corpus for the bigram/trigram models. The models were trained on 10,082 sentences from the corpus (34,010 word tokens; 1,740 word types). We wanted to compare the cross-entropy of grammatical and ungrammatical polar interrogatives. For that purpose, we created two novel sets of sentences. The first one contained grammatically correct polar interrogatives and the second one contained the ungrammatical version of each sentence in the first set. The sentences were created using a random algorithm that selected words from the corpus, and created sentences according to syntactic and semantic constraints. We tried to prevent any possible bias in creating the test sentences. The test sets only contained relevant examples of polar interrogatives of the form: "*Is* / NP/ (who/that)/ *is* / A_i/ B_i?", where A_i and B_i represent a variety of different material including VP, PARTICIPLE,NP, PP, ADJP (e.g.: *"Is the lady who is there eating?*"; "*Is the dog that is on the chair black?*"). Each test set contained 100 sentences. We estimated the mean cross-entropy per sentence by calculating the average cross-entropy of the 100 sentences in each set. Then we compared the likelihood of pairs of grammatical and ungrammatical sentences by comparing their cross-entropy and choosing the version with the lower value. We studied the statistical significance of the results using paired t-test analyses.

Results

We found that the mean cross-entropy of grammatical sentences was lower than mean cross entropy of ungrammatical sentences. We performed a statistical analysis of the cross-entropy difference, considering all pairs of grammatical and ungrammatical sentences. The cross-entropy difference was highly significant (t(99), $p<0.0001$) (see Table 1). These results show that grammatical sentences have a higher probability than ungrammatical ones. In order to compare each grammatical-ungrammatical pair of sentences, we defined the following criterion: When deciding between each grammatical vs. ungrammatical polar interrogative example, choose the one that has lower cross-entropy (the most probable one).

Table 1: Comparison of mean cross-entropy in Exp.1.

	Mean cross-entropy		Mean difference	t(99) p-value
	Gram.	Ungramm.		
Bigram	22.92	23.73	0.83	< 0.0001
Trigram	21.81	23.07	1.26	< 0.0001

A sentence is defined as correctly classified if the chosen form is grammatical. Using that criterion, we found that the percentage of correctly classified sentences using the bigram model is 92% and using the trigram model is 95%. Figure 1 shows the performance of the models according to the defined classification criterion. Of the 100 test sentences, the trigram model only misclassified the following five: *Is the lady who is here drinking?; Is the alligator that is standing there red?; Is the jacket that is on the chair lovely?; Is the one that is in the kitchen scared?; Is the phone that is in the office purple?*

The bigram model in addition to the above five sentences also misclassified the next three test sentences: *Is the bunny that is in the car little; Is the baby who is in the castle eating?; Is the bunny that is sleeping black?*

It is possible to calculate the probability of a sentence from the cross-entropy value. Figure 2 shows the comparison of mean probability of grammatical and ungrammatical sentences. We found that the mean probability of grammatical polar interrogatives is almost twice the mean probability of ungrammatical polar interrogatives according to the bigram model and it is more than twice according to the trigram model.

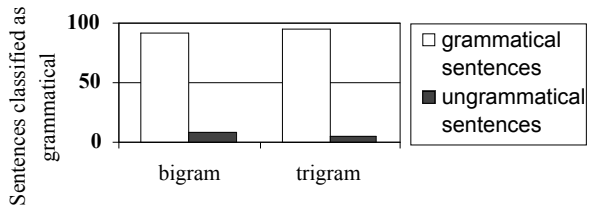


Figure 1: Number of sentences classified correctly (white bars) and incorrectly as grammatical (gray bars)

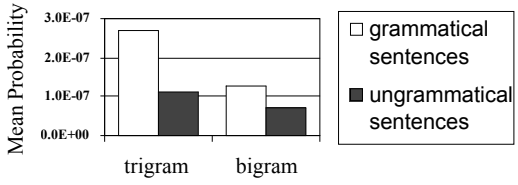


Figure 2: Mean probability of grammatical sentences vs. mean probability of ungrammatical sentences.

Experiment 2: Testing Sentences with Auxiliary Fronting Produced by Children

Although Experiment 1 shows that there is sufficient indirect statistical information available in child-directed speech to differentiate reliably between the grammatical and ungrammatical *aux*-questions that we had generated, it could be argued that the real test for our approach is whether it works for actual sentences produced by children. We therefore tested our models on a small set of sentences elicited from children under experimental conditions.

Crain & Nakayama (1987) conducted an experiment designed to elicit complex *aux*-questions from 3- to 5-year-old children. The children were involved in a game in which they asked questions to Jabba the Hutt, a creature from Star Wars. During the task the experimenter gives an instruction to the child: *'Ask Jabba if the boy who is watching Mickey Mouse is happy'*. Children produced sentences like a) *'Is the boy who is watching Mickey Mouse happy?'* but they never produced sentences like b) *'Is the boy who watching Mickey Mouse is happy?'*. The authors concluded that the lack of structure-independent errors suggested that children entertain only structure-dependent hypotheses, supporting the existence of innate grammatical structure.

Method

Models Same as in Experiment 1.

Materials Six example pairs were derived from the declarative sentences used in Crain & Nakayama[1](1987):

6. *The ball that the girl is sitting on is big*
7. *The boy who is unhappy is watching Mickey Mouse*
8. *The boy who is watching Mickey Mouse is happy*
9. *The boy who is being kissed by his mother is happy*
10. *The boy who was holding the plate is crying*
11. *The dog that is sleeping is on the blue bench*

The grammatical and ungrammatical *aux*-questions were derived from the declaratives in 6-11. Thus, the sentence '*Is the dog that is sleeping on the blue bench?*' belonged to the grammatical test set whereas the sentence '*Is the dog that sleeping is on the blue bench?*' belonged to the ungrammatical test set. Consequently, grammatical and ungrammatical test sets contained 6 sentences each.

Procedure The bigram/trigram models were trained on the Bernstein-Ratner (1984) corpus as in Experiment 1, and tested on the material derived from Crain & Nakayama (1987).

Results

Consistently with Experiment 1, we found that the mean cross-entropy of grammatical sentences was significantly lower than the mean cross entropy of ungrammatical sentences both for bigram and trigram models ($t(5)$ $p<0.013$ and $p<0.034$ respectively). Table 2 summarizes these results.

[1] As some of the words in the examples were not present in the Bernstein-Ratner corpus, we substitute them for semantically related ones: Thus, the words: "mother", "plate", "watching", "unhappy" and "bench" were replaced respectively by "mommy", "ball", "looking at", "crying" and "chair".

Using the classification criterion defined in Experiment 1, we found that all six sentences were correctly classified using the bigram model. That is, according to the distributional information of the corpus, all grammatical *aux*-questions were more probable than the ungrammatical version of them. When using the trigram model, we found that five out of six sentences were correctly classified.

Table 2: Comparison of mean cross-entropy in Exp.2.

	Mean cross-entropy		Mean difference	t(5) p-value
	Gram.	Ungramm.		
Bigram	26.99	27.89	0.90	< 0.013
Trigram	25.97	26.86	0.89	< 0.034

Experiment 3: Learning to Produce Correct Sentences with Auxiliary Fronting

While Experiments 1 and 2 establish that there is sufficient indirect statistical information in the input to the child to differentiate between grammatical and ungrammatical questions involving auxiliary fronting—including questions produced by children—it is not clear whether a simple learning device may be able to exploit such information to develop an appropriate bias toward the grammatical forms. To investigate this question, we took a previously developed SRN model of language acquisition (Reali, Christiansen & Monaghan, 2003), which had also been trained on the same corpus, and tested its ability to deal with *aux*-questions.

Previous simulations by Lewis & Elman (2001) have shown that SRNs trained on data from an artificial grammar were better at predicting the correct auxiliary fronting in *aux*-questions. An important question is whether the results shown using artificial-language models are still obtained when dealing with the full complexity and the general disorderliness of speech directed at young children. Thus, we seek to determine whether a previously developed connectionist model, trained on the same corpus, is sensitive to the same indirect statistical information that we have found to be useful in bigram/trigram models. SRNs are simple learning devices that have been shown to be sensitive to bigram/trigram information.

Method

Networks We used the same ten SRNs that Reali, Christiansen & Monaghan (2003) had trained to predict the next lexical category given the current one. These networks had initial weight randomization in the interval [-0.1; 0.1]. A different random seed was used for each simulation. Learning rate was set to 0.1, and momentum to 0.7. Each input to the network contained a localist representation of the lexical category of the incoming word. With a total of 14 different lexical categories and a pause marking boundaries between utterances, the network had 15 input units. The network was trained to predict the lexical category of the next word, and thus the number of output units was 15. Each network had 30 hidden units and 30 context units. All networks were simulated using the Lens simulator in a Unix environment. No changes were made to the original networks and their parameters.

Materials We trained and tested the networks on the Bernstein-Ratner corpus similarly to the bigram/trigram models. Each word in the corpus corresponded to one of the 14 following lexical categories from CELEX database (Baayen, Pipenbrock & Gulikers, 1995): nouns, verbs, adjectives, numerals, infinitive markers, adverbs, articles, pronouns, prepositions, conjunctions, interjections, complex contractions, abbreviations, and proper names. Each word in the corpus was replaced by a vector encoding the lexical category to which it belonged. We used the two sets of test sentences used in Experiment 1, containing grammatical and ungrammatical polar interrogatives respectively. However, as the network was trained to predict lexical classes, some test sentences defined in Experiment 1 mapped onto the same string of lexical classes. For simplicity, we only considered unique strings, resulting in 30 sentences in each test set (grammatical and ungrammatical).

Procedure The ten SRNs from Reali, Christiansen & Monaghan (2003) were trained on one pass through the Bernstein-Ratner corpus. These networks were then tested on the *aux*-questions described above. To compare network predictions for the ungrammatical vs. the grammatical *aux*-questions, we measured the networks' mean squared error recorded during the presentation of each test sentence pair.

Results

We found that in all ten simulations the grammatical set of aux-questions produced a lower error compared to the ungrammatical ones. The mean squared-error per next lexical class prediction was 0.80 for the grammatical set and 0.83 in the ungrammatical one, this difference being highly significant ($t(29)$ $p < 0.005$). Out of the 30 test sentences, 27 grammatical sentences produced a lower error than its ungrammatical counterpart. On the assumption that sentences with the lower error will be preferred, SRNs would pick the grammatical sentences in 27 out of 30 cases.

It is worth highlighting that the grammatical and ungrammatical sets of sentences were almost identical, only differing on the position of the fronted "is" as described in Experiment 1. Thus, the difference in mean squared error is uniquely due to the words' position in the sentence. Despite the complexity of child-directed speech, these results suggest that simple learning devices such as SRNs are able to pick up on the existing distributional properties showed in Experiment 1. Moreover, differently to Experiment 1, here we explored the distributional information of the lexical classes alone and thus the network was blind to the possible information present in word-word co-occurrences.

Conclusion

In the corpus analyses, we showed that there is sufficiently rich statistical information available *indirectly* in child-

directed speech for generating correct complex *aux*-questions—even in the absence of any such constructions in the corpus. We additionally demonstrated how the same approach can be applied to explain results from child-acquisition studies (Crain & Nakayama, 1987). These results indicate that indirect statistical information provides a possible cue for generalizing correctly to grammatical auxiliary fronting.

Whereas the corpus analyses indicate that there are statistical cues available in the input, it does not show that there are learning mechanisms capable of utilizing such information. However, previous results suggest that children are sensitive to the same kind of statistical evidence that we found in the present study. Saffran, Aslin & Newport (1996) demonstrated that 8 month-old children are particularly sensitive to transitional probabilities (similar to our bigram model). Sensitivity to transitional probabilities seems to be present across modalities, for instance in the segmentation of streams of tones (Saffran, Johnson, Aslin, & Newport, 1999). These and other results on infant statistical learning (see Gómez & Gerken, 2000) suggest that children have mechanisms for relying on implicit statistical information. SRNs are simple learning devices whose learning properties have been shown to be consistent with humans' learning abilities. Even though it was originally developed in a different context (Reali, Christiansen & Monaghan, 2003), our SRN model proved to be sensitive to the indirect statistical evidence present in the corpus, developing an appropriate bias toward the correct forms of *aux*-questions.

In conclusion, this study indicates that the poverty of stimulus argument may not apply to the classic case of auxiliary fronting in polar interrogatives, previously a corner stone in the argument for the innateness of grammar. Our results further suggest that the general assumptions of the poverty of stimulus argument may need to be reappraised in the light of the statistical richness of the language input to children.

Acknowledgments

This research was supported in part by a Human Frontiers Science Program Grant (RGP0177/2001-B).

References

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium. Univ. of Pennsylvania, Philadelphia, PA.

Bernstein-Ratner, N. (1984). Patterns of vowel Modification in motherese. *Journal of Child Language.* 11: 557-578.

Chen S.F. & Goodman J. (1996). An Empirical Study of Smoothing Techniques for Language Modeling. *Proceedings of the 34th Annual Meeting of ACL*.

Chomsky N. (1957). *Syntactic Structures*. Mouton and co.: The Hague.

Chomsky N. (1965). *Aspects of the Theory of Syntax*. Boston, MA: MIT Press.

Chomsky N. (1980). *Rules & Representation*. Cambridge, MA: MIT Press.

Christiansen, M.H. & Dale, R.A.C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 220-225). Mahwah, NJ: Lawrence Erlbaum.

Crain, S. & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*. 63: 522-543.

Crain, S. & Pietroski, P. (2001). Nature, nurture and Universal Grammar. *Linguistics and Philosophy*. 24: 139-186.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*. 14: 179-211.

Gómez, R.L., & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*. 4: 178-186.

Jurafsky, D. and Martin, J.H. (2000) *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.

Legate, J.A. & Yang, C. (2002) Empirical re-assessment of stimulus poverty arguments. *Linguistic Review*. 19: 151-162.

Lewis, J.D. & Elman, J.L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 359-370). Somerville, MA: Cascadilla Press.

Mintz, T.H. (2002) Category induction from distributional cues in an artificial language. *Memory & Cognition*. 30: 678-686.

Morgan, J. L., Meier, R.P. & Newport, E.L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*. 19: 498–550.

Piatelli-Palmarini, M. (ed.), 1980. *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Pullum G.K. & Scholz B. (2002) Empirical assessment of stimulus poverty arguments. *Linguistic Review*. 19: 9-50.

Reali, F., Christiansen, M.H. & Monaghan, P. (2003). Phonological and Distributional Cues in Syntax Acquisition: Scaling up the Connectionist Approach to Multiple-Cue In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 970-975). Mahwah, NJ: Lawrence Erlbaum.

Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science*. 22: 425-469.

Saffran, J.R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*. 12: 110-114.

Saffran, J.R., Aslin, R. & Newport, E.L. (1996) Statistical learning by 8- month-old infants. *Science*. 274: 1926-1928.

Saffran, J.R., Johnson, E.K., Aslin, R. & Newport, E.L. (1999) Statistical learning of tone sequences by human infants and adults. *Cognition*. 70: 27-52.

Modeling Complex Tasks: An Individual Difference Approach

John Rehling (rehling@andrew.cmu.edu)
Marsha Lovett (lovett@cmu.edu)
Christian Lebiere (cl@cmu.edu)
Lynne Reder (reder@cmu.edu)
Baris Demiral (baris@andrew.cmu.edu)
Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

It is the usual case in cognitive modeling that a model's output is compared to the average of a number of subjects, in which case the enterprise of modeling is apparently to capture the behavior of the typical individual. Our approach is to administer two simple tasks to each subject, using performance on those tasks as measures of individual ability. Those measures are then used as the values for parameters in an ACT-R model of a more complex task, so that the model can predict individual performance on that task.

Introduction

Work in cognitive modeling, when it seeks validation in the performance of human subjects, is almost unanimously concerned with the average performance of many subjects. For many purposes, however, it is desirable to be able to model or predict *individual* performance. We present here the first work to use a fine-grained cognitive model to predict individual performance in a complex task.

The ACT-R architecture, the basis of a great deal of work in cognitive modeling, has a detailed, well-developed theory of cognition – perception, learning, performance, and so on (Anderson and Lebiere, 1998). The architecture by necessity contains a number of parameters that can be used to fix levels of performance in, e.g., memory, to realistic levels. The ACT-R community has by custom sought universal values for these parameters wherever possible, finding values work across tasks, optimizing how well the model fits the data of the average subject. These parameters are each *meaningful*, each parameter determining the model's behavior in one specific way. For example, there is a parameter called W that determines the sum of the activations of all the pieces of information that may be retrieved at any point in time. It therefore controls the model's working memory capacity. Extensive empirical work in ACT-R modeling (again, of the kind where the model was meant to predict the average subject) found that a W value of 1.0 produces very good fits with subject data.

It was later postulated, however, that the W parameter could be meaningfully varied in order to model individual differences in working memory capacity (Lovett, Reder, & Lebiere, 1999). This was later demonstrated empirically by using individual performance in one simple memory task to measure the W value that best fit the individual's performance. The diagnostic memory task is called MODS, or the modified digit span task. In each MODS trial, subjects are presented strings of digits to be read aloud in synchrony with a metronome beat and are required to remember the final digit from each string for later recall. After a certain number of digit strings are thus presented, a recall prompt cues the subject to report the memory digits in the order they were presented. Each subject's MODS score was used to estimate their individual W value, which was then plugged into an ACT-R model of a separate working memory task, and the model output was used to predict *individual* data on that second task (Daily, Lovett, & Reder, 2001).

Previous and concurrent work by other groups suggested a number of positive characteristics that might be combined into a single, more powerful methodology. ACT-R parameters had been manipulated (Taatgen, 2001) in a model of individual differences in learning, although the "individuals" in that work were simulated, and corresponded to *types* of individuals, not to actual subjects. In that work, performance in a complex task was related to individual difference parameters across the simulated individuals. Earlier work in modeling also accounted for relationships between ability in one task and ability in another, but making assessments on the group, not individual, level (Just, Carpenter, and Shell, 1990). A complementary approach measures individual performance on complex tasks, and utilizes statistical methods such as intercorrelation matrices, allowing predictions of individual performance on one task based upon measurements of performance on other tasks in the matrices, making no use of any particular theory of cognition (for example, Ackerman and Kanfer, 1993).

All of this previous work, we felt, pointed towards a methodology that combined a number of positive features from these complementary approaches into a

single, more comprehensive modeling paradigm. In the methodology we envisioned, one or more simple tasks could be administered to an individual, allowing us to estimate their individual parameters; then, by plugging the individual's parameter values into the ACT-R architecture, we could predict the individual's performance on any task for which an ACT-R model exists. Because ACT-R models produce predictions with a grain size of tens to hundreds of milliseconds, this provides us with a *detailed* model of individual performance, offering the potential for predictions on the trial level, or predictions of novel measures of performance that emerge from lower-level detail – potentially allowing predictions of almost any measure that can be made of subjects. Because our approach builds atop the platform of the rich ACT-R theory, it is realistic to expect that these individualized model runs will be somewhat meaningful in their details, not just a way to arrive at a final, aggregate performance metric of some kind.

In order to take the next step beyond the Daily, Lovett, & Reder study that involved only two simple memory tasks, we decided to pick a more complex, interactive task. In order to capture a broader spectrum of individual differences, we chose to measure two parameters per subject: the W parameter as well as a measure of perceptual and motor ability, henceforth referred to as P/M. This is not a part of the standard ACT-R architecture, but seemed to be an important kind of individual variation. Thus far, we have used only one parameter, which represents as though they were one individual perception and motor speed. We allow that those may covary freely among individuals, but we have so far had success using the one parameter alone for this.

The AMBR Task

Given the preceding considerations, we chose as our more-complex task the AMBR simulation, an air traffic control task that already had a foundation as a test bed for cognitive models in a project organized by the Air Force Research Laboratory (Gluck and Pew, 2001). This task already had an ACT-R model implemented (Lebiere, Anderson, and Bothell, 2001), which not only facilitated our project, but also provided a gauge of the modularity of our approach; ideally, we would be able to plug parameters into this off-the-shelf model and obtain good results without modifying it in any other way.

The task places the subject in the role of an air traffic controller whose job is to process aircraft (AC) as they enter and leave the airspace zone, central in the simulated radar display, for which he or she is responsible. This primarily consists of issuing, via a graphical interface, two commands to an AC as it enters one's zone from a neighboring zone of airspace, and issuing two commands to an AC as it departs for another zone. The same AC must thereby be issued a total of four commands if it passes into and subsequently out of the central zone during a scenario. In some cases, the AC will only enter the central zone, or only depart the central zone, during the duration of a scenario, in which case that AC will require only a total of two commands. In addition, a fifth type of command is required if an AC requests a speed change, which requires the subject to make a trivial judgment as to whether or not the AC is on course to catch, from behind, any other AC; if so, the speed change request should be denied, and otherwise, it should be accepted. AC arrivals can be detected both from the radar display and from text messages appearing in windows to the side of the display. Speed change requests can be cued only via text messages. The departure of an AC from the central zone can be detected only via the radar display. Under the assumption that AC are at different altitudes, however, collisions cannot take place in this simulation, nor do AC land or take off in the simulation. The subject is scored based on issuing all commands in a timely fashion that permits AC to move freely without ever reaching the border of the central zone while still awaiting one of the required commands. If an AC does reach the zone border without having received all necessary commands, it will go into a hold, thereby turning the AC red in the display, halting the AC's motion, and penalizing the subject 1 point. The score at the end of the run is the sum of the errors the subject makes, lower score thereby signifying better performance. Subjects were also penalized for making interface errors of the sort that the model never made. Subject and model performance levels can thereby be compared on the basis of hold errors. A static image of the display is visible in Figure 1.

We were required to modify one aspect of the AMBR task in order to eliminate uncontrolled strategic variation among the subjects. AMBR's original implementation has a more baroque scoring system where some errors lead to penalties of 1 point and other errors up to 50 points. In response to that scoring system, some subjects tried to avoid all errors while other subjects opportunistically allowed low-penalty errors when that helped them avoid any occurrences of high-penalty errors. That strategic variation was noticed only when some data had been collected; this is an indication of the subtle difficulties that can arise when modeling tasks at the level of complexity of AMBR. The difficulty of producing a suitably correct Cognitive Task Analysis was roundly reported by the four cognitive modeling groups involved in the AFRL's AMBR modeling project. Unconstrained by the need to need to coordinate with other groups, we changed the task.

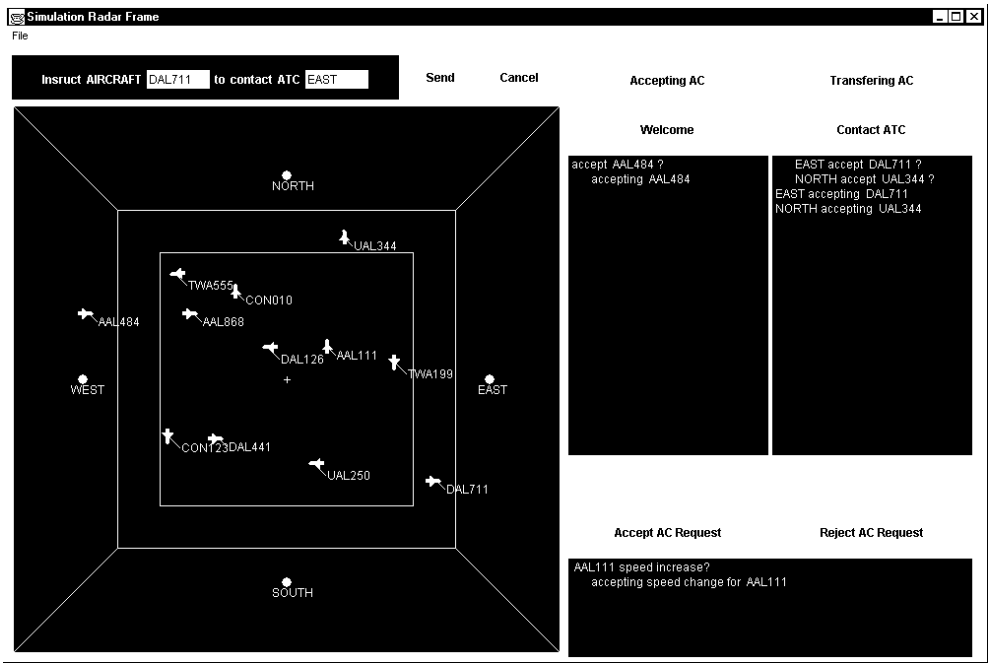


Figure 1: The AMBR Display

Sources of individual difference

Having introduced AMBR as our complex task, we acknowledge that the term "complex" is a relative one, and in seeking a complex task for our work, we were actually seeking an appropriate kind of complexity. We distinguish between distinct kinds of individual difference factors, postulating that **architectural differences** are those differences that pertain to relatively permanent characteristics of the individual, not shaped by particular episodes in the individual's experience. (We make no claims about how development shapes architectural differences throughout an individual's life.) **Knowledge-based differences**, on the other hand, can arise through specific instances of learning declarative information; the state of an individual's knowledge can only be described (or tested) in a very expansive manner, and this is not our enterprise. A third type of individual difference, **strategic differences**, could be broken down into either of the two previously mentioned types. It is not our goal to measure the encyclopedic total of an individual's knowledge, but we do anticipate that certain differences in how an individual chooses a strategy for a given task will depend upon and emerge from architectural differences. Cases where we can predict strategic differences based upon architectural differences will serve to validate our approach. We recognized that we would have trouble, however, with any task that invited strategic variation between subjects that could not be predicted from architectural differences. In such a case, our individual-difference approach would risk the same pitfalls that a non-individual-approach can lead to when subject strategies vary (Newell, 1973; Siegler, 1987).

Initial results

In two distinct experiments, with two sets of subjects, we applied the methodology of administering initial tests to measure the W and P/M parameters. The P/M parameter was actually calculated based upon the speed of mouse clicks in the AMBR training. Our procedures for calculating the parameters produce values of W and P/M which both have population means of about 1.0 and standard deviations of about 0.2 (for Carnegie Mellon undergraduates). High W means better working memory capacity, while high P/M means slower perceptual and motor responses – it is a multiplier, so that P/M = 1.2 means responses 20% slower than average). Therefore, where we find significant effects, W correlates negatively with error counts and P/M positively.

Subjects were trained on the AMBR task until they understood it quite thoroughly, and then participated in a number of AMBR scenarios, the data from which we compared to individualized model runs for each subject. Experiment 1 featured 10 AMBR scenarios, each 9 minutes long, and alternating between very easy and very difficult. Experiment 2 had 9 scenarios, each 4.5 minutes long, varying evenly along a continuum in terms of difficulty from easy to difficult. As an informal measure of difficulty, we have taken the number of AC per scenario times the average speed of those AC, divided by the scenario length. Using the idiosyncratic units of our simulations, Experiment 1 scenarios had difficulty ratings of 26 (easy) and about 180 (hard). Experiment 2 scenarios ranged in difficulty from 40 to 200.

It is instructive to note the analysis that would be performed if this were not an individual difference study. Aggregate group performance, measured in hold errors, as a function of scenarios was predicted well by the aggregate model runs (Experiment 1: r = 0.975; Experiment 2: 0.929). This was almost the same analysis, from the very same ACT-R model, presented in Lebiere, Anderson, and Bothell (2001), and used to argue for a good fit between subjects and model.

The model correctly predicted that the AMBR task, as originally conceived, is more sensitive to variation in P/M than in W. This is seen clearly in the correlations of subject holds with P/M (r = 0.658) and W (r = -0.266). Not only can the model be used to generate predictions for specific subjects, but it can also be used to probe the effects of one parameter by varying that parameter while holding the other one neutral (at the population mean of 1.0). (Note that holding one parameter fixed while varying another among the subjects is a very difficult practical matter.) This use of the model shows a strikingly greater effect upon holds from P/M than W. This is in agreement with data on the actual air traffic controller task (which, it should be noted, has several distinct differences from the AMBR task, not the least of which being that it involves voice communication, not a graphical user interface alone), which documents that only a small number of errors are due to memory failures (Billings and Cheaney, 1981).

Studies of AMBR traces reveal that the reason for this is that hold errors are primarily an outgrowth of time pressure when the time demands on a subject exceed the time that is available. For 3 of the 5 types of command in AMBR, the subject is shown the name of an AC in the text cue, and must click on the AC as part of the subsequent action sequence. Memory becomes a factor in AMBR performance primarily in that if a subject cannot remember the location of an AC based upon its name, then the display must be searched for the AC. This turns out to be a small factor because visual search is fast – slower than memory, perhaps, but the difference is on the order of a fraction of a second, while clicking in a command sequence takes several seconds whether the AC location is remembered or not. W, then, is logically a small factor in the original AMBR task, and for a small portion of the variance.

Designing for science

In order to improve upon the studies described above, we designed a follow-up study that modified the scenario difficulty, the measures of performance that we used, and even the task itself. It was obviously necessary to decrease scenario difficulty into the range for which the model produced a good fit to subject data. This also allowed us to use one performance measure that is more sensitive than hold errors – the reaction time between an action's cue and the subject's response to that cue (we used the time for an action sequence to end, meaning the third or fourth click in all). In order to emphasize the W effect relative to the P/M effect (since cognitive modeling, and not motor/kinesthetic modeling is our chief interest), we modified the task so as to create a greater penalty for failures in recall. We did this by removing AC names from the display by default, and showed the name of an AC only when the subject clicked on the AC. Moreover, only one AC name could be seen at a time, and this would appear after a delay. This change meant that the speedy visual searches of the earlier experiments would be impossible, and any failure to recall an AC's location would entail an excruciatingly slow manual search. This task modification also had the merit of giving us data on searches, and let us emphasize a performance measure that calculated what on what proportion of commands a subject found the correct AC on the first click. In other ways, Experiment 3 was similar to Experiments 1 and 2. Each subject was to participate in 5 AMBR scenarios that were easy – hold errors confound reaction time, so we needed them to be fairly rare in order to use RT as a performance measure. In the units of scenario difficulty mentioned earlier, all Experiment 3 scenarios gauged 17 or lower.

Before the study began, we ran the model, which was revised to allow for the task modification involving name-hiding, on the Experiment 3 scenarios, and it seemed not to work correctly. Instead of performing manual searches for AC names, it would *guess* which AC it was looking for and click through the entire action sequence without bothering to verify that it had clicked the right AC. While work on the model, to fix this "problem" was underway, the first subjects ran in the experiment. They behaved the same way. We had set the delay that one must wait, after clicking on an AC, for its name to appear, too long, and subjects preferred to hope that they had guessed right correctly rather than perform the laborious verification process. ACT-R came to the same conclusion based upon the undesirably large cost associated with clicks that required several seconds before the desired consequence took place. We modified the task again, shortening the delay before the name appeared, and both the model and the subjects performed manual search in the way we had hoped. This demonstrates one possible application of our approach – tasks (experimental or otherwise) can be *designed* with the model's predictions taken as a serious indicator of subject performance, individual or otherwise.

Experiment 3 produced the subject characteristics we had sought. Our three measures of individual

performance correlated significantly with W (Holds: r = -0.444; RT: r = -0.314; First-clicks: r = 0.314). P/M had about as large an impact on performance (Holds: r = 0.508; RT: r = 0.485; First-clicks: r = -0.172), but, as we desired, it did not dominate as in the first two experiments.

The result most central to our intent was the prediction of individual performance with model output (Holds: r = 0.461; RT: r = 0.436; First-clicks: r = 0.406). These correlations are distinctly less than what is often possible when averaging multiple model runs against the average of many subjects, but are very much in line with the kinds of correlations found in task intercorrelation matrix approaches (Ackerman and Kanfer, 1993; Joslyn and Hunt, 1998).

To demonstrate the possibility of precise, instance-level predictions, we looked at model predictions across all three experiments, as to whether or not, for each scenario, an individual subject would commit at least one hold error. The model predicted correctly 91.7% of the time, as detailed below in Table 1.

	Subject scenarios with errors	**Subject scenarios with no errors**
Model scenarios with errors	205	4
Model scenarios without errors	21	70

Table 1: Prediction of Error Situations

Future directions

For a variety of goals, both applied and scientific, it is and will be desirable to be able to predict individual performance on a fine-grained level. It seems certain that the methodology we are exploring will be expanded upon and utilized for such applications in the future. At present, it is possible to point to the range and extent of our successes and note the particular difficulties that individual difference modeling entails.

One avenue to explore is to involve a larger number of individual difference factors. ACT-R has many parameters built into it, and future work may be able to predict individual performance more accurately by making use of pre-tests besides the two we now use.

Because our model is fine-grained, it permits many measures of performance, on the subject, scenario, command, or click level. Ways in which the model fits, or alternately does not fit, subject data highlights many areas where future work is required. For example, we have observed in the subject data from Experiment 3 some phenomena of interest that the model does not predict. These include a correlation between higher W and the frequency with which a subject completes a sequence of command clicks *without* waiting for the AC name to appear. We believe that we can capture this with additional refinements to the model, taking advantage of ACT-R's utility-learning mechanisms. A second discrepancy between the subject data and the model predictions are that the model does not recall AC locations as well as the subjects do, and we believe that this stipulates that the model should include rehearsals of AC location between the time that information is learned and when it is needed. A third difference is that subjects often respond to Welcome commands, which are always the second of a pair of commands regarding a given AC, much faster than the model does. In fact, some subjects respond much faster than other subjects in this regard, and it is clear that strategic variation has intruded into our study – something that is difficult to prevent absolutely with a task of AMBR's complexity. In upcoming experiments, we will try to instruct all subjects to anticipate Welcome commands when they can, and will change the model so that it does so as well.

Subject phenomena that are not captured by the model, we believe, stem from the problem of deriving a valid Cognitive Task Analysis, which is known to be difficult for a novel, complex task. It is striking how much simpler AMBR is than many tasks (for example, *real* air traffic control), and yet how challenging it is to model it precisely. It has not only been a challenging task to which to extend the individual difference methodology from memory to more complex tasks; it is also at the right level of complexity for the next stages of work as we try to model it still more accurately and over a variety of task modifications.

Acknowledgments

This research was supported in large part by ONR Grant N00014-02-10020. Thanks to Susan Chipman for her support.

References

Ackerman, P. L. & Kanfer, R. (1993). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. *Journal of Applied Psychology*, *78(3)*, 413-432.

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Billings, C. and Cheaney, E. (1981) The information transfer problem: summary and comments. *Information Transfer Problems in the Aviation System*. NASA Technical Paper 1875. NASA, California.

Carpenter, P.A., Just, M.A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, 97 (3), 404-431.

Daily, L. Z., Lovett, M. C. & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, 25, 315-353.

Gluck K. A., & Pew, R. W. (2001). Overview of the Agent-based Modeling and Behavior Representation (AMBR) model comparison project. *Proceedings of the 10th Computer Generated Forces and Behavior Representation Conference*, Orlando, FL.

Joslyn, S. & Hunt, E. (1998). Evaluating individual differences in response to time-pressure situations. *Journal of Experimental Psychology: Applied*, *4(1)*, 16-43.

Lebiere, C., Anderson, J. R., & Bothell, D. (2001). Multi-tasking and cognitive workload in an ACT-R model of a simplified air traffic control task. *Proceedings of the 10th Computer Generated Forces and Behavior Representation Conference*, Norfolk, VA.

Lovett, M. C., Reder, L. M., & Lebiere, C. (1999). Modeling individual differences in a digit working memory task. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 460-465). Mahwah, NJ: Erlbaum.

Newell, A. (1973). *You can't play 20 Questions with nature and win: Projective comments on the papers of this symposium.* In W. Chase (Ed.), Visual information processing. New York: Academic Press.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example of children's addition. *Journal of Experimental Psychology: General*, 106, 250-264.

Taatgen, N. A. (1999). A model of learning task-specific knowledge for a new task. In *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp. 730-735). Mahwah, NJ: Erlbaum.

Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and its Relationship to Discourse Comprehension

Daniel C. Richardson (richardson@psych.stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305, USA

Rick Dale (rad28@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853, USA

Abstract

While their eye movements were being recorded, participants spoke extemporaneously about a TV show whose cast members they were viewing. Later, other participants listened to these speeches while their eyes were tracked. Within this naturalistic paradigm using spontaneous speech, a number of results linking eye movements to speech comprehension, speech production and memory were replicated. More importantly, a cross-recurrence analysis demonstrated that speaker and listener eye movements were coupled, and that the strength of this relationship positively correlated with listeners' comprehension. Just as the mental state of a single person can be reflected in patterns of eye movements, the commonality of mental states that is brought about by successful communication is mirrored in a similarity between speaker and listener's eye movements.

Introduction

Imagine standing in front of a painting, discussing it with a friend. As you talk, your eyes will scan across the image, moving approximately three times a second. They will be drawn by characteristics of the image itself, areas of contrast or detail, as well as features of the objects or people portrayed. Eye movements are driven both by properties of the visual world and processes in a person's mind. Your gaze might also be influenced by what your friend is saying, what you say in reply, what is thought but not said, and where you agree and disagree. If this is so, what is the relationship between your eye movements and those of your friend? How is that relationship related to the flow of conversation between you?

Language use often occurs within rich visual contexts such as this, and the interplay between linguistic processes and visual perception is of increasing interest to psycholinguists and vision researchers (Henderson & Ferreira, 2004). As yet, however, such processes have been limited to experiments that examine the eye movements of the speaker or the listener in isolation. Language use, more often than not, occurs within a richer social context as well.

Direct eye contact between conversants plays an interesting, crucial role in coordinating a conversation (Bavelas, Coates, & Johnson, 2002), and in conveying various attitudes or social roles (Argyle & Cook, 1976). The focus of the current experiment, however, is cases such as those introduced at the outset, where conversants are not looking at each other, but at some visual scene that is the topic of the conversation. More common examples might be discussing a diagram drawn on a whiteboard, figuring out together how to do something on a computer, or talking during a movie.

Uniquely poised between perception and cognition, eye movements can reveal cognitive processes such as speech planning, language comprehension, memory, mental imagery and decision making. The current experiment investigates whether the eye movements of a speaker and a listener to a visual common ground can provide insight into a discourse.

Eye movement Research

Eye movements of a speaker

If a speaker is asked to describe a simple scene, they will fixate the objects in the order in which they are mentioned, around 900ms before naming them (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998). Since such pictures can be identified rapidly, it is argued that during this time speakers are not just retrieving words but selecting and planning which to use.

Eye movements of a listener

Eye movement research has shown that there is a tight interdependence between speech recognition and visual perception. Eye movements to potential referents for a word can provide evidence for a lexical item being recognized before the word is finished being spoken. The link between visual and linguistic processing can also be seen in eye movements that disambiguate syntactic structures (Tanenhaus, Spivey Knowlton, Eberhard, & Sedivy, 1995) and anticipate the future agents of actions (Kamide, Altmann, & Haywood, 2003). Recent studies of the eye-movements of a participant engaged in a conversation with another naïve participant reveal a remarkable sensitivity to the referential domains established by the task, the visual context and the preceding conversation (Brown-Schmidt, Campana, & Tanenhaus, 2004). Qualitatively, eye movement research reveals a very close, time-locked integration between visual and linguistic processing (Tanenhaus, Magnuson, Dahan, & Chambers, 2000). Although fixation times are heavily modulated by context, as a very rough quantitative guide, research suggests that listeners will fixate an object around 400-800ms after the name onset.

Eye movements of a thinker

Since participants will make systematic eye movements to entirely empty and uninformative regions of space when retrieving information from memory (Richardson & Kirkham, in press; Richardson & Spivey, 2000) or listening to a story (Spivey & Geng, 2001) it is clear that they can be governed by cognitive as well as perceptual processes. Influencing how the eye moves across an image can have profound effects on mental processes. Researchers have recorded the eye movements of participants interpreting an ambiguous picture in a particular way, or solving a difficult deductive problem from a diagram. Using low level visual cues, a second set of participants were then influenced to attend to the same regions of the picture. The second set of participants were more like to form the same interpretation of the ambiguous picture (Pomplun, Ritter, & Velichkovsky, 1996), and remarkably, were more likely to solve the deductive problem (Grant & Spivey, 2003). If forced similarity between participants' eye movements can result in similar cognitive states, then will the similar cognitive states that are brought about by successful verbal communication result in similar eye movement patterns between speaker and listener?

Experiment

Speech production and speech comprehension have previously been studied in separate eye tracking paradigms. Yet if both are indeed closely linked to eye movements, the eye movement patterns of two people engaged in a natural, unscripted conversation may bare some relationship to each other. Moreover, it raises the intriguing possibility that the strength of the relationship between conversants' eye movements will parallel the success of their linguistic relationship.

The current experiment approximates a conversation between naïve conversants by asking participants to speak spontaneously, with neither a script nor a rehearsal for an extended period of time about a TV show, whose characters were displayed in front of them. These speeches were then played back to other participants who were looking at the same display. Crucially, both the speakers' and the listeners' eye movements were tracked throughout. The listeners' comprehension was then measured by a series of content questions. Thus in addition to extending various eye movement-language results to natural, spontaneous speech, the current experiment was able to investigate a number of entirely novel hypotheses regarding the linkage between speaker and listener eye movements, and its relation to the listener's comprehension.

Methods

The first four participants recruited to take part in this experiment were designated as speakers, and the remainder were listeners. The methods for both stages will be described below.

Participants

40 Stanford undergraduates took part in the experiment in exchange for course credit.

Apparatus

An ASL 504 remote eye tracking camera was positioned at the base of a 17" LCD stimulus display. Participants were unrestrained, and sat approximately 30" from the screen. The camera detected pupil and corneal reflection position from the right eye, and the eye tracking PC calculated point-of-gaze in terms of co-ordinates on the stimulus display. This information was passed every 33ms to a PowerMac G4 which controlled the stimulus presentation and collected looking time data. Prior to the experimental session, the participants went through a 9 point calibration routine, which typically took between 2 and 5 minutes.

Speakers' voices were recorded by microphone, and listeners made responses using the two buttons of a mouse held in their lap.

Design – Speakers

The intention was to record participants speaking spontaneously about a TV show while looking at a picture of the cast members. In the first case, a picture of the 6 principal characters of the cast of the TV sitcom *Friends* was used. The characters were shown individual in 6 separate pictures. Potential speakers were asked if they knew they show and would like to talk about it, and two speakers were selected who were knowledgeable and reasonably gregarious. Speakers were instructed to 'Talk about the show for a couple of minutes. You could talk about the relationships between the characters, your opinion of them, or your favourite episode'. In the second case, two participants were shown a 5 minute scene from *The Simpsons* during which they undergo family therapy. These participants were then shown a picture of the five family members and their therapist. The participants were asked to 'Describe what went on in the scene and what you thought about it'.

As they spoke, the speakers' eye movements were tracked and their voices were recorded by microphone. These recordings were trimmed so that they were all roughly one minute long, and the text was transcribed for later analysis.

Design - Listeners

Participants listened while looking at the same picture of the six cast members that had been in front of the speaker. Since there could not be systematic looks to the cast members if the participant did not recognize any of them, participants were first asked if they were familiar with either show. On this basis, the listeners were presented with one or both of the *Friends* and *Simpsons* stimuli, and were randomly assigned one of the two speakers.

Listeners heard a minute of speech, and then a screen appeared warning them that the question period was about to start. In the four question trials, participants saw six solid grey circles or squares in the locations where pictures of the individual cast members had previously appeared. After a 1000ms pause, they heard a question and responded yes or

no using the two mouse buttons. There followed a 2000ms ISI during which the screen was blank.

The questions were recorded by the experimenter and were of the form, "Did the speaker say…?". The questions were designed such that they could not be answered on the basis of knowledge about *Friends* or *The Simpsons* alone, but were specific to the information mentioned (or not) by that particular speaker. The correct answer to half the questions was yes and half no.

Data Coding

Roughly half of our listeners were familiar with both TV shows and half knew the characters from only one. All analyses are based on 49 usable listener-speaker dyads. A further 9 cases were dropped due to problems with the equipment or the calibration procedure.

The eye movements of the speaker and of the listener during the minute of speech were analyzed in exactly the same way. The eye movement data relayed which, if any, of the six pictures were being fixated every 33ms. The data were cleaned for blinks and saccades across a picture - only stable fixations longer than 99ms were analyzed – and then expressed in terms of a sequences of gaze onsets and offsets in the six pictures.

The speakers' recordings were transcribed with onset times for each word spoken. In addition, words were flagged if they were names of any of the six characters pictured. Listener responses to the questions were coded for accuracy, and their looking times to each of the pictures while answering were calculated.

Results and Discussion

This experiment provided precise timing information about speakers' speech and gaze onsets, and listeners' gaze onsets. This information can depicted graphically in what we call a 'scarf plot', which represents a transcript of the speech together with the timing of word onsets and the eye movements of both speaker and listen. Figure 1 shows an nine second segment of a scarf plot for one speaker-listener dyad. Such eye movement data can be statistically analyzed and compared with the objective measure of the listeners' understanding of the speech provided by their performance answering four comprehension questions.

Before the detailed inferential analyses begin, it is useful to get a rough sense of the behavior being studied. On average, speakers used 160 words, only 12 of which were the names of the characters depicted. It is important to note that the speeches were not edited for content, and include all the deviations, hesitations and repetitions that are typical of just a minute of normal, spontaneous speech.

Speakers and listeners switched their gaze between pictures around 120 times. For each occasion, they spent about 500ms looking at the picture. Since the average eye fixation lasts 200-300ms, it is reasonable to assume that this represents two fixations within the same picture.

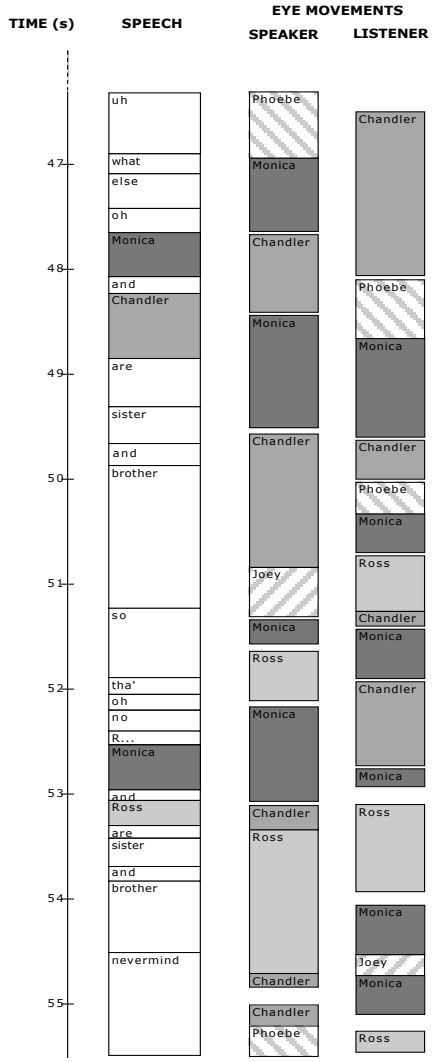


Figure 1. Scarf Plot of a 9 second segment of one dyad. The speaker's words are shown on the left, with nouns highlighted. The speaker's and listener's eye movements are shown in the middle and right columns respectively. Time is on the y axis, increasing down the page.

Speaker Fixations Prior To Naming

For each occasion that the speaker named character X, their eye movement data were consulted to find the point at which X was last previously fixated. The difference between the gaze onset and the name onset was computed for every name used by every speaker. On average, a character was fixated 860 ms prior to being named.

This lag is exactly in the range reported by the speech production literature (Griffin & Bock, 2000), where typically participants are explicitly instructed to describe a simple picture. We have found a lag of the same magnitude with spontaneous, natural speech, when participants are describing not what is front of them per se, but things that are not depicted - stories, opinions, relationships – that relate to those characters.

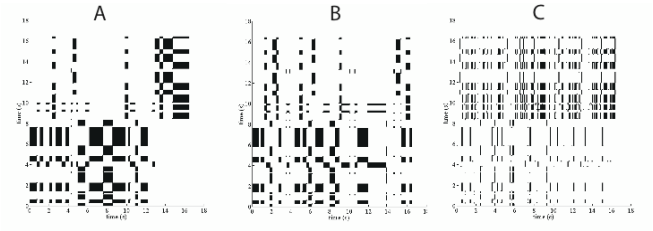


Figure 2. Example CRPs

Relationship Between Speaker And Listener Eye Movements

To what degree were speaker and listener looking at the same thing at the same time?

We quantified this question by generating categorical cross-recurrence plots between the speaker and listener time series of fixations (Dale & Spivey, in submission). These plots permit visualization and quantification of recurrent patterns of states between two time series (see Shockley, Santana & Fowler, 2003, for a fuller introduction; see Eckmann, Kamphorst & Ruelle, 1987; Zbilut & Webber, 1992 for foundational treatises). In our case, the cross-recurrence plot portrays the extent to which dyad fixations are overlapping temporally.

To begin, windows of a given length are moved along each time series, forming individual windows at every time index. The windows of each time series are then compared to *all* those of the other time series (comparing *every* time index). At time index *i* for the first time series and *j* for the second, if their windows are sufficiently similar, a point *(i, j)* is recorded on a two-dimensional plot. By comparing every window in the first to the second time series, we can generate a full plot of points in which the two time series are close to each other – a cross-recurrence plot.

For simplicity, we used a window size of 1 for our analysis. By using a categorical metric (see Dale & Spivey, in submission, for details), we have the criterion that dyad fixations are recurrent if falling on the same object for 33ms. We generated plots using this metric between every speaker-listener pair. Figure 2 shows example cross recurrence plots between a speaker and (a) a listener who answered all comprehension questions correctly (b) a listener who answered few correctly, and (c) a listener with their eye movement data placed in a random order. There are three things to notice here. Firstly, the good listener has higher density in their plot, indicating more points of recurrence with the speaker. Secondly, both listeners have more structured plots compared to the randomized series. Lastly, one can see that for the two real listeners there is a higher density in the region on and below the *i=j* diagonal. This indicates that the speaker and listeners' eye movements overlapped more when the listeners' eye movements lagged behind the speakers.

We employed a further analysis to find out exactly what temporal lag between the listener and the speaker would produce the greatest degree of recurrence, or overlap, between their eye movement patterns. Listener time series were successively lagged by 330ms. On the line defined by $i = j$ in the plot (the *line of incidence*), any point indicates that *in the same temporal context* fixations are recurrent. Thus, by lagging the listeners' time series, and recording maximal recurrence along the line of incidence within each lag, we get a measure of the extent to which dyads' eye movements are related. Though our chosen window size is small, the results are quite compelling.

Figure 3 shows the degree of recurrence between speaker and listener at different time lags, averaged across all 49 dyads. We also randomized listeners' eye movement data and calculated its recurrence with the speakers'. This randomized series serves as a baseline of looking 'at chance' at any given point in time, but with the same overall distribution of looks to each picture as the real listeners.

A 2 (listeners/randomized listener) x 40 (lag times) ANOVA revealed a significant main effect of listener type ($F(1,45)=785.5$, $p<.0001$) and a main effect of lag ($F(40,1800)=25.2$, $p<.001$). Moreover, there was a significant interaction between the factors ($F(40,1800)=24.7$, $p<.001$).

Clearly, the real listeners are not looking around these displays randomly. Rather their eye movements are linked to the speakers', and this relationship has a temporal character. More precisely, the maximum recurrence between the speakers and listeners, the lag time at which their eye movements overlap the most, is at 1650ms

These results are exactly what one would expect from the combination of the speech production and speech comprehension eye movement literature. Typically, speakers will fixate an item 900ms before naming it and listeners will fixate an object around 800ms after the name onset. Very roughly this would suggest we would find a lag of 900+800=1700ms between speaker gaze onsets and listeners'. This derived value corresponds both to the exact lag that produces a maximum recurrence value, 1650ms, and the general region of higher recurrence in the 1000-2000ms range.

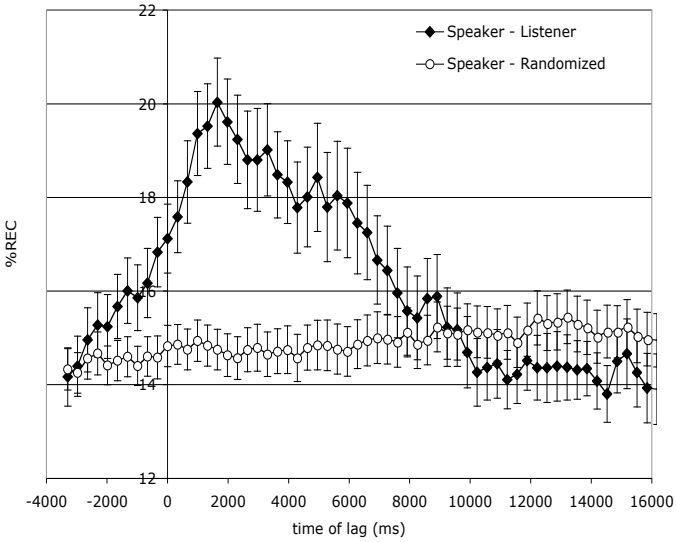


Figure 3. Cross recurrence at different time lags

The speech production and comprehension literatures, however, deal with cases where an object or person is explicitly named. Perhaps it is the case that the differences between critical and non-critical gaze onset lag distributions observed here are due mainly to the occasions when the speaker planned and spoke out loud a name of one of the characters pictured.

This question was addressed by examining a subset of the data. The name-subset includes only speaker fixations to person X that were immediately prior to the speech onset of name X. As noted previously, since there were on average 12 cases of name use, this constituted about 10% of the 120 fixations the average speaker made.

Figure 4A plots the recurrence at different time lags for the name subset of our data. The 2 (listeners/randomized listener) x 40 (lag times) ANOVA revealed a significant main effect of listener type ($F(1,45)=192.3$, $p<.0001$) and a main effect of lag ($F(40,1800)=28.1$, $p<.001$). As before, there was a significant interaction between the factors ($F(40,1800)=27.5$, $p<.001$).

For the subset of speaker fixations that precede a name, there is a highly pronounced difference between the speaker and the listener and the speaker and randomized looking. Once more, the greatest extent of this difference is just before 2000ms. Again, this would be predicted by the speech production and comprehension eye movement literatures. Is it the case, then, that the current experiment has simply replicated these name-use results using spontaneous speech?

To answer this question the data excluded from the name analysis above were analyzed in isolation. Figure 4B plots the 'non name dataset' that corresponds to the 90% of speaker fixations to person X which were not immediately followed by X being named out loud. The ANOVA showed a similar pattern of results: main effect of listener type ($F(1,45)=559$, $p<.0001$), a main effect of lag ($F(40,1800)=25.8$, $p<.001$). and a significant interaction between the factors ($F(40,1800)=25.0$, $p<.001$). Although subtracting the cases of name use from the full data set appeared to attenuate somewhat the differences between critical and non-critical gaze onset lags, it is certainly the case that these distributions still differ. In other words, it is not just the when the speaker names a character that speaker and listener eye movements are linked. It must be other properties of the discourse (implicit reference, anaphor, topics, agents, for example) which drive the speakers eye movements while they are being planned, and a few seconds later, influence the listener's eye movements once they are spoken.

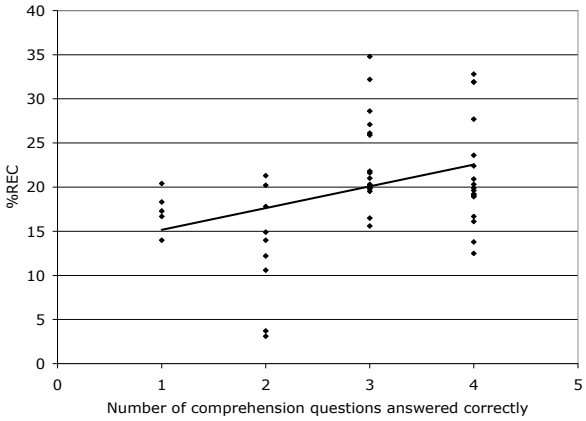


Figure 5. Correlation Between Speaker-Listener Eye Movements Coupling and Listener Comprehension

Speaker-Listener Eye Movement Linkage and Listener Comprehension

The degree to which eye movements were linked in a given speaker-listener dyad were compared with the listener's comprehension of what had been said. For each dyad, we computed the degree of recurrence (REC%) at a lag of 1650ms between speaker and listener. This is the lag that produced the greatest recurrence across our whole data set, and hence serves as a baseline to compare the linkage between individual speaker-listener dyads. The performance of listeners answering four comprehension questions was taken as an objective measure of how well they had comprehended the one minute of speech.

A regression analysis was performed on this data, and found that a linear fit had $r^2=0.14$. Although it may not account for a large portion of the variance in participants' behaviour, an ANOVA shows that this relationship is significant ($F(1,47)=7.39$, $p<.01$).

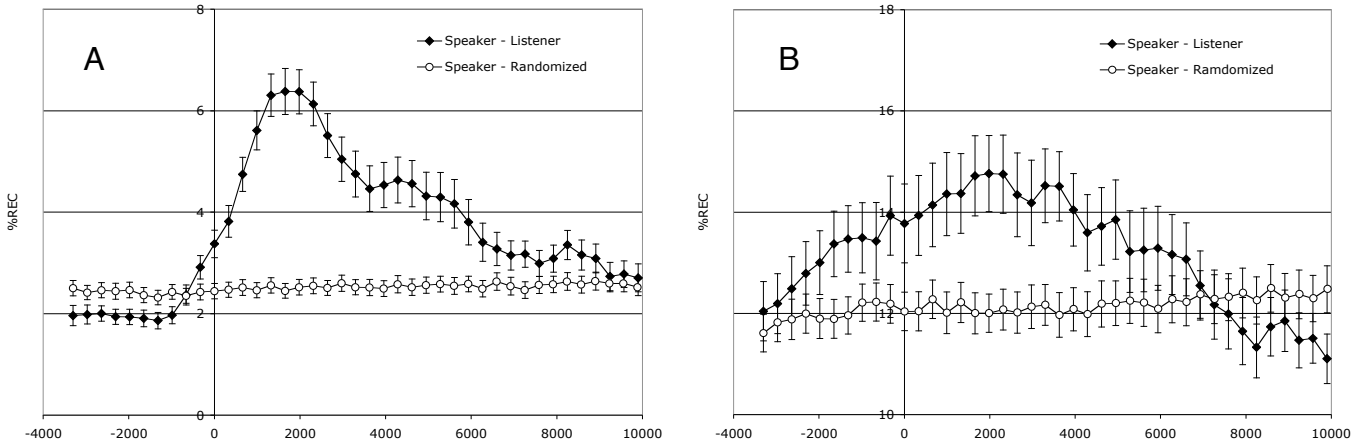


Figure 4. Cross recurrence at different time lags for (a) name fixations, (b) non-name fixations

General Discussion

The current experiment uses a naturalistic paradigm that elicits and presents spontaneous speech. The language-use in this experiment is grounded in the visual items presented on the display, but is not a description of them per se, or an explicit instruction relating to their presence or appearance. Nevertheless, this single paradigm replicates several results obtained in more constrained circumstances concerning the relationship between eye movements, speech production, and speech comprehension.

More importantly, this experiment provides what could be the first demonstration that during the production and comprehension of a spontaneous discourse, the eye movements of a speaker and a listener are coupled. Moreover, this relationship between eye movement patterns is not driven by cases in which the speaker explicitly names people who are depicted. It seems to be that the planning of more diverse types of reference and foregrounding may be influencing the speaker's eye movements, and, a few seconds later via the speech stream, influencing the listener's eye movements. Crucially, the strength of relationship between the speaker's and the listener's eye movements appears to predict the degree to which the listener successfully comprehended the speech.

Instances of new paradigms such as this inevitably raise many questions for future research. Is it the case that a tight coupling between speaker and listener eye movements is an overall indication of listener attentiveness, which also predicts listener comprehension? Or is it that by rapidly bringing their eyes to bear on the same item as the speaker, good listeners receive appropriate visual information that supports the verbal input? Or perhaps it is not so much that moving the eyes closely in step with a speaker brings in visual content, but rather it is an indication (or a cause) that the listener is using spatial information to cognitively structure the information in the same way as the speaker?

The close relationship between speaker and listener eye movements and the success of the discourse clearly aligns with a view of language use as a joint activity (Clark, 1996), in which successful communication is brought about by a successful coordination of information in the common ground. The human eye only receives detailed information from 2º of its visual field: therefore, if the speaker and listener are looking at exactly the same thing, then they are certainly sharing a higher, common ground.

Acknowledgments

The authors are indebted to Natasha Kirkham, Herb Clark and Michael Spivey.

References

Argyle, M., & Cook, M. (1976). Gaze and Mutual Gaze. Cambridge: Cambridge University Press.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. Journal of Communication, 52(3), 566-580.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2004). Real-time reference resolution by naïve participants during a task-based unscripted conversation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), World-situated language processing: Bridging the language as product and language as action traditions. Cambridge: MIT Press.

Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.

Dale, R. & Spivey, M. J. (submitted). *Data visualization of complex behavioral structure across time.* Manuscript submitted for publication.

Eckmann, J.-P., Kamphorst, S.O., Ruelle, D. (1987). Recurrence lots of dynamical systems. *Europhysics Letters, 5*, 973-977.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. Psychological Science, 14(5), 462-466.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. Psychological Science, 11(4), 274-279.

Henderson, J. M., & Ferreira, F. (Eds.). (2004). The integration of language, vision, and action: Eye movements and the visual world. New York: Psychology Press.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory & Language, 49(1), 133-156.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. Cognition, 66(2), B25-B33.

Pomplun, M., Ritter, H., & Velichkovsky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. Perception, 25(8), 931-948.

Richardson, D. C., & Kirkham, N. Z. (in press). Multi-modal events and moving locations: evidence for dynamic spatial indexing in adults and six month olds. Journal of Experimental Psychology: General.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. Cognition, 76, 269-295.

Shockley, K., Santana, M. V. & Fowler, C.A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 326-332.

Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. Psychological Research/Psychologische Forschung, 65(4), 235-241.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. Journal of Psycholinguistic Research, 29(6), 557-580.

Tanenhaus, M. K., Spivey Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632-1634.

Zbilut, J. P. & Webber, C. L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A, 171*, 199-203.

Working Memory and Inhibition as Constraints on Children's Development of Analogical Reasoning

Lindsey E. Richland (lengle@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Robert G. Morrison (robertmorrison@xunesis.org)
Xunesis
PO Box 269187, Chicago, IL 60626-9187

Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Abstract

We developed a picture-mapping task (Richland Picture Analogies, RPA) to examine the roles of inhibition and working-memory load on children's development of analogical reasoning. Children of ages 3-4, 6-8, and 13-14 were instructed to use relational correspondences between source and target pictures to select the target object corresponding most directly to a specified source object. The study examined age trends in children's proficiency with analogical reasoning. Relational complexity and perceptual distraction were manipulated to investigate how maturational constraints interact with each other and with age. Results indicate that children's development of the capacity to reason analogically interacts with increases in working-memory capacity and inhibitory control.

Children's higher-order reasoning skills are central to their ability to transfer knowledge from an initial learning context to future environments. This process enables children to understand novel situations and contexts, to build on their everyday learning experiences and to develop a flexible body of knowledge (Gentner, Holyoak & Kokinov, 2001; Gentner & Rattermann; 1991; Holyoak, Junn & Billman, 1984). Following Gentner (1983), analogy is defined as a conceptual strategy in which a source object is represented as similar to a target object, and correspondences are mapped between the two analogs. Although there is wide agreement that this conceptual process is central to children's everyday learning, the mechanisms underlying and constraining the development of analogical reasoning are not yet well understood.

The process of constructing an analogy requires a reasoner to represent source and target analogs, maintain both representations in working memory (WM; Hummel & Holyoak, 1997, 2003), and construct a mapping between elements of the source and target based upon correspondences between relations in each (Gentner, 1983; Holyoak & Thagard, 1989). Critically, the relational correspondences may compete with more superficial perceptual or semantic similarities between individual objects, requiring inhibitory control when relational and more superficial responses conflict (Gentner & Toupin, 1986; Morrison et al., 2004; Viskontas et al., in press).

Proposed Developmental Mechanisms

Researchers have proposed three developmental mechanisms to explain age-related changes in children's performance on analogical reasoning tasks: increased domain knowledge, a relational shift, and increased WM capacity for manipulating relations.

Goswami (1992, 2001) proposed domain knowledge as the primary mechanism underlying developmental changes in analogical reasoning. According to her relational primacy hypothesis, analogical reasoning is available as a capacity from early infancy, but children's analogical performance increases with age due to increased knowledge about relevant relations. This hypothesis was developed in reaction to Piagetian studies suggesting that children are unable to reason analogically prior to achieving formal operations, approximately at age 13 or 14 (Piaget, Montangero & Billeter, 1977). Piaget's tasks frequently involved uncommon relations, such as "steering mechanism", which would likely have been unfamiliar to younger children. In contrast, research has since shown children can reason analogically at much

younger ages (e.g. Gentner 1977, Holyoak et al., 1984). Goswami and Brown (1989) argued that children as young as 3 years old were successful on analogical reasoning tasks when they demonstrated knowledge about the relevant relations. Goswami and Brown presented children with complex versions of analogy tasks in which two physical, causal relations (e.g., cutting and wetting) were imposed on a source object "a" to become source object "b." Children were required to map on the basis of these relations to complete an analogy of the form a:b::c:d. The investigators found that children were fairly competent on these problems with 2 relational changes when they showed knowledge of the relations.

These data provided some evidence that domain knowledge is related to successful analogical reasoning, but the methodology of this study has been criticized.

Rattermann and Gentner (1998) found that when a substantial perceptual distractor was included in the Goswami and Brown stimuli, children younger than age five were likely to select a perceptual match in spite of knowledge of the relations and explicit analogy instructions. Gentner and Rattermann (1991; Rattermann & Gentner, 1998) posited that a "relational shift" occurs between the ages of four and five. Before the relational shift, they argue that children primarily attend to perceptual similarity and will reason on the basis of perceptual features if available. Following the relational shift, children can and do reason on the basis of relational features even when faced with perceptual distractors. The authors suggest that domain knowledge is integral to the relational shift, though the mechanism is not explicitly postulated.

An alternative explanation for the relational shift is that children younger than age five were unable to inhibit their responses to perceptual similarity, although they were aware that the task required attention to relational similarity. It is well-established that children's inhibition capacity develops with age (Diamond, Kirkham & Amso, 2002), and follows similar age-related patterns as does analogical reasoning. Accordingly, development of children's inhibitory capacity, one aspect of the human working memory system, may underlie children's patterns of success and failure on analogical reasoning tasks.

Finally, WM constraints have been proposed to explain developmental change in analogical reasoning. In particular, relational complexity has been argued to constrain children's performance on analogical reasoning tasks (Halford, 1993). Two primary definitions of relational complexity have been advanced. Zelazo et al, (2003, 1998; Frye & Zelazo, 1998; Frye et al., 1996) define complexity as the number of hierarchical rules that must be maintained in working memory in order to accomplish a task, a view proposed as Cognitive Complexity and Control (CCC) theory. For example, in the Dimensional Change Card Sort (DCCS) task, children were asked to follow a rule to sort by color (e.g., "if red ... here" and "if blue ... here") and a rule to sort by shape (e.g. "if rabbit... here" and "if boat...here"). Children ages 3-4 were successful on these sorting tasks when performing them separately, but failed when required to integrate these two within a higher-order rule.

Halford (1993; Andrews & Halford, 2003; Halford et. al, 2002) has argued that relational complexity is more generally a constraint on the number of distinct units of information that must be processed in parallel while being maintained in WM in order for a reasoner to complete a task. Using this metric of relational complexity, Halford has argued for a developmental continuum in children's relational complexity capacity such that until approximately age four, children can process binary relations (a relationship between two objects) but not ternary relations (relationships among three objects, equivalent to the integration of 2 binary relations).

The three hypotheses are not mutually exclusive, and the relationships among the empirical factors emphasized by each model have not been fully examined. The present project examines the interactions among the constraints at the heart of the three models: the role of domain knowledge, inhibition of perceptual distraction, and relational complexity.

Picture Analogy Task

We developed a set of materials for a picture-analogy task suitable for children across a wide age range. The general structure of the stimuli was modeled after those developed by Markman and Gentner (1993), with inclusion of additional controls and using content accessible to young children. Our picture set (Richland Picture Analogies, or RPA task; available from first author upon request) was designed to examine the impact of relational complexity and perceptual distraction (i.e., need for inhibition) on children from age 3 yrs, while controlling for domain knowledge. The RPA stimuli depict relational motion verbs of the type learned early in children's vocabulary acquisition (e.g., Golinkoff et. al., 1996; Golinkoff et al, 1995; Gentner, 1978). Relations were motion verbs with perceptually available meanings that are familiar to young children by the age of 3 (e.g., "kiss", "chase" and "feed"). The objects used to represent these relations were items regularly encountered by preschool age children, including humans, animals, and dolls. Counterbalanced versions of each picture set factorially varied number of relevant relations (1 or 2) and presence vs. absence of a perceptual distractor in a 2x2 design. Perceptual distractors were either exact matches to the source object located within the target picture or were slight variations of the same object (e.g., a cat chasing and a cat sitting). In the no distractor conditions, a neutral object replaced the featural match. The spatial location was held constant. Unlike in the Markman and Gentner stimuli, distractors were never placed in key relational roles (allowing perceptual and relational errors to be coded separately), and the number of objects in each picture was controlled.

Method

Participants

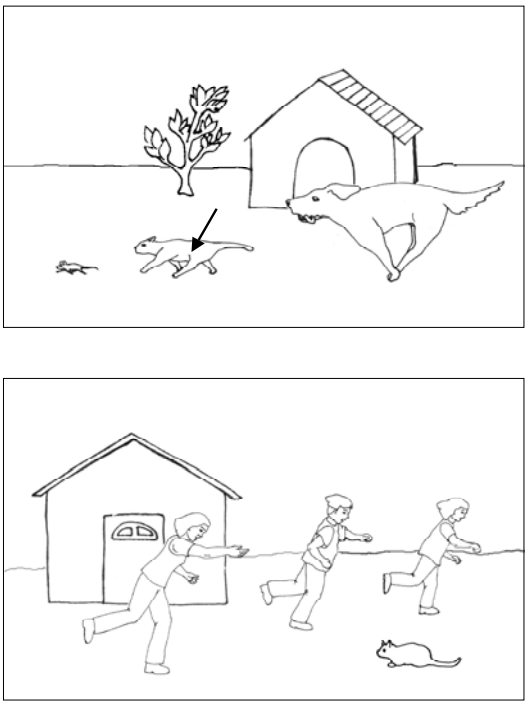
The participants were 68 children: 22 aged 3-4 years, 21 aged 6-8, and 25 aged 13-14. They were enrolled in preschool, elementary, and junior high school programs in the New York City and Los Angeles areas.

Materials and Design

The RPA task consists of 20 pairs of source and target pictures in which objects in the paired pictures depict the same relationship using unique objects. On a single page, participants viewed the two pictures in a set. An arrow pointed to a source object in the top picture, and the participant was asked to select the corresponding object in the bottom picture (cf. Markman & Gentner, 1993). For the example in Figure 1, the top picture represents "dog chasing cat chasing mouse" and the bottom picture represents "woman chasing boy chasing girl". If an arrow pointed to the cat, the correct relational response would be the boy in the bottom picture. All pictures contained extra items not depicting the relevant relationship, and the number of total objects was standardized across pictures per condition. Most image sets contain a total of five objects.

Four versions of each picture set were constructed in order to manipulate two variables in a 2x2 design. The first variable was the presence or absence of a perceptual distractor in the target picture, defined by strong featural similarity to an object in the source picture. The featural distractor was either an identical match to an object in the source picture or was the same object in a slightly different position. For example, in Figure 1 (top) the cat is depicted sitting in the target picture but is not involved in the chase. The featural distractor is never involved in the relational structure of the target picture. In Figure 1 (bottom), the correct relational response is the boy; however the participant must inhibit the featural match to make this choice. When present, the featural distractor spatially replaces an alternative object in the target picture. As a control to ensure that the featural distractors were indeed perceptually distracting, ten undergraduates were asked to select the most perceptually similar object to the target in the 2R-D version of each stimuli. Participants selected the intended featural match 96% of the time, indicating that the manipulation of perceptual similarity is valid.

The second variable was the number of relations, one or two, that participants were required to process simultaneously in order to accurately select a target object. When two relations were involved, the correct target object was both agent and recipient of a relation. For example, in Figure 1 the top picture represents "dog chases cat" and "cat chases mouse", whereas the bottom picture depicts "mom chases boy" and "boy chases girl". If the participant only considered one of the relations in each picture, there would be two equally plausible answer choices, and participants would be expected to perform at a 50% level at best. In this example the boy is the correct relational response because he (uniquely) is both being chased and is chasing. Making this determination requires integration of two binary relations in each picture.



Problem 2-2a

Figure 1. Sample stimuli, two relations with distractor (R2-D). The cat in the top picture (both chaser and chased) maps relationally to the boy in the bottom picture.

The 2x2 repeated-measures portion of the design generated four conditions: one relation, no featural distractor (R1-N), one relation with featural distractor (R1-D), two relations, no featural distractor (R2-N), two relations, featural distractor (R2-D). Packets of picture pairs for each participant were organized such that five examples of each condition were included in a random order. The assignment of specific picture pairs to each of the four conditions was counterbalanced across participants in each age group. The three age groups constituted an additional between-subjects factor. The dependent variable was participants' object choice within target pictures.

Procedure

The task was administered to participants in paper form. All participants were given two sample problems,

one involving one relation and the other involving two relations. The instructions stated that "a certain pattern exists in both the top picture and the bottom picture, and the child's job is to find this pattern." Following the first sample problem, (a 1R-D problem), it was explained that "some pictures have two parts of the pattern like that one, and others have three parts" (demonstrated subsequently in the 2-relation sample problem). The child was taught that an object in the top picture would be highlighted by an arrow, and they were to point or draw an arrow to the corresponding object in the bottom picture. For both sample problems, children were asked to point to the correct answer and then were given feedback. Feedback was repeated until they gave the correct answer. If they failed initially on both sample problems, their performance on the first 5 problems was used as criteria for exclusion. If participants failed on more than 3 problems, their data was excluded from analysis.

The problems were presented in random order following the sample problems. The task was administered to the 13-14 year old participants in groups; all other children were tested individually by a single experimenter.

Results

Figure 2 presents the proportion of correct relational responses for each of the four picture conditions as a function of age. An analysis of variance (ANOVA) was performed to examine the effects of age, relational complexity, and distractor condition on children's proportion of correct relational choices. The ANOVA revealed main effects of age, $F(2, 65) = 78.15$, $p<.001$, featural distraction, $F(1, 65) = 26.07$, $p<.001$, and relational complexity, $F(1, 65) = 24.83$, $p< .001$. These results establish that the RPA task is sensitive to age, that the picture manipulations were effective at creating distraction and increasing WM load, and that these constraints actively impede children's analogical reasoning.

Interactions were examined among age, relational complexity, and distraction. The interaction between age and distractor condition was reliable, $F(2, 65) = 3.15$, $p=.05$, whereas that between age and relational complexity was not, $F(2, 65) = .57$, $p= .57$. Importantly, the 3-way interaction was significant, $F(2, 65) = 3.28$, $p< .05$.

The pattern of interaction was investigated using repeated-measures ANOVAs for each age group separately. Results show that for the youngest children, ages 3-4, there was a main effect of relational complexity, $F(1, 21) = 4.44$, $p< .05$, a main effect of distractor, $F(1, 21) = 14.08$, $p<.01$, and a significant interaction between relational complexity and distraction, $F(1, 21) = 4.21$, $p=.05$. For the 6-7 yr old children there was a main effect of relational complexity, $F(1, 20) = 10.43$, $p< .01$ and of distraction, $F(1, 20) = 10.31$, $p< .01$, but no reliable interaction between these variables, $F(1, 20) = 2.71$, $p = .116$. Data for the 13-14 yr olds revealed a main effect of relational complexity, $F(1, 24) = 17.66$, $p< .001$ but not of distraction, $F(1, 24) = 2.21$, $p= .15$, nor was there a reliable interaction, $F(1, 24) = 1.67$, $p=.21$.

These data reveal that young children responded correctly well above chance on the one relation, no distractor condition; however, their accuracy fell when either a distractor or an added level of relational complexity (or both) was added. With age this pattern remained similar for 6-7 year olds, but as children reached adolescence, the negative effects of distractor and relational complexity were minimized.

Chance was calculated conservatively as the percent likelihood that a subject would select the correct relational match within the set of reasonable choices. These included relational errors and featural errors, but not extraneous objects. With this criteria, chance differed by condition reflecting the differential number of potential errors ranging from 50% (2 relevant possible answers) for 1R-N to 25% (4 relevant possible answers) for 2R-D. Paired *t*-tests revealed that the youngest children were above chance on all conditions (1R-N: $t(21) = 2.71$, $p< .05$; 2R-N: $t(21) = 2.43$, $p< .05$; 2R-D: $t(21) = 2.35$, $p< .05$) except for the 1R-D condition ($t(21) = 1.10$, $p=.29$).

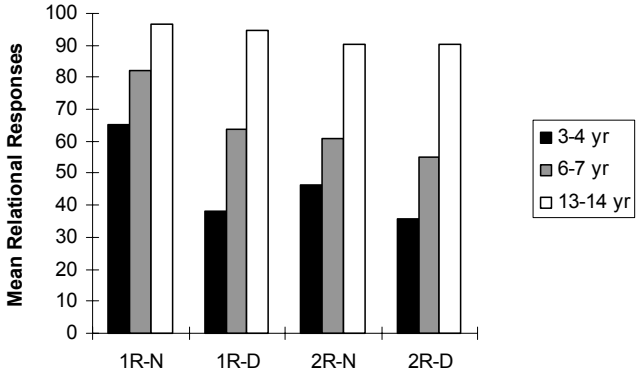


Figure 2. Proportion correct relational responses as a function of distraction and number of relations across age groups.

Error analysis

Children's responses were categorized into four types (see Table 1). Reponses were coded as either (1) relationally correct; (2) relational errors (an object in the correct relation but wrong role); (3) featural errors (the featural match in distractor conditions, or an unrelated object in the corresponding spatial location in no-distractor conditions); or (4) other errors. A repeated-measures ANOVA was performed to examine the relationship between age and participants' featural errors across the four picture conditions. Children's choice of

the featural match on the distractor conditions was compared with their choice of a non-featural, matched object in the same spatial location for the no distractor conditions. The main effect of age was reliable, $F(2, 65) = 49.78$, $p < .001$, as was the main effect of distractor, $F(1, 65) = 126.54$, $p < .001$, confirming that the featural match was an effective distractor. There was also a significant interaction between age and distractor, $F(2, 65) = 20.15$, $p < .001$, supporting the hypothesis that perceptual inhibition is a developmental constraint on analogical reasoning. No other interactions were reliable.

Table 1. Proportion of each response type across age and condition.

	Age	R1-N	R1-D	R2-N	R2-D
Correct Relational	3-4	65	38	46	36
	6-7	82	64	61	55
	13-14	97	95	90	90
Featural Errors	3-4	8	46	11	46
	6-7	0	25	4	27
	13-14	0	5	0	8
Relational Errors	3-4	15	9	6	9
	6-7	13	7	19	8
	13-14	2	1	4	4
Other Errors	3-4	9	4	21	6
	6-7	5	5	15	10
	13-14	2	0	4	5

A separate repeated-measures ANOVA was performed on relational errors. Note that there was one possible relational error choice in the R1 conditions, and two such possible error choices in the R2 conditions. The ANOVA revealed main effects of age, $F(2, 65) = 23.41$, $p < .001$, as well as relational complexity, $F(1, 65) = 59.56$, $p < .001$. There was also a significant interaction between the presence of a distractor and age on children's relational errors, $F(2, 65) = 5.85$, $p < .01$. At younger ages children made relational errors more frequently when there was no perceptual distractor available as an option. This finding suggests that young children unsure about the correct answer first attempted to make a feature-based selection; if no perceptually similar choice was available, then they made a guess among objects participating somehow in the relevant relation.

Discussion

Data from the RPA task at ages 3-4, 6-7 and 13-14 provide insight into the roles of relational knowledge, the relational shift, and maturational capacity in children's development of analogical reasoning. Patterns in participants' correct relational responses revealed main effects of age, distraction, and relational complexity, supporting the validity of the task manipulations. These main effects support theories of analogical reasoning development based on relational complexity and the relational shift.

Conversely, because the 3-4 year olds' performance on the 1R-N condition was high, their subsequent increases in errors in conditions with featural distraction or relational complexity provide support against the theory that domain knowledge alone is the mechanism underlying age-related development of children's analogical reasoning.

Interactions between age, distraction, and relational complexity indicate that in spite of children's capacity to perform analogical mapping based on these relations, as evidenced by their success on the R1-N condition, maturational factors may interact to constrain children's capacity to perform successfully on picture analogies that require more WM or perceptual inhibition.

Further, the error patterns suggest that perceptual distraction may be a primary constraint on children's reasoning and relational complexity a secondary constraint. Error analysis provided support for the claim that participants' patterns of failure were associated with age-related inhibition and relational complexity constraints. Participants were likely to make featural errors when the perceptual distractor was present, highlighting the validity of the distraction manipulation within the task. Supporting the relational shift hypothesis, at 3-4 yrs children were more likely to make featural responses when available than relational errors, even for the 2R-D condition, suggesting that inhibition was a more powerful constraint than relational complexity. However, relational errors were also made by children of all age groups, in highest numbers in the 2R-N condition, indicating that relational complexity is an important constraint on young children's analogical reasoning but may operate secondarily to featural distraction. One possible explanation for this is that inhibition is a core mechanism necessary for the WM system to operate on multiple relations (see Viskontas, in press)

The mechanism underlying featural distraction proposed by Rattermann and Gentner (1998; Gentner & Rattermann, 1991) is domain knowledge; however, this hypothesis is not supported by the current data, as the pictures were simple and counterbalanced across all conditions. The alternative explanation based on an inhibition mechanism is supported by the great difference between children's performance on the R1-N and R1-D conditions, as well as the similarity between the R1-D and R2-D conditions.

In sum, the RPA task provides a new paradigm for using children's interpretations of picture analogies to gather information about children's development of analogical reasoning, and specifically reveals interactions between the roles of perceptual inhibition/ distraction and relational complexity across age.

Acknowledgments

The authors wish to thank the Spencer Foundation (Dissertation Fellowship: Lindsey Richland), the National Institute of Mental Health (MH-64244-01A1; Robert Morrison), Xunesis (www.xunesis.org; Robert Morrison) and the Institute of Education Science (R305H030141; Keith Holyoak) for their generous support. We also thank Ann Fink for drawing the pictures used in the RPA materials. The RPA task is available from the first author upon request.

.References

Bassok, M. (2001). Semantic alignments in mathematical word problems. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 401-433). Cambridge, MA: MIT Press.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20,* 493-523.

Frye, D., & Zelazo, P. D. (2003). The development of young children's action control and awareness. In J. Roessler & N. Eilan (Eds*.), Agency and self-awareness: Issues in philosophy and psychology*. Oxford: Oxford University Press.

Frye, D., Zelazo, P. D., Brooks, P. J., & Samuels, M. C. (1996). Inference and action in early causal reasoning. *Developmental Psychology, 32*, 120–131.

Frye, D., Zelazo, P. D., & Burack, J. A. (1998). Cognitive complexity and control: Implications for theory of mind in typical and atypical development. *Current Directions in Psychological Science, 7*, 116–121.

Gentner, D. (1977). If a tree had a knee, where would it be? Children's performance on simple spatial metaphors. *Papers and Reports on Child Language Development*, *13*, 157-164.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Inetrrelations in development* (pp. 225-277). New York: Cambridge University Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277-300.

Gentner, D., Holyoak,, K., & Kokinov, B. (2001). *The analogical mind: Perspectives from cognitive science.* Cambridge, MA: MIT Press.

Gentner, D., & Holyoak, K. (1997). Reasoning and learning by analogy: Introduction. *American Psychologist, 52,* 32-24.

Gick, M.L., & Holyoak, K. L. (1980). Analogical problem solving. *Cognitive Psychology, 15,* 306-355.

Gick, M.L. & Holyoak, K. L. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Golinkoff, R. M., Jacquet, R., Hirsh-Pasek, K., & Nandakumar, R. (1996). Lexical principles may underlie the learning of verbs. *Child Development*, 67, 3101-3119.

Golinkoff, R. M., Hirsh-Pasek, K., Mervis, C. B., Frawley, W., & Parillo, M. (1995). Lexical principles can be extended to the acquisition of verbs. In M. Tomasello & W. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 185-222). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holyoak, K. J., Junn, E. N., & Billman, D. (1984). Development of analogical problem-solving skill. *Child Development*, *55*, 2042-2055.

Holyoak, K., Novick, L., & Melz, E.R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol 2: Analogical connections* (pp. 113-180). Norwood, NJ: Ablex.

Hummel, J.E., & Holyoak, K.J. (1997). Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review, 104,* 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in Frontotemporal Lobar Degeneration. *Journal of Cognitive Neuroscience*, *16*, 260-271.

Piaget, J., Montangero, J., & Billeter, J. (1977). La formation des correlats. In J. Piaget (Ed.) *Recherches sur l'abstraction reflechissante I* (pp. 115-129). Paris: Presses Universitaires de France.

Ross, B. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629-639.

Viskontas, I.V., Morrison, R.G., Holyoak, K.J., Hummel, J.E., & Knowlton, B.J., (in press) Relational integration, inhibition and analogical reasoning in older adults. *Psychology and Aging.* Zelazo, P. D., & Frye, D. (1998). Cognitive complexity and control: The development of executive function. *Current Directions in Psychological Science*, 7, 121–126.

A Neural Model of Episodic and Semantic Spatiotemporal Memory

Gerard J. Rinkus (rinkus@comcast.net)
468 Waltham St.
Newton, MA USA

Abstract

A neural network model is proposed that forms sparse spatiotemporal memory traces of spatiotemporal events given single occurrences of the events. The traces are distributed in that each individual cell and synapse participates in numerous traces. This sharing of representational substrate provides the basis for similarity-based generalization and thus semantic memory. Simulation results are provided demonstrating that similar spatiotemporal patterns map to similar traces. The model achieves this property by measuring the degree of match, *G*, between the current input pattern on each time slice and the expected input given the preceding time slices (i.e., temporal context) and then adding an amount of noise, inversely proportional to *G*, to the process of choosing the internal representation for the current time slice. Thus, if *G* is small, indicating novelty, we add much noise and the resulting internal representation of the current input pattern has low overlap with any preexisting representations of time slices. If *G* is large, indicating a familiar event, we add very little noise resulting in reactivation of all or most of the preexisting representation of the input pattern.

Introduction

Any realistic cognitive model must exhibit both episodic and semantic memory. And, as emphasized by Ans, Rousset, French, & Musca (2002), it must demonstrate these properties for the spatiotemporal (or, sequential) pattern domain. Thus, the model must be able to recall, without significant interference, large numbers of spatiotemporal patterns, which we will call episodes, given only single presentations of those episodes. Furthermore, it must exhibit human-like similarity-based generalization and categorization properties that underlie many of those phenomena classed as semantic memory.

We propose a sparse, distributed neural network model, TESMECOR (Temporal Episodic and Semantic Memory using Combinatorial Representations), that performs single-trial learning of episodes. The degree of overlap between its distributed memory traces increases with the similarity of the episodes that they represent. This latter property provides a basis for generalization and categorization and thus, semantic memory. The model achieves this property by computing, on each time slice, the similarity, *G*, between the expected and actual input patterns and then adding an amount of noise inversely proportional to *G* into the process of choosing an internal representation (IR) for that time slice. When expected and actual inputs match completely, no noise is added, allowing those IR cells having maximal input via previously modified weights to be reactivated (i.e., fully deterministic recall). When they completely mismatch, enough noise is added to completely drown out the learned, deterministic inputs, resulting in activation of an IR having little overlap with preexisting traces.

The opposing purposes of episodic memory and pattern recognition (i.e., semantic memory)—i.e., remembering what is unique about individual instances vs. learning the similarities between instances—has led other researchers to propose that the brain uses two complementary systems. McClelland et al (1995) and O'Reilly & Rudy (1999) propose that the purpose of the hippocampus is to rapidly learn new specific information whereas the purpose of neocortex is to slowly integrate information across individual instances thus coming to reflect the higher-order statistics of the environment. The hippocampus then repeatedly presents its newly acquired memory traces to neocortex, acting as trainer facilitating the gradual transfer of information to neocortex during the period of memory consolidation. We point out that TESMECOR is not such a two-component model. Rather, it is a monolithic model, i.e., it has a single local circuit architecture and processing algorithm (envisioned as an analog of the cortical mini-column) that satisfies the opposing needs.

Episodic Spatiotemporal Memory

Rinkus (1995) introduced a neural network model, TEMECOR, of episodic memory for spatiotemporal patterns. As shown in Figure 1, the model's. Layer 1 (L1) consists of binary feature detectors and its layer 2 (L2) consists of competitive modules (CMs). The L2 cells are nearly completely connected via a horizontal matrix (H-matrix) of binary weights.

The model operates in the following way. On each time step, a pattern is presented to L1. On that same time step, one L2 cell is chosen at random to become active in each CM corresponding to an active L1 cell. In addition, the horizontal weights from the L2 cells active on the prior time slice to those that become active on the current time are increased to their maximal value of one. In this way, spatiotemporal memory traces are embedded in the H-matrix. Later on, if we reinstate a set of L2 cells that was coactive in the past while learning an episode, the remainder of that episode will be read out in time. That is, the model recalls spatiotemporal memories.

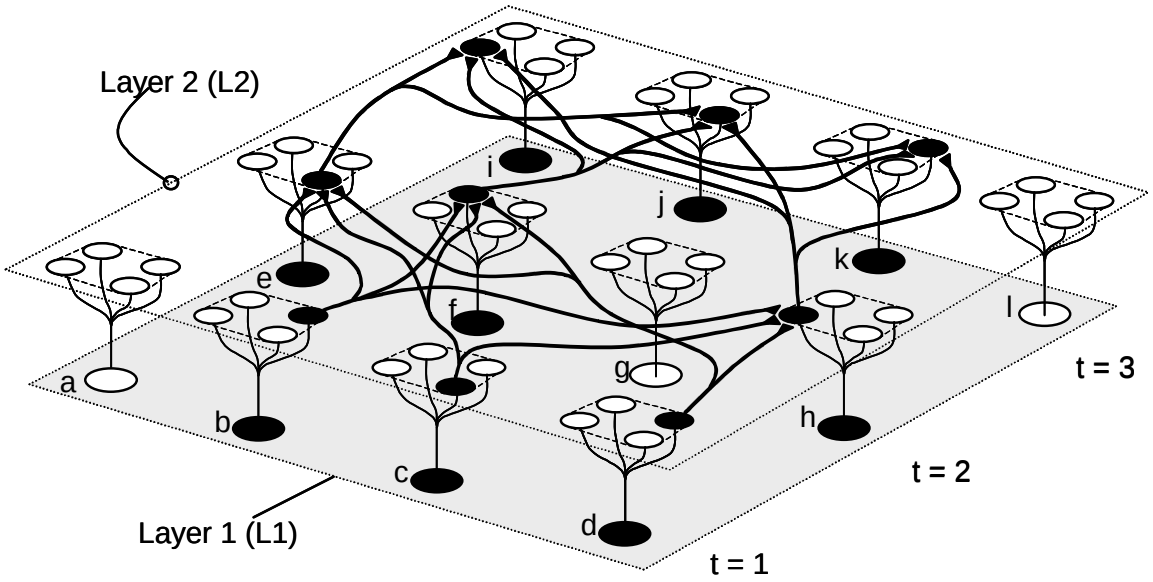


Figure 1: TEMECOR architecture showing how spatiotemporal memory traces are laid down amongst the horizontal connections of Layer 2. Features {b,c,d} are active at t = 1, {e,f,h} at t = 2, and {i,j,k} at t = 3. Each L2 cell has horizontal connections to all other L2 cells except those in its own CM. Only the connections increased while processing this particular spatiotemporal pattern (episode) are shown. Note that although this figure shows each time slice of the episode being handled by a separate portion of the network, this is purely to keep the figure uncluttered. In fact, all L1 cells and all L2 CMs are eligible to become active on every time slice.

TEMECOR exhibits high capacity, as shown in Figure 2, as well as other essential properties of episodic memory, e.g., single-trial learning. The model's beneficial properties derive principally from its use of a sparse distributed, or *combinatorial*, representational framework, a framework underlying many other models—Willshaw, Buneman & Longuet-Higgins, 1969; Lynch, 1986; Palm, 1980; Moll & Miikkulainen, 1995; Coultrip & Granger, 1994. The key to its high capacity is that by randomly choosing winners in the CMs, it minimizes the average overlap amongst the memory traces.

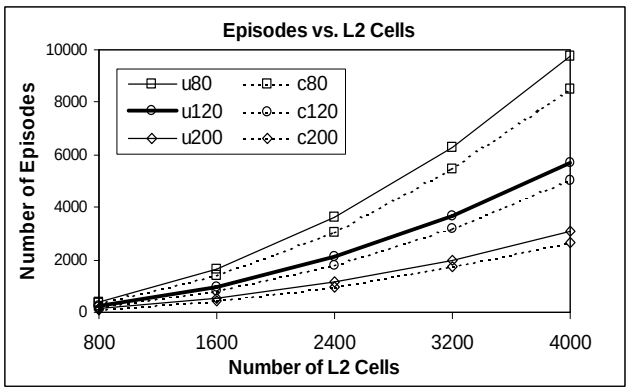


Figure 2: Capacity Results

Table 1 provides the data for the bold curve in the figure. It gives the maximal capacity, E, and other statistics for networks of increasing size, L. All episodes had T = 6 time slices and each time slice had S = 20 (out of M = 100) active features, chosen at random. The bottom row of the table shows that a network containing 4000 L2 cells, i.e., 100 CMs having K = 40 cells each, can store 5693 such episodes.

Table 1: Capacity Test Results

E	E/L	F	K	L	V	R_{set}	H
237	0.30	285	8	800	36	96.3	52.3
943	0.59	1132	16	1600	71	97.0	52.1
2104	0.88	2524	24	2400	105	97.0	51.8
3691	1.15	4430	32	3200	138	97.2	51.4
5693	1.42	6831	40	4000	171	97.4	50.9

Table 1 was generated as follows. For each K, the maximal number of episodes, E, which could be stored to criterion average recall accuracy, 96.3%, was determined. Recall accuracy, R_e, for a given episode e, is defined as:

$$R_e = (C_e - D_e)/(C_e + I_e) \quad (1)$$

where C_e is the number of L2 cells correctly active during recall of e^{th} episode, D_e is the number of deleted L2 cells, and I_e is the number of intruding L2 cells. The table reports R_{set}, the average of the R_e values for a whole set of episodes. All episodes were presented only once.

The other statistics in Table 1 are as follows. E/L is the ratio of stored episodes to the number of cells in L2, which increases linearly. F is the average number of instances of each feature across the entire set of episodes. V is the average number of times each L2 cell in a given CM became active to represent the corresponding feature. H is the percentage of weights increased, which is nearly

constant, at just over 50%, across rows. As we allow the fraction of weights to increase beyond 50%, more episodes are stored, but with a lower average recall accuracy due to the increase in intrusion errors resulting from saturation of the weights.

While TEMECOR exhibited the major properties of episodic memory it was not initially intended to model semantic memory and, due to its completely random method of choosing sparse internal representations at L2, it did not exhibit the generalization and categorization properties that underlie semantic memory. The successor version of the model, TESMECOR was developed to address this shortcoming (Rinkus, 1996).

Semantic Spatiotemporal Memory

TESMECOR is shown in Figure 3. It has some architectural differences with the original version (essentially, relaxations of some of the original's structural constraints) and a greatly modified winner selection process. The H-matrix of L2 is as before but the vertical projection is generalized. There is no longer a 1-to-1 correspondence between L1 cells and L2 CMs. Rather; each L1 cell connects to a fraction of all the L2 cells chosen at random in simulations. In TESMECOR, all CMs are active on every time slice. In addition, the bottom-up, or forward, connections (F-weights) and the top-down, or reverse, connections (R-weights) are now modeled separately and are modifiable.

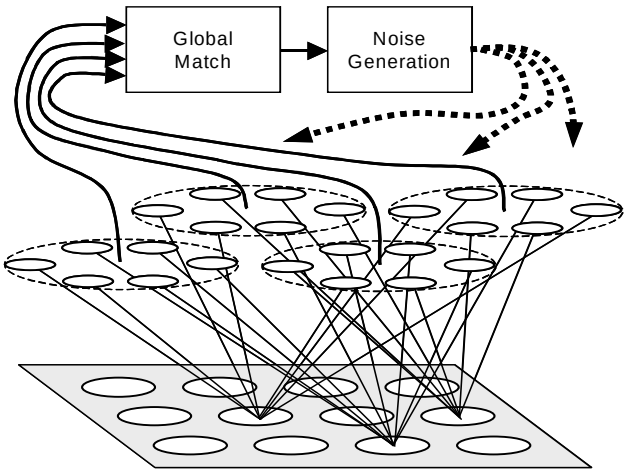


Figure 3: TESMECOR architecture.

The most significant change between TEMECOR and TESMECOR however is in the processing algorithm. Specifically, TESMECOR adds circuitry implementing spatiotemporal matching operations, both locally within each CM and globally over the entire L2. On each time slice, the global degree of match between the actual current input and the expected input, given the spatiotemporal context of the current input, modulates the amount of noise injected into the process of selecting which L2 cells will become active. The smaller the match, the more noise that is added and the greater the difference between the internal representation (IR) that would have become active purely on the basis of the deterministic inputs reflecting prior learning and the IR that actually does become active. The greater the match, the less noise added and the smaller the difference between the most highly implicated IR (on the basis of prior learning) and the actually chosen IR.

Figure 4 illustrates the basic principles by which the model computes, on each time slice, the degree of match, *G*, between its expected input and the actual input and then uses *G* to determine how much noise to add to the internal representation selection scheme. Figure 4a shows a pattern, A, presenting at t = 1. The H-weights are increased (represented by the dotted lines) from the active L1 cells onto an internal representation, IR_A, comprised of the three L2 cells that emerge as winners in their respective CMs. For purposes of this example, these three winners can be assumed to be chosen at random.

Figure 4b shows another pattern, B, presenting at t = 2. As with IR_A, IR_B can be assumed to be chosen at random. Here, we see the both H- and F-weights being increased.

Figure 4c shows another trial with pattern A presenting at t = 1. This time, IR_A becomes active due to the deterministic effects of the previously increased weights (which are now shown as solid lines). The cells of IR_A now send out signals via the H-matrix which will arrive at the other CMs at t = 2.

At this point, it is convenient to portray the t = 2 time slice in two steps. Figures 4d and 4e show these two steps. Figure 4d shows the signals arriving via the H-matrix at the same time that that signals arrive via the F-matrix from currently active L1 cells. Thus, the L2 cells in the three CMs on the right simultaneously receive two vectors each carrying possibly different expectations about which IR should become active (or equivalently, different hypotheses about what the current state of the world is). It is these two vectors that TESMECOR compares. In this particular case, the three cells of IR_B are receiving full support via the H-matrix. In other words, the temporal context says that IR_B should become active. However, these cells are receiving only partial support (two out of four L1 cells) via the F-matrix. Indeed, this is a novel input, pattern C, which has presented. Thus, the current spatial context does not contain sufficient information (given this network's history of inputs) to clearly determine what IR should become active. We represent this less-than-maximal support for IR_B by the gray shading of its cells. Because of this mismatch, i.e., $G < 1.0$, we add some noise into the winner selection process. The final result is that a different L2 cell than the one most strongly implicated by the deterministic inputs ends up winning the competition in one of the three CMs (the bottom right-hand one) active at t = 2. Thus, Figure 4e shows a new IR, IR_C, representing the novel pattern, C.

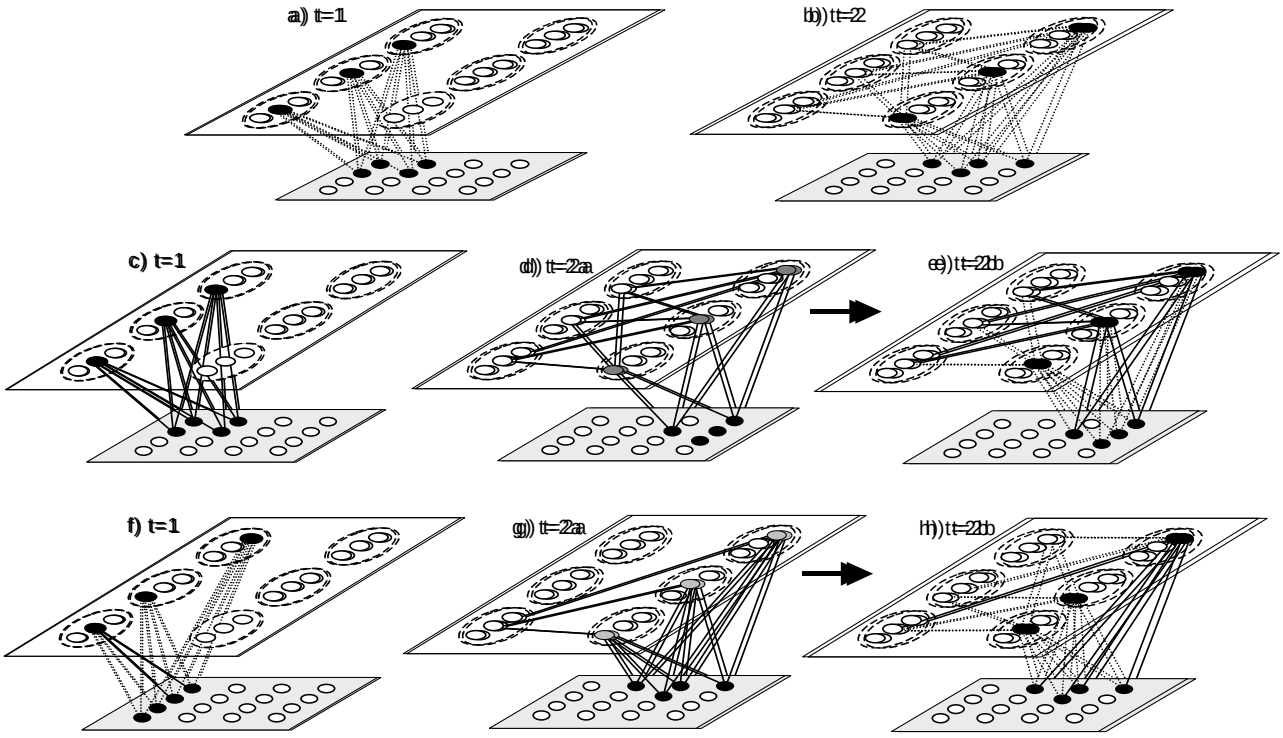


Figure 4: Sketch of TESMECOR's spatiotemporal pattern comparison and noise-modulated internal representation selection scheme. See text for explanation. As in Figure 2, the division of the L2 CMs into separate groups for different time slices is purely to avoid clutter. In the model's actual operation, all CMs are active on every time slice.

Figures, 4f, 4g, and 4h, show another possible scenario. This time, we will again present pattern B at t = 2. However a novel pattern, D, having only two features in common with A, presents at t = 1. As this is the first time slice of this new trial, there is no prior context vector active in the H-matrix. For concreteness, let's assume that this degree of mismatch causes a new winner to be chosen in two of the three CMs active at t = 1, resulting in a new IR, IR_D. When B presents at t = 2, the F-vector lends maximal support for IR_B but the H-vector has great uncertainty; only 1/3 of the maximal possible horizontal input arrives at the cells of IR_B. This seems like even a worse match than in Figure 4d (shown by an even lighter shading of the IR_B cells than in Figure 4d). Consequently, more noise is added to the winner selection process. Let's assume that this degree of mismatch leads to a new winner in two of the three CMs active at t = 2, resulting in a new IR, IR_{B*}, for pattern B.

With this example of the desired behavior in mind, we now give TESMECOR's processing algorithm, which is computed on each time slice for each L2 cell.

1. $$\psi_{i,t} = \sum_{j \in \Gamma_t} w_{ji} \tag{2}$$

2. $$\Psi_{i,t} = \frac{\psi_{i,t}}{\max(\max_{j \in CM}(\psi_{j,t}), {}^{F}\Theta_t)} \tag{3}$$

3. $$\phi_{i,t} = \sum_{j \in \Delta_{t-1}} w_{ji} \quad , t > 0 \tag{4}$$

4. $$\Phi_{i,t} = \frac{\phi_{i,t}}{\max(\max_{j \in CM}(\phi_{j,t}), {}^{H}\Theta_t)} \quad , t > 0 \tag{5}$$

5. $$\chi_{i,t} = \begin{cases} \Psi_{i,t}{}^{u} \Phi_{i,t}{}^{v} & , t > 0 \\ \Psi_{i,t}{}^{w} & , t = 0 \end{cases} \tag{6}$$

6. $$\mathrm{X}_{i,t} = \frac{\chi_{i,t}}{\max(\max_{j \in CM}(\chi_{j,t}), {}^{\chi}\Theta)} \tag{7}$$

7. $$\pi_{k,t} = \max_{j \in CM_k} \mathrm{X}_{j,t} \quad , 1 \le k \le Q \tag{8}$$

8. $$G_t = \sum_{k=1}^{Q} \pi_{k,t} \Big/ Q \tag{9}$$

9. $$p_{i,t} = \frac{f(\mathrm{X}_{i,t}, G_t)}{\sum_{j \in CM} f(\mathrm{X}_{i,t}, G_t)} \tag{10}$$

In step 1, each L2 cell, *i*, computes its total weighted input, $\psi_{i,t}$, from the set, Γ_t, of currently active L1 cells. In step 2, the ψ values are normalized within each CM. That is, we find the maximum ψ value, in each CM and divide all the individual values by the greater of that value and

F-matrix threshold, ${}^{F}\Theta_t$. ${}^{F}\Theta_t$ is needed to ensure that small feedforward signals are not amplified in subsequent normalization steps. ${}^{F}\Theta_t$ is a parameter that can vary from one time slice to the next but we omit discussion of this detail in this paper due to space limitations.

Steps 3 and 4 perform analogous operations for the horizontal inputs. In step 3, *i*, computes its total weighted input, $\phi_{i,t}$, from the set, Δ_{t-1}, of L2 cells active on the prior time slice. In step 4, the ϕ values are normalized within each CM. That is, we find the maximum ϕ value, in each CM and divide all the individual values by the greater of that value and an H-matrix threshold, ${}^{H}\Theta_t$. ${}^{H}\Theta_t$ is needed to ensure that small H values are not amplified in subsequent normalization steps. ${}^{H}\Theta_t$ also varies from one time slice to the next but again, space limitations force us to omit discussion of this detail. Note that steps 3 and 4 are only applicable on non-initial time slices (t > 0) of episodes.

Step 5 works differently on the first time slices of episodes than on the rest. When t > 0, we multiply the two pieces of evidence, $\Psi_{i,t}$ and $\Phi_{i,t}$, that cell *i* should become active but we do this after passing them through separate exponential filters. Since $\Psi_{i,t}$ and $\Phi_{i,t}$, are both between 0 and 1, the final $\chi_{i,t}$ values output from this step are also between 0 and 1. The exponential filters effect a generalization gradient: the higher the exponents, *u* and *v*, the sharper the gradient and the more sensitive the model is to differences between inputs (i.e., the finer the spatiotemporal categories it would form) and the less overlap between the internal representations chosen by the model. When t = 0, we do not have two vectors to compare. Instead, we simply pass the Ψ values through an exponential gradient-controlling filter. The three different exponent parameters, *u*, *v*, and *w*, simply let us fine-tune the model's generalization gradients. For example, we might want the model's sensitivity to featural similarity to be stricter at the beginning of episodes than on the successive time slices of episodes; thus we would set *w* higher than *u*.

In step 6, we normalize the combined evidence vector, again subject to a threshold parameter, ${}^{\chi}\Theta_t$, that prevents small values from erroneously being amplified. In step 7, we simply determine the maximum value, $\pi_{i,t}$, of the $X_{i,t}$ values in each CM. These π values constitute local, i.e., within each CM, comparisons between the model's expected and actual inputs. In step 8, we compute the average of these local comparison results across the *Q* CMs of L2, resulting in the model's global comparison, G_t, of its expected and actual inputs.

In step 9, we convert the $X_{i,t}$ values back into a probability distribution whose shape depends on G_t. We want to achieve the following: if G_t is 1.0, indicating that the actual input has perfectly matched the model's expected input, then, in each CM, we want to choose, with probability 1.0, the cell belonging to the IR representing that expected input. That cell, in each CM, is the one having the highest X value. Since, in general, other cells in that cell's CM could have non-zero or even high X values, we need to filter the values by an expansive nonlinearity, *f*, so that the cell with the maximal X value maps to a probability, $p_{i,t}$, of 1.0 and the rest of the cells end up mapping to $p_{i,t} = 0.0$. On the other hand, if $G_t = 0$, indicating that the actual input is completely unexpected in the current temporal context given all of the model's past experience, then we want to make all the cells, in any given CM, be equally likely to be chosen winner. Thus, in this case, *f* should be a compressive nonlinearity that maps all cells in the CM to $p = 1/K$, where *K* is the number of cells in the CM. Without going into details, the function, *f*, is a sigmoid that meets the above goals. In the last stage of step 9, we simply choose the winner in each CM according to the resulting distribution.

To summarize, on each time slice, every L2 cell compares two evidence vectors, the H-vector, reflecting the sequence of patterns leading up to the present time slice (temporal context), and the F-vector, reflecting the current spatial pattern (spatial context). These vectors are separately nonlinearly filtered and then multiplicatively combined. The combined evidence vector is then renormalized and nonlinearly filtered before being turned into a probability distribution that governs the final selection of L2 cells to become active. Note that this basic scheme can be extended to simultaneously compare other evidence vectors as well. This is one of our intended lines of future research: specifically, we will examine incorporating a hippocampal component to the model, which will provide a third evidence vector to the L2 cells.

The concept of controlling the embedding of internal representations (IRs) based on comparing the expected and actual inputs is common to other cognitive models, e.g., Grossberg (1987). However, TESMECOR's use of distributed IRs, rather than singleton IRs, requires a generalized comparison scheme. Specifically, with distributed IRs, there exists a range of possible degrees of overlap between IRs. We want to use that range to represent the spatiotemporal similarity structure of the environment to which the model has been exposed. Therefore, rather than having a single threshold for judging the similarity of the current input and expected inputs (e.g., ART's vigilance parameter), TESMECOR's continuous-valued similarity measure, *G*, is used to inject a variable amount of noise into the IR-selection process, which in turn allows for selecting IRs whose degrees of overlap are correlated with their spatiotemporal similarities.

Simulation Results

In this section, we provide the results of preliminary investigations of the model demonstrating that it performs similarity-based generalization and categorization in the spatiotemporal pattern domain.

The four simulations described in Table 2 were performed as follows. In the learning phase, E episodes were presented, once each. Each episode consisted of 5 time slices, each having 20 (out of 100) randomly selected features present. Then, perturbed versions, differing by *d* = 2, 4, 6, or 8 (out of 20) features per time slice from the original episodes

were generated. The model was then tested by presenting the first Z time slices of the perturbed episodes as prompts. Following the prompt time slices, the model entered a *free-running* mode (i.e. cutting off any further input) and processing continued from that point merely on the basis of signals propagating in the H-projection.

Table 2: Generalization/Categorization Results

Simulation	**E**	***d***	**Z**	R_{set}
1	27	2	1	92.3%
2	13	4	1	98.0%
3	7	6	1	98.3%
4	13	8	2	82.7%

These results indicate that the model was extremely good at locking into the trace corresponding to the most-closely-matching original episode. The accuracy measure, R_{set} (eq. 1) measures how close the recall L2 trace is to the L2 trace of the most-closely-matching original episode. The accuracy for simulation 4 (82.7%) may seem low. However, if the accuracy measure is taken only for the final time slice of each episode then it is close to 100% for all four simulations. The view taken herein is that given that the pattern to be recalled are spatiotemporal, the most relevant measure of performance is the measure of accuracy on the last time slice of the test episode. If the model can "lock into" the correct memory trace by the end of the recalled trace, then that should be sufficient evidence that model has recognized the input as an instance of a familiar episode.

Table 3: Per-Time-Slice L2 Accuracy for the Test Trials of Simulation 4 of Table 2

Episode	**T=1**	**T=2**	**T=3**	**T=4**	**T=5**
1	0.9	0.9	1.0	1.0	1.0
2	0.82	1.0	1.0	1.0	1.0
3	0.67	1.0	1.0	1.0	1.0
4	0.82	0.9	1.0	1.0	1.0
5	0.67	0.82	1.0	1.0	1.0
6	0.67	0.9	1.0	1.0	1.0
7	0.9	1.0	1.0	1.0	1.0
8	0.74	1.0	1.0	1.0	1.0
9	0.74	1.0	1.0	1.0	1.0
10	0.67	0.82	1.0	1.0	1.0
11	0.54	0.67	0.22	0.0	0.0
12	0.48	0.21	0.0	0.0	0.0
13	0.82	0.9	1.0	1.0	1.0

Table 3 shows the details of the simulation 4 in Table 2. Specifically, it shows the L2 accuracy on each time slice of each episode during the recall test. For each recall trial the model received a prompt consisting of degraded versions of the first two time slices of the original episode—specifically, 4 out of 20 features were substituted on each time slice (for a total of 8 featural differences). In all but two cases, the model 'locks into' the L2 trace corresponding to the most-closely-matching original episode (i.e., the episode from which the degraded prompt was created.

These simulations provide preliminary evidence that TESMECOR exhibits generalization, and in fact categorization, in the spatiotemporal domain, while at the same time exhibiting episodic memory since the episodes are learned with single trials.

Acknowledgments

Thanks to Daniel H. Bullock, Michael E. Hasselmo, Michael A. Cohen and Frank Guenther, for discussions that provided many helpful ideas during the formulation of this theory and preparation of the doctoral dissertation from which this work was derived.

References

Ans, B., Rousset, S., French, R.M., & Musca, S. (2002) Preventing Catastrophic Interference in Multiple-Sequence Learning Using Coupled Reverberating Elman Networks. *Proc. of the 24th Annual Conf. of the Cognitive Science Society*. LEA, NJ.

Carpenter, G. & Grossberg, S. (1987) Massively parallel architectures for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*. **37**, 54-115.

Coultrip, R. L. & Granger, R. H. (1994) Sparse random networks with LTP learning rules approximate Bayes classifiers via Parzen's method. *Neural Networks*, **7**(3), 463-476.

Lynch, G. (1986) Synapses, Circuits, and the Beginnings of Memory. The MIT Press, Cambridge, MA.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419-457.

Moll, M. & Miikkulainen, R. (1995) Convergence-Zone Episodic Memory: Analysis and Simulations. Tech. Report AI95-227. The University of Texas at Austin, Dept. of Computer Sciences.

O'Reilly, R. C. & Rudy, J. W. (1999) Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function. TR 99-01. Institute of Cognitive Science, U. of Colorado, Boulder, CO

Palm, G. (1980) On Associative Memory. *Biological Cybernetics*, **36**. 19-31.

Rinkus, G. J. (1995) TEMECOR: An Associative, Spatiotemporal Pattern Memory for Complex State Sequences. *Proc. of the 1995 World Congress on Neural Networks*. LEA and INNS Press. 442-448.

Rinkus, G. J. (1996) A Combinatorial Neural Network Exhibiting both Episodic Memory and Generalization for Spatio-Temporal Patterns. Ph.D. Thesis, Graduate School of Arts and Sciences, Boston University.

Willshaw, D., Buneman, O., & Longuet-Higgins, H. (1969) Non-holographic associative memory. *Nature*, **222**, 960-962

Promoting Flexible Problem Solving: The Effects of Direct Instruction and Self-Explaining

Bethany Rittle-Johnson (bethany.rittle-johnson@vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University
230 Appleton Place, Peabody #512, Nashville, TN 37203

Abstract

How do people learn flexible problem-solving knowledge, rather than inert knowledge that is not applied to novel problems? Both the source of the knowledge – instructed or invented – and a central learning process – engaging in self-explanation – may influence the development of problem-solving flexibility. Seventy-seven third- through fifth-grade students learned about mathematical equivalence under one of four conditions that varied on two dimensions: 1) prompts to self-explain and 2) invention vs. instruction on a procedure. Both self-explaining and direct instruction helped students to learn a correct problem solving procedure. Self-explanation promoted transfer, whereas direct instruction had both positive and negative effects on transfer. Overall, self-explanation is an important learning mechanism underling the acquisition of flexible problem solving with or without direct instruction.

Introduction

Everyday, we are faced with new problems to solve. How do we complete our income tax return, write a new resume, or find a new route home given recent road construction? When faced with a problem repeatedly, we often develop procedures for solving the problem, i.e. step-by-step methods for solving the problem. Ideally, we learn flexible, relatively abstract procedures that we can appropriately apply to a variety of tasks so that we do not need to invent new procedures when task conditions shift. Flexible, abstract, knowledge is also a key characteristic of expertise (Chi, Feltovich, & Glaser, 1981). Thus, understanding how people develop flexible, abstract knowledge is crucial for understanding learning and development and for designing learning environments to support flexibility.

Unfortunately, people of all ages and across a large range of domains often gain inert knowledge instead – knowledge that is not applied to new situations (see Bransford, Brown, & Cocking, 2001 for a review). For example, physics students typically fail to use knowledge of physics principles, such as Newton's Laws, to solve everyday problems (Halloun & Hestenes, 1985). Indeed, even scientists sometimes fail to use their scientific knowledge to solve mundane tasks (Lewis & Linn, 1994).

How do people learn flexible knowledge, rather than simply gaining inert knowledge, and how can we support this learning? In the current study, two processes were evaluated: 1) The source of new knowledge – invention or direct instruction and 2) A potential mechanism underlying flexible learning - generating self-explanations for why and how things work.

Invention vs. Instruction

Where do new procedures come from? Typically, we invent a procedure through problem exploration or we learn a procedure from others (e.g. via imitation or direct instruction). Major theories of learning and philosophies of education differ in their emphasis on the sources of new procedures. The current paper focuses on one source of knowledge from other people – direct instruction – and compares it to inventing procedures on ones own.

Invention and learning from direct instruction can both lead to learning of the target behavior or knowledge (e.g. Judd, 1908). However, a major concern with discovery learning is that a substantial proportion of learners never invent a correct procedure or engage in correct ways of thinking (Mayer, 2004).

Another critical issue is the relative effectiveness of each source of knowledge for supporting flexible, generalizable knowledge. Direct instruction on a procedure can lead people to learn the procedure by rote, to make nonsensical errors and to be unable to transfer the procedure to solve novel problems (e.g. Brown & Burton, 1978; Hiebert & Wearne, 1986), whereas when people invent procedures, they often use the procedures flexibly in new situations (Hiebert & Wearne, 1996). Thus, there appears to be a trade-off between instruction improving problem solving on a restricted range of problems but potentially harming flexible problem solving on a broader range of problems. The current study evaluates the pros and cons of direct instruction versus encouragement to invent a procedure on a single task and evaluates the role of self-explaining as a learning mechanism under both conditions.

Self-Explaining

A potential mechanism underlying the impact of instruction and invention on procedural flexibility (and learning more generally) is learners' attempts to generate explanations for why and how things work. Successful learners typically generated explanations while studying worked-examples to problems. These explanations included identification of gaps in understanding and linkages to previous examples or sections in the text (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). Subsequent research indicates that learners ranging from 5-years-old to adulthood in domains ranging from number conservation to computer programming can learn more if they are prompted to generate self-explanations (Aleven & Koedinger, 2002; Bielaczyc, Pirolli, & Brown, 1995; Chi, de Leeuw, Chiu, & LaVancher, 1994). These findings are cooborated by findings from classroom-based research on individual differences and on cross-cultural

differences in teaching practices (Stigler & Hiebert, 1997; Webb, 1991). In all cases, generating explanations is associated with greater learning. Thus, generating explanations to explain how or why things work is a critical learning process across the lifespan and across domains.

However, the informal theories or working hypotheses that learners develop are not always correct. Rather, learners often develop incorrect theories, and engaging these incorrect theories is critical to supporting learning (Bransford et al., 2001). Evidence from a variety of domains indicates that learners' incorrect informal theories are resistant to change and often persist after formal instruction that contradicts their theories (e.g. Halloun & Hestenes, 1985).

Prompting students to generate explanations for incorrect, as well as correct, solutions, beliefs, etc, may be one method for helping learners overcome incorrect prior knowledge. For example, Siegler (2002) found that prompting students to explain both correct and incorrect solutions led to greater procedural flexibility than only explaining correct solutions. In the current study, learners in the self-explanation condition were prompted to explain both correct and incorrect solutions to maximize the effectiveness of the explanation condition.

Prompting students to self-explain has been used in conjunction with a variety of sources of new information (reading a text, studying worked example, problem-solving with feedback), but has not been used in combination with direct instruction on a correct procedure nor has prior research directly compared the role of self-explaining under different sources of new knowledge. Prompting students to engage in an effective learning process after direct instruction may help students to understand and generalize the procedure. Similarly, prompts to self-explain may help students to invent and generalize correct procedures when they do not receive direct instruction (Siegler, 2002).

The current study

These issues were evaluated in the context of children learning to solve problems that tap the idea of mathematical equivalence. Mathematical equivalence is a fundamental concept in both arithmetic and algebra. Unfortunately, most children in elementary and middle school do not seem to understand equivalence, and this poses a major stumbling block for students' success in algebra (Kieran, 1981). Novel problems such as 3+4+5=3+__ challenge students' naïve understanding of equivalence in a familiar arithmetic context, and approximately 70% of fourth- and fifth-graders do not solve these problems correctly (Alibali, 1999; Rittle-Johnson & Alibali, 1999). In the current study, third through fifth graders learned to solve these mathematical equivalence problems under one of four conditions based on two factors: 1) direct instruction on a correct procedure vs. prompts to invent a new way to solve the problems and 2) prompts for self-explanations vs. no prompts.

Method

Participants. Initial participants were 121 third- through fifth- grade students from an urban, parochial school serving a working- to middle-class population. In line with previous studies using mathematical equivalence problems, 34 students (29%) solved at least half of the mathematical equivalence problems correctly at pretest and thus were excluded from the study. One student was excluded because he did not take the pretest, and 9 students were excluded because they were absent on the day of the delayed posttest, so they could not be included in the repeated-measure analyses. The final sample consisted of 37 third-graders, 22 fourth-graders, and 18 fifth-graders.

Design. Students were randomly assigned to one of four conditions based on crossing two factors: 1) instruction on a correct procedure or prompts to invent a procedure and 2) prompts to self-explain correct and incorrect solutions or no prompts. There were 20 participants in the *instruction + explain* condition, 21 students each in the *invent + explain* and *instruction-only* conditions, and 15 students in the *invent-only* condition (unequal group sizes due to random differences in absenteeism at the delayed posttest).

Procedure. Students completed the pretest in their classrooms. Students who solved at least half of the mathematical equivalence problems incorrectly participated in a one-on-one intervention session. During the intervention session, there were 3 phases: warm-up, intervention (instruction problems and practice problems, all with accuracy feedback), and follow-up. At the end of this session, students completed the immediate paper-and-pencil posttest. Approximately 2 weeks later, students completed the delayed paper-and-pencil posttest in their classrooms.

Intervention session. All problems presented during this session were in standard format (see Table 1). At the beginning of the session, students solved two warm-up problems, explained how they had solved each problem, and were told whether they had solved the second problem correctly to motivate students to try to figure out correct ways to solve the problems. During the intervention phase, all students solved 8 problems. On all problems, students explained how they solved the problem and then were told if they had solved it correctly. The first two problems were the instructional problems and were both in the format A+B+C=A+_ (*A+ problems)*. For students in the instruction conditions, the experimenter explained a correct, add-subtract, procedure for solving the problem. For the problem 4+9+6 = 4+__, the experimenter said: "You can add the 4 and the 9 and the 6 together before the equal sign (gesture a "circle" around the 3 numbers), and then subtract the 4 that's over here, and that amount goes in the blank. So, try to solve the problem using this strategy." Students in the invention conditions were asked to try to figure out a new way to solve the problems.

Table 1: Procedural Knowledge Problem Types.

Problem Type	Problem Format
Standard	A+B+C=A+__ (A+) A+B+C=__+C (+C)
No repeated addend	A+B+C=D+__ A+B+C=__+D
Subtraction too	A+B-C=A+__ A+B-C=__ - C
Swap sides	A+__=A+B+C __+C=A+B+C

Next, students solved 6 practice problems that alternated between the two standard problem formats (see Table 1). After solving each of these problems, students were also told the correct answer to the problem. Students in the explain conditions were then prompted to try to explain a correct and an incorrect solution. They were shown the solution that two students at another school had gotten – one correct and one incorrect – and were asked to explain both how each student had gotten the answer and why each answer was correct or incorrect. The intervention trials were presented on a laptop computer that recorded accuracy and solution times. At the end of the intervention, students solved two follow-up problems without feedback and explained their solutions.

Assessments. The pretest, immediate posttest and delayed posttest were identical except that only a subset of the procedural knowledge problems were presented at pretest. The *procedural knowledge* assessments contained 4 types of problems, as shown in Table 1. Letters stand for numbers and indicate when a number was repeated within a problem. The standard problem formats were used during the intervention. One instance of each of the standard and no repeated addend problems was presented on the pretest. One instance of each of the problems in Table 1 were presented on the posttests. Students were encouraged to show their work when solving the problems. The 5 items on the *conceptual knowledge* assessment are shown in Table 2.

Table 2: Conceptual Knowledge Assessment Items

Item	Coding (2 pts)
Define equal sign	Mention "the same" or "equal" (2pts)
Rate definitions of equal sign: Rate 4 definitions as "always, sometimes or never true"	Rate "two amounts are the same" as "always true" (2 pts) or "sometimes true" (1 pt)
Group Symbols: Place symbols such as =, +, <, & 5 into three groups	Group =, >, and < together (2 pts)
Recognize use of equal sign in multiple contexts: Indicate whether 8 problems such as 8=2+6 and 3+2=6-1 make sense	7 or 8 correct (2 pts); 6 correct (1 pt)
Correct Encoding: Reproduce 4 equivalence problems from memory	Correctly reproduce problem (.5 point each)

Table 3: Procedures for Solving Equivalence Problems

Procedures	Sample student explanation
Correct Procedures	
Equalize	"I added 8 plus 7 plus 3 and I got 18 and 8 plus 10 is 18."
Add-subtract	"I did 8 plus 7 equals 15 plus 3 equals 18 and then 18 minus 8 equals 10"
Grouping	"I took out the 8's and I added 7+3."
Incorrect Procedures	
Add all	"I added 8+7+3+8, which is 26"
Add to equal sign	"8 plus 7equals15, plus 3 is18."
Incorrect Grouping	"I added 8 plus 7."

Coding. On the procedural knowledge assessments, students' percent correct was used (arithmetic slips were ignored). Students' verbal explanations during the intervention were used to code students' procedure use on those problems (see Table 3). On the conceptual knowledge assessment, each item was scored from 0-2 points for a possible total of 10 points (see Table 2).

Results

The effects of condition were assessed on three outcomes: procedural learning, procedural transfer, and conceptual knowledge. Procedural learning was assessed 3 times: verbally at the end of the intervention and on the immediate and delayed posttests. Procedural transfer and conceptual knowledge were assessed twice – on the immediate and delayed posttest. Results were evaluated for each outcome using repeated-measures ANOVAs with time of assessment as a within-subject factor and instruction vs. invention and prompts to explain (yes/no) as between subject factors. Pretest conceptual and procedural knowledge were included in all analyses as covariates.

Procedural Learning

First, consider procedural learning, which was assessed using problems identical in form to those presented during the intervention (see Table 1). As shown in Figure 1, generating explanations and, to some extent, receiving instruction led to greater accuracy on the learning problems. There was a main effect for explaining, F (1, 71) = 6.11, p = .02, η_p^2 = .08, and a marginal effect for instruction, F (1,71) = 2.84, p = .10, η_p^2 = .04, and no interaction between the two conditions and no effects of time of assessment.

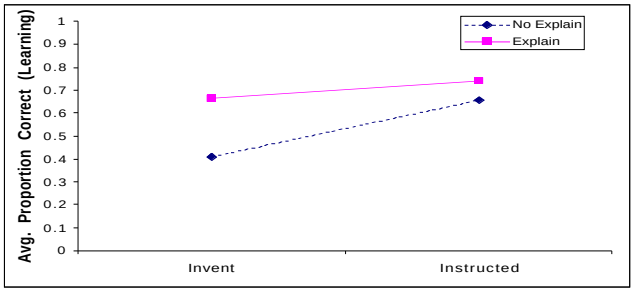


Figure 1: Effect of condition on learning problems

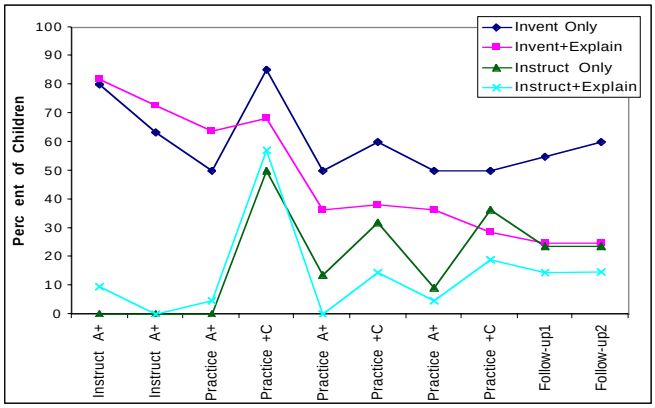


Figure 2: Incorrect procedures: Trial-by-trial use during the intervention and follow-up, by condition.

Inspection of students' procedure use during the intervention provided insights into learning pathways (see Table 3 for a description of each procedure). First, consider students' use of incorrect procedures. As shown in Figure 2, during the instruction phase (first two problems), students in the instruction condition quickly abandoned their incorrect procedures whereas most students in the invention conditions persisted in using incorrect procedures. During the practice phase, when students first encountered a problem in a different format (+C problem), there was a sharp return to using incorrect procedures. Students in the instruction condition continued to struggle with the +C problems, especially if not prompted to self-explain. Students in the invent-only condition struggled across problems – at least 50% continued to use incorrect procedures, whereas students who were prompted to self-explain steadily decreased in use of incorrect procedures.

Next, consider students' use of the correct instructed procedure – add-subtract (see Figure 3). Students in the instruction conditions quickly learned the add-subtract procedure and many students persisted in using this procedure across a majority of problems. However, a third of the students did not apply the procedure when the surface structure of the problem changed (+C problems), even though the procedure required only a very minor adaptation. Of students in the invent conditions, about 15-20% invented and used this procedure. Next, consider the other commonly used correct procedure – grouping (see Figure 4). Students in the invent conditions gradually increased their use of this procedure. Prompts to explain also helped

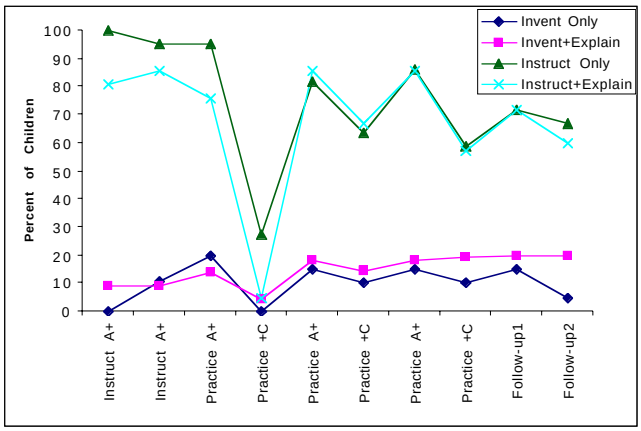


Figure 3: Use of Add-Subtract procedure trial-by-trial during the intervention and follow-up, by condition.

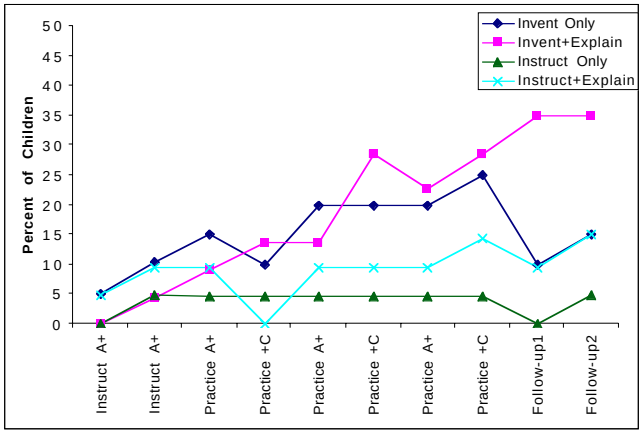


Figure 4: Use of Grouping procedure trial-by-trial during the intervention and follow-up, by condition.

students to invent and maintain use of the procedure on the follow-up problems. Finally, students in the invent conditions also used the equalizer procedure on 12% of intervention trials, whereas students in the instruct+explain condition used it on 6% of trials and students in the instruct-only condition used it on less than 1% of trials.

Overall, over half of students in the invent-only group did not learn a correct procedure through feedback alone. Less than a quarter of students in the other conditions had similar difficulty. Rather, direct instruction quickly led children to adopt a correct procedure, and prompts to explain helped students to invent new procedures.

Procedural Transfer

Next consider students' ability to transfer their procedures to novel problems (see Figure 5). Overall, there was a main effect of explaining $F(1, 71) = 3.93$, $p = .05$, $\eta_p^2 = .05$ and no overall effect of instruction, interaction between the two, or effect of test time. Prompts to explain supported transfer of procedures to novel problems, but instruction vs. invention did not have a general effect on transfer (although see below for an important caveat). Inspection of success on individual problems suggested that the impact of

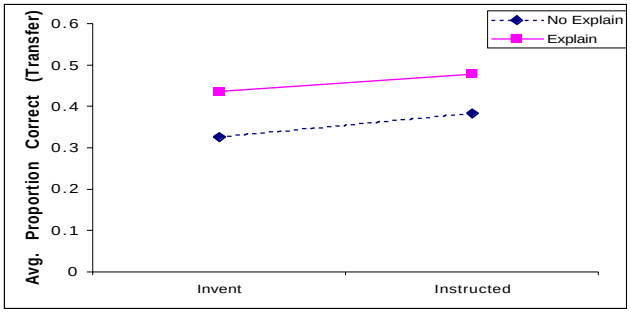


Figure 5: Effect of condition on transfer performance

condition varied by problem type. To evaluate this, a second repeated-measures ANOVA was conducted with transfer problem type (no repeated addend, including subtraction, swapping sides; see Table 1) as an additional within-subject factor. Indeed, there was an interaction between problem type and instruction, F (2, 142) = 9.37, p < .001, η_p^2 = .12, but no interaction of explaining with problem type. Follow-up analyses indicated that instruction improved performance on problems without a repeated addend, F (1, 71) = 7.89, p = .006, η_p^2 = .10. The instructed procedure did not need to be modified to solve problems without a repeated addend. Instruction had no reliable effect on the other problem types. However, focusing on the most difficult individual problem (A+B-C=_-C), receiving instruction actually harmed performance, regardless of explaining, F (1,71) = 13.12, p = .001, η_p^2 = .16. For example, on the delayed posttest, very few of the students who received instruction solved the problem correctly, whereas at least a third of students in the invent conditions solved the problem correctly (see Figure 6). This may be because the invented grouping procedure is easier to apply to this problem than add-subtract.

Conceptual Improvement

Finally, there was no effect of condition on gains in conceptual knowledge. Although students as a whole made small gains in conceptual knowledge from pretest to immediate posttest (m = 2.9 vs. 3.1 out of 10), t(76) = 2.0, p =.05, and made even great gains after a delay (m = 3.8), t(76) = 5.7, p < .001, the amount of gain did not vary by condition.

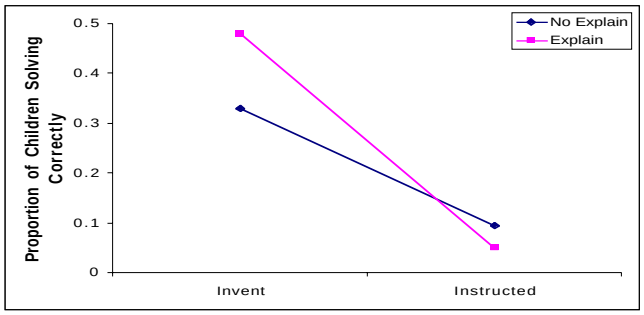


Figure 6: Effect of condition on hardest transfer problem, A+B-C=_-C, at delayed posttest

Discussion

Self-explanation is a critical learning mechanism that leads to greater procedural flexibility. The current findings converge with past findings that better learners spontaneously produce self-explanations and that prompting learners to generate explanations leads to greater learning (Aleven & Koedinger, 2002; Bielaczyc et al., 1995; Chi et al., 1989; Chi et al., 1994). The current findings expand past research by demonstrating that self-explanation is an important learning mechanism regardless of instruction. For students who received direct instruction on a correct procedure, prompts to self-explain had little influence on the use of the instructed procedure. Rather, the prompts promoted generation of additional correct procedures. Indeed, 57% of students in this condition used at least two correct procedures during the intervention, compared to only 24% of students who received instruction but were not prompted to explain. Using multiple procedures is a common feature of development and is beneficial to performance (Siegler, 2002). Prompts to explain under invention conditions also promoted invention of a correct procedure (Siegler, 2002).

Overall, students in the explain conditions were better able to solve transfer problems, regardless of instruction. Analysis of students' explanations revealed that students rarely explained the rationale for why a solution was correct. Approximately 8% of explanations included mention of equal sides or the importance of the equal sign. Rather, most why explanations where ambiguous or described the procedure for solving the problem. Combined with the finding that explanations did not influence conceptual learning, this suggests that prompts to self-explain on a problem-solving task promote exploration of alternative procedures but not reflection on conceptual-underpinnings of the procedure. Self-explanations are one promising mechanism for explaining why some learners make improvements in conceptual understanding after learning a new procedure while others do not (Rittle-Johnson & Alibali, 1999), but the current study does not support this hypothesis.

The current findings have important implications for the debate between use of direct instruction vs. encouragement to invent procedures. There are serious limitations to relying on people inventing correct procedures without guidance on effect learning processes. Half of the students in the current study never invented a correct procedure when receiving feedback on the correct answers alone. Some prior research has suggested that feedback is critical to supporting invention during exploration, but even this level of support was insufficient for many learners (e.g Lacher, 1983). In comparison, as in previous studies, direct instruction supported rapid adoption of a narrowly used procedure (e.g. Alibali, 1999). A third of the students in the instruction groups failed to generalize the add-subtract procedure even when receiving feedback during the intervention, and instruction only supported transfer to a very similar problem that required no adaptation to the

procedure. On the hardest problem, having received instruction interfered with problem solving. Overall, direct instruction by itself appears to be a quick route to inert knowledge. However, promoting engagement in effective learning processes, such as self-explaining, helped students to avoid many of the downsides of both invention and direct instruction.

Overall, it is not the source of procedure, but rather engagement in a fundamental learning process, self-explanation, that is important for promoting flexible problem-solving.

Acknowledgments

This research was supported by a small research grant from Peabody College, Vanderbilt University. A special thanks to Jennifer Behnke and Gayathri Narasimham for collecting the data and to Kathryn Swygart, Betsy Thomas and Stephanie Crisafulli for help coding and entering the data.

References

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*(2), 147-179.

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35*(1), 127-145.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*(2), 221-252.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2001). *How People Learn: Brain, Mind, Experience, and School* (Expanded Edition ed.). Washington, D.C.: National Academy Press.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*(2), 155-192.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53*, 1045-1055.

Hiebert, J., & Wearne, D. (1986). Procedures over concepts: The acquisition of decimal number knowledge. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 199-223). Hillsdale, NJ: Erlbaum.

Hiebert, J., & Wearne, D. (1996). Instruction, Understanding, and Skill in Multidigit Addition and Subtraction. *Cognition and Instruction, 14*(3), 251-283.

Judd, C. H. (1908). The relation of special training to general intelligence. *Educational Review, 36*, 28-42.

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12*(3), 317-326.

Lacher, M. B. (1983). Effects of feedback, instruction, and initial performance level upon training and persistence of verbal rehearsal. *Journal of General Psychology, 108*(1), 43-54.

Lewis, E. L., & Linn, M. C. (1994). Heat energy and temperature concepts of adolescents, adults, and experts: Implications for curricular improvements. *Journal of Research in Science Teaching, 31*(6), 657-677.

Mayer, R. E. (2004). Should there be a thre-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59*(1), 14-19.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*(1), 175-189.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Garnott & J. Parziale (Eds.), *Microdevelopment: A process-oriented perspective for studying development and learning* (pp. 31-58). Cambrdige: Cambridge University Press.

Stigler, J. W., & Hiebert, J. (1997). Understanding and Improving Classroom Mathematics Instruction. *Phi Delta Kappan, 79*(1), 14-21.

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22*(5), 366-389.

The effect of stimulus familiarity on modality dominance

Christopher W. Robinson (robinson.777@osu.edu)
Center for Cognitive Science
The Ohio State University
207D Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

When unfamiliar non-speech sounds and visual input co-occur, they often compete for attention, with auditory input overshadowing visual information for infants and young children (Robinson & Sloutsky, in press; Sloutsky & Napolitano, 2003). The current study investigated whether labels and familiar sounds also compete for attention with corresponding visual information in infancy. The results indicate that, unlike unfamiliar, non-speech sounds, labels do not compete for attention with corresponding visual information at 16-months of age: 16-month-olds ably encoded both auditory and visual information. At the same time 8-month-olds only encoded the labels. When infants were familiarized to the same non-speech sounds that overshadowed visual input in Robinson and Sloutsky's study, 16-month-olds encoded both auditory and visual information, whereas, 8-month-olds continued to encode only the sounds. These findings, in conjunction with the findings of Robinson and Sloutsky (in press) and Sloutsky and Napolitano (2003), point to an important developmental progression in processing of auditory and visual information.

Introduction

Language plays an important role in conceptual development. When two entities share a common label, children are more likely to perceive these entities as being more similar to each other (Sloutsky & Lo, 1999), more likely to group these entities together (Sloutsky, Lo, & Fisher, 2001), and more likely to make inferences from one entity to the other (Gelman & Markman, 1986; Sloutsky, et al., 2001).

The effect of linguistic input on categorization appears very early in development. Even 9-months-olds were purported to benefit from linguistic input when forming object categories (Balaban & Waxman, 1997). In particular, it has been argued that "…from the onset of acquisition, object naming and object categorization are linked. Infants across the world begin the task of word learning equipped with a broad, universal expectation that directs them to link novel words to commonalities among objects." (Waxman, 2003, p. 213).

For example, in Balaban and Waxman's (1997) study, 9-month-olds who heard labels or content-filtered speech (which retained the original prosodic pattern) were more likely to categorize entities at the basic-level than children who only heard sounds. Therefore, it appears that hearing the same linguistic input associated with different exemplars helps infants group these exemplars together. Labels can also help infants detect differences between objects (Xu, 2002). Here, 9-month-olds are more likely to differentiate two objects when the two objects are associated with different labels. Thus, hearing the same label associated with different exemplars helps infants group these objects together, and hearing different labels helps infants differentiate the objects.

Various mechanisms have been proposed in an attempt to explain the importance of linguistic input on conceptual development. Language-specific explanations suggest that children understand that entities belong to categories, and labels highlight categories (Gelman & Markman, 1987). Labels may also be weighed heavier than other features such as appearance because children may be attentive to the prosody of human speech (Balaban & Waxman, 1997). From a general-auditory explanation, labels may initially be weighed heavier than other features because labels are presented to the auditory modality. Moreover, auditory information receives privileged processing early in development (Robinson & Sloutsky, in press; Sloutsky & Napolitano, 2003).

In support of a general-auditory explanation, Sloutsky and Napolitano (2003) demonstrated that modality preference changes throughout development: Four-year-olds are more likely to attend to auditory input, whereas adults are more

likely to attend to visual input. This finding suggests that the greater attention to auditory information may explain, in part, the effects of labels.

More recently, Robinson and Sloutsky (in press) extended these findings with infants as young as 8-months of age. Here, infants were familiarized to an auditory-visual compound stimulus ($AUD_{old}VIS_{old}$). After familiarization, infants were presented with four different test trials ($AUD_{old}VIS_{old}$ and $AUD_{new}VIS_{new}$), which served as within subjects controls and ($AUD_{new}VIS_{old}$ and $AUD_{old}VIS_{new}$), which were used to determine if infants were primarily attending to auditory, visual, or both auditory and visual components during familiarization. If infants attend to a specific component during familiarization, looking should increase when that component changes at test. In sum, infants increased looking when either the auditory component or both components changed ($AUD_{new}VIS_{old}$ and $AUD_{new}VIS_{new}$); however, infants at 8-, 12-, and 16-months of age did not increase looking when only the visual component changed ($AUD_{old}VIS_{new}$). This finding suggests that infants were primarily attending to the auditory input during familiarization. At the same time, infants amply encoded the visual component when it was presented in isolation, which suggests that the auditory component overshadowed the visual component.

These results point to auditory dominance early in development and they have several important implications. Most importantly, auditory dominance effects can provide a coherent account for many of the previous findings. Recall that it has been argued that common labels help infants detect commonalities between objects, and different labels help children differentiate objects. Although infants in Robinson and Sloutsky (in press) study were presented with non-speech sounds, the pattern of results are identical to what would be expected if infants were presented with linguistic labels (i.e., the same visual stimulus that was presented during familiarization was perceived as new when paired with a new sound and a new visual stimulus was perceived as old when paired with the old sound). In short, it seems possible that under both speech and non-speech auditory input conditions, infants rely primarily on the auditory information.

The aim of Experiment 1 was to test this hypothesis by investigating whether linguistic labels, similar to the non-speech sounds (Robinson & Sloutsky, in press), overshadow visual input. In particular, if linguistic input is weighed heavier than visual input because it represents auditory information then non-speech sounds and labels should reveal similar patterns of results.

Experiment 1

Method

Participants Nineteen 8-month-olds (5 boys and 14 girls, *M* = 249 days, *Range* = 231 - 280 days) and nineteen 16-month-olds (6 boys and 13 girls, *M* = 489 days, *Range* = 470 - 501 days) participated in this experiment. Parents' names were collected from local birth announcements, and contact information was obtained through local directories. All children were full-term (i.e., > 2500g birth weight) with no auditory or visual deficits, as reported by parents. A majority of infants were Caucasian. Seven infants were not included due to fussiness, and 10 infants were excluded because they did not reach the training criterion indicated below.

Apparatus Infants were seated on parents' laps approximately 100 cm away from a 152 cm x 127 cm projection screen, which was located approximately 5 cm above the infant's eye level. A Sony DCR-TRV40 camcorder was used to capture infants' fixations and was projected to one of two Dell flat panel monitors in the observation room. An NEC GT2150 LCD projector was mounted on the ceiling approximately 30 cm behind the infant (130 cm away from the projection screen). Two Boston Acoustics 380 speakers were 76 cm apart from each other and mounted in the wall. The speakers and camcorder were concealed by black felt and located directly below the projection screen. Two small lights were located behind the infant to ensure that the room was dimly lit throughout the entire procedure. In an adjacent room, a Dell Dimension 8200 computer with *Presentation* software was used to present stimuli to the infants, as well as to record the onset and offset of infant's visual fixations. Fixations were recorded online by pressing a button on an Excalibur 10-button gamepad when infants were looking at the stimulus and releasing the button when infants looked away from the stimulus. A second Sony DCR-PC120 camcorder was used to record the video stream of the infant from the monitor indicated above, as well as to record the image of the stimulus presentation on a second Dell flat panel monitor. This split screen recording was used to establish interrater reliability.

Stimuli Each infant was familiarized to an auditory-visual compound stimulus ($AUd_{old}VIS_{old}$) and tested on four auditory/visual combinations ($AUD_{new}VIS_{old}$, $AUD_{old}VIS_{new}$, $AUD_{new}VIS_{new}$, and $AUD_{old}VIS_{old}$). The auditory components consisted of two infant-directed nonsense labels (vika and kuna), which were presented at 65-68 dB. The visual components

consisted of two three-shape patterns (circle, pentagon, triangle, and cross, octagon, square), and were projected to 25 cm x 7 cm in size. Previous research has demonstrated that infants can discriminate these visual stimuli when presented in isolation; however, they are overshadowed by unfamiliar non-speech sounds (Robinson & Sloutsky, in press; Experiment 2).

Procedure The procedure consisted of 10 familiarization trials, 2 test trials, 3 retraining trials, and 2 more test trials. Each familiarization trial consisted of a compound stimulus that appeared for 1000 ms and disappeared for 500 ms. Each stimulus appeared five times during each trial (7500 ms trial duration). After familiarization, infants were present with 4 different test trials ($AUD_{new}VIS_{old}$, $AUD_{old}VIS_{new}$, $AUD_{new}VIS_{new}$, and $AUD_{old}VIS_{old}$). Test trials were 12 s in duration and were randomized so that each test stimulus had an equally likely chance of appearing as the first test trial, last test trial, etc. The retraining trials were the same as familiarization trials and were used to remind infants of the familiarization stimulus. Retraining trials always appeared between the first two and last two test trials. Fixations were recorded online by an experimenter for all training, test, and retraining trials. A random sample of 25% of the infants were coded offline by experimenters who were blind to the auditory and visual components presented to infants. No differences were found between subjects coded on- and offline.

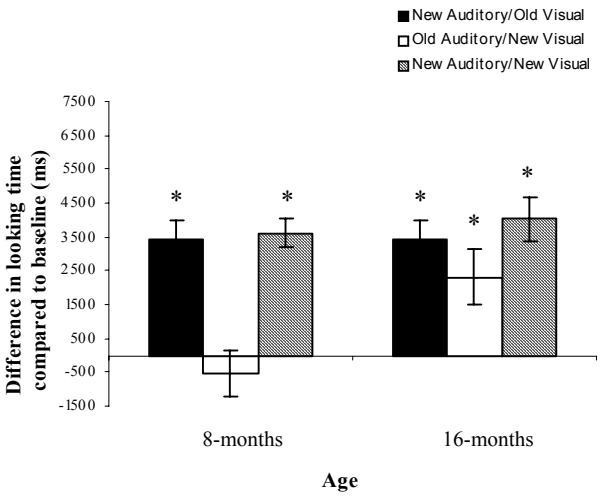
Results and Discussion

Training Criterion. Only infants who demonstrated a novelty preference at test were included in additional analyses (i.e., looking to $AUD_{new}VIS_{new} > AUD_{old}VIS_{old}$). As reported above, 10 infants did not reach this criterion.

Test Trials. Analysis of test trials focused on whether infants were primarily attending to auditory and/or visual input during familiarization. A difference score was calculated by taking the accumulated looking to each test stimulus and subtracting it from baseline (e.g., the effect of changing the auditory component = $AUD_{new}VIS_{old} - AUD_{old}VIS_{old}$). Thus, positive numbers indicate that looking increased as a function of changing a specific stimulus component, which suggests that infants encoded that modality during training. As can be seen in Figure 1, at 8- and 16-months of age, looking increased when the auditory component changed and when both auditory and visual components changed, one-sample ts > 0, ts > 5, ps < .001. In contrast, only the 16-month-olds increased looking when the visual stimulus changed, one-sample t > 0, t (18) = 2.88, p < .01.

A 2 (Age: 8-months, 16-months) x 3 (Test Trial: $AUD_{new}VIS_{old}$, $AUD_{old}VIS_{new}$, $AUD_{new}VIS_{new}$) revealed an effect of Test Trial and also confirmed the Age x Test Trial interaction, Fs > 5, ps < .01. At 8-months of age, changing the auditory component had a larger effect than changing the visual component, $DIFF_{AUDnewVISold}$ = 3435 ms > $DIFF_{AUDoldVISnew}$ = -552 ms, paired t (18) = 5.87, p < .001. This difference, however, attenuated at 16-months of age ($DIFF_{AUDnewVISold}$ = 3403 ms = $DIFF_{AUDoldVISnew}$ = 2318 ms), paired t (18) = 1.40, p > .1.

Figure 1. Effects of changing labels and visual stimuli in Experiment 1



Note: *Difference score > 0, p < .01. Error bars represent standard errors.

It is important to note that, although the nonsense labels overshadowed visual input at 8 months of age, these same visual stimuli were ably encoded by 8-months-olds when presented in isolation (Robinson & Sloutsky, in press). In contrast, 16-month-olds encoded both the auditory and visual components. This pattern of results is strikingly different from those reported by Robinson & Sloutsky. In particular, when the same visual stimuli were paired with unfamiliar non-speech sounds (laser and static sounds), 8- , 12-, and 16-month-olds only encoded the auditory component. Thus, the results from the current experiment, in conjunction with Robinson & Sloutsky, demonstrate that both speech and non-speech sounds overshadow visual input at 8-months of age. In contrast, by 16-months of age children encode both auditory and visual components; however, only when the auditory input consists of speech sounds. While revealing interesting developmental differences in effects of label on processing of visual information, the current study did not elucidate the nature of these effects.

Experiment 2

The goal of Experiment 2 was to determine whether the effect of label stems from language-specific properties or from general-attentional effects. From a language-specific perspective, the different pattern of results at 16-months of age between Experiment 1 with those reported in Robinson & Sloutsky (in press) could stem from privileged processing of linguistic input. In particular, it is possible that linguistic information does not compete for attention with corresponding visual information, which allowed 16-month-olds to process both auditory and visual information. However, it is also possible that human speech represents a familiar class, and even familiar non-speech sounds do not compete for attention with corresponding visual input. Although very few empirical studies, if any, have compared processing of familiar sounds with linguistic input early in development, there is preliminary neurophysiological evidence with adults suggesting that familiar non-speech sounds are processed in the brain similarly to words (Cycowicz & Friedman, 1998). Thus, the goal of Experiment 2 is to determine if stimulus familiarity can account for differences between Experiment 1 and Robinson & Sloutsky (in press).

Method

Participants Twenty 8-month-olds (10 boys and 10 girls, *M* = 252 days, *Range* = 245 - 269 days) and ten 16-month-olds (4 boys and 6 girls, *M* = 490 days, *Range* = 474 - 504 days) participated in this experiment. Recruitment procedures and demographics were identical to Experiment 1. Two infants were not included due to fussiness, and 13 infants were excluded because they did not demonstrate a novelty preference (i.e., $AUD_{new}VIS_{new} > AUD_{old}VIS_{old}$).

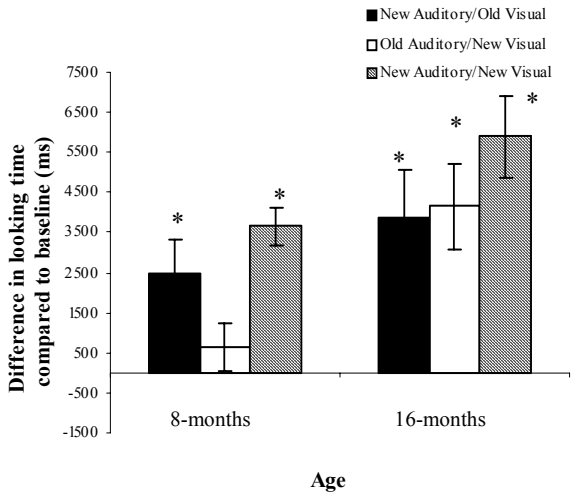
Stimuli and Procedure With two exceptions, the procedure was identical to Experiment 1. First, the nonsense labels were replaced with non-speech sounds (laser sound and static sound). Note that these same sounds overshadowed the three-shape patterns in Robinson & Sloutsky (in press). Second, and most importantly, children were familiarized to the non-speech sounds prior to the actual experiment. In the current experiment children sat on parent's laps and heard each non-speech sound 10 different times. As with the actual experiment, the auditory stimulus was presented at 65-68 dB, and each auditory stimulus lasted for 1000 ms. Auditory stimuli were presented in pairs and pseudorandomized so that infants heard the same stimulus at least twice in a row and no more than 4 times in row. In addition, the non-speech sounds were not associated with the three-shape patterns or any visual stimulus. This ensured that children in Experiments 1 and 2, and children in Robinson & Sloutsky (in press) all had equal experience with the three-shape patterns. After infants heard each sound 10 times, infants were given a 4 minute distracter task in which they looked at realistic pictures of animals. After the distracter task, infants were then presented with the main experiment.

Results and Discussion

As in Experiment 1, a difference score was calculated by taking the accumulated looking to each test stimulus and subtracting it from baseline. As can be seen in Figure 2, the pattern of results are very similar to Experiment 1. That is, both age groups increased looking when either the auditory component changed or when both auditory and visual components changed, one-sample *t*s > 0, *t*s > 3, *p*s < .01, and only the 16-month-olds increased looking when the visual stimulus changed, one-sample $t > 0$, $t(9) = 3.86$, $p < .01$.

A 2 (Age: 8-months, 16-months) x 3 (Test Trial: $AUD_{new}VIS_{old}$, $AUD_{old}VIS_{new}$, $AUD_{new}VIS_{new}$) revealed an effect of Test Trial, $F(2, 56) = 7.72$, $p < .001$. Here, children looked longer when both components changed ($DIFF_{AUDnewVISnew}$ = 4401 ms) than when only the auditory component changed ($DIFF_{AUDnewVISold}$ = 2940 ms) or when only the visual component changed ($DIFF_{AUDoldVISnew}$ = 1819 ms), paired *t*s > 2.5, $p < .01$. The above analyses also revealed an effect of Age, $F(1, 28) = 5.93$, $p < .05$, with 16-month-olds (*M* = 4631 ms) accumulating more looking across test trials than 8-month-olds (*M* = 2264 ms).

Figure 2. Effects of changing familiar sounds and visual stimuli in Experiment 2



Note: *Difference score > 0, $p < .01$. Error bars represent standard errors.

General Discussion

The results from the two experiments in conjunction with Robinson & Sloutsky (in press) demonstrate that unfamiliar non-speech sounds, familiar non-speech sounds, and nonsense labels all overshadow visual input at 8-months of age. That is, 8-month-olds do not discriminate visual stimuli when these images are paired with auditory input; however, they ably discriminate the same images when presented in isolation (Robinson & Sloutsky, in press). In contrast, 16-month-olds encode both the auditory and visual components; however, only when the visual stimuli are paired with labels or familiar sounds. Interestingly, the non-speech sounds that children heard in Experiment 2 were the same non-speech sounds that overshadowed the three-shape patterns in Robinson and Sloutsky's study. These findings demonstrate that, at 16-months of age, just hearing an auditory stimulus a few times affects the way children attend to auditory and visual input. These findings also demonstrate that familiar sounds and labels have similar effects on processing of auditory and visual information at 8- and 16-months of age.

Overall, the current study expands previous research concerning the development of attention, the role of familiarity in the auditory modality, and possible mechanisms underlying the effect of labels on conceptual development.

One potential explanation of the developmental differences found in the current study concerns the notion that attentional biases and attentional resources change considerably throughout development. There is a growing body of research demonstrating that younger children are more likely than adults to demonstrate a preference for auditory input and more likely to encode only one modality (Robinson & Sloutsky, in press; Sloutsky & Napolitano, 2003). Currently, there are several possible mechanisms that may explain this developmental pattern. First, it is possible that young children lack attentional resources that are needed for simultaneously processing auditory and visual input. However, it is also possible that young children either habituate to and/or process auditory information faster than visual information. Future research will need to address this issue.

The current study also introduces the notion that familiar sounds and labels may play a similar role early in development. Although there is neurophysiological work demonstrating that familiar sounds are processed in the brain similarly to words (Cycowicz & Friedman, 1998), the current study provides behavioral evidence for this notion in infancy. One interesting question concerns the idea that labels may represent a familiar class of auditory stimuli. This would explain why labels and familiar sounds have similar effects in the adult brain, as well as in the current study.

At a more general level, it is well known that linguistic input plays a large role in conceptual development. However, it is uncertain how and when labels become special. Even as young as 9-months of age, hearing the same label associated with different exemplars helps infants group these objects together, and hearing different labels helps infants differentiate objects (Balaban & Waxman, 1997; Xu, 2002). Interestingly, 8-month-olds in the current study demonstrated the same pattern of results when presented with unfamiliar non-speech sounds (Robinson & Sloutsky, in press), labels, and familiar sounds. This suggests that young children may initially rely on various types of auditory information (sounds and labels), and this initial preference for auditory input may help bootstrap labels into a special status.

Acknowledgments

This research has been supported by a grant from the National Science Foundation (BCS # 0078945) to Vladimir M. Sloutsky.

References

Balaban, M.T., & Waxman, S.R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, *64*, 3-26.

Cycowicz, Y.M., Friedman, D. (1998). Effect of sound familiarity on the event-related potentials elicited by novel environmental sounds, Brain and *Cognition*, *36*, 30-51.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.

Napolitano, A. V., & Sloutsky, V. M. (2003). Flexible attention and modality preference in young children. In R. Alterman & D. Kirsh (Eds.), Proceedings of the XXV Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.

Robinson. C.W. & Sloutsky, V.M. (in press). Auditory dominance and its change in the course of development. *Child Development*.

Sloutsky, V. M., & Lo, Y. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. Developmental Psychology, 6, 1478-1492.

Sloutsky, V. M., Lo, Y., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*, 1695-1709.

Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*, 822-833.

Waxman, S. R. (2003). Links between object categorization and naming: Origins and emergence in human infants. In D. H. Rakison & L. M. Oakes (Eds.), Early category and concept development: Making sense of the blooming, buzzing confusion (pp. 213-241). London, UK: Oxford University Press.

Xu, F. (2002). The role of language in aquiring object kind concepts in infancy. *Cognition*, *85*, 223-250.

The transformation of scientific information through artifacts

L. Fernando Romero (fer@asu.edu)
Psychology in Education, #0611
Tempe, AZ 85287-0611 USA

Sarah K. Brem (sarah.brem@asu.edu)
Psychology in Education, #0611
Tempe, AZ 85287-0611 USA

Abstract

We use artifact analysis to describe the process of scientific dissemination in a community-based program that informs parents and professional caregivers about early childhood development. We define this program as a network of information management and our unit of analysis are the sociocultural activities of dissemination, and the artifacts that shape them. Drawing upon activity theory, social networks theory, and distributed practice, we describe and analyze the impact, evolution, and sociocultural nature of understandings, goals, values, artifacts, actions, events, and organizational elements. Our data were collected through observations, field notes, focus groups, artifact collection, and stimulated recall interviews. Results suggest that as artifacts move from one environment to another, their role changes, often resulting in a loss or distortion of information. We describe how and why these problems are overlooked and the potential problems they may create.

Studies of scientific dissemination are rich sources of information about cognitive processes situated in a sociocultural context. The dissemination process has been almost completely the domain of large corporate, government, or academic entities—universities, pharmaceutical companies, the National Institute of Health, media networks, and the like. The role of lay people and their communities has been largely one of end-user, with the assumption that they could be expected to act as recipients of information rather than disseminators; a passive role at the bottom of the organizational structure (Epstein, 1996)

The importance of community involvement in education, advocacy, and decision making has been growing over the past decade (Minkler & Wallerstein, 2002). At the local level, the project team identifies community needs through community engagement (Minkler & Wallerstein, 2002), an approach to research and intervention characterized by its use of the community as a unit of identity, action and analysis. Communities may be formed around geography, socioeconomic status, shared emotions, or common goals. Facilitators are community members who bring scientific information to the attention of local end-users, translate concepts and terms, and help end-users apply the information in making personal decisions.

Facilitators can also inform disseminators and scientists about end-user interests and needs; thus, ideally, information can flow in both directions. However, facilitators need aid in finding and organizing information, contextualizing scientific findings, applying them to local situations, providing emotional support, and serving as advocates and spokespeople. In short, facilitators need support to provide support, in terms of content, culturally relevant delivery, and information management. There is growing evidence for community-based dissemination, scientific communication that is culturally responsive, accounts for audiences' prior knowledge and ability/willingness to acquire new knowledge, and is flexible enough to fit diverse goals, resources, and interests. (e.g., Minkler & Wallenstein, 2003; Shonkoff & Phillips, 2000; Wilcox, Hadley, & Bacon, 1998).

This creates interesting questions regarding reasoning in community engagement setting, especially in regard to facilitators and outreach personnel. They are, on a number of dimensions, in limbo. Regarding the scientific content, they are neither experts nor novices; they usually have some teaching or outreach experience, but they often have never functioned in this role with this population before. They are engaged in scientific dissemination, but they are not part of the groups usually studied in the context of dissemination studies, such as scientists, media, or teachers.

Research on community-based interventions also offers interesting opportunities for dissemination research. A main reason is that facilitators are engaged in a process that requires a quick turnaround; their training may last few weeks or months and they are soon ready to work in the field. Updates, refresher courses, and additional training are put to work within a similar timeframe. This allow us to watch the inflow and outflow of information in a way that we cannot with dissemination agents whose timeline involves years of training or experience, such as a journalist, scientist, or social worker.

In short, community engagement and community facilitators are playing an increasingly important role in scientific dissemination, they are unusual in a number of ways, and they also provide opportunities to watch the dissemination process in a compressed format. Of course, this can both create unusual patterns and behaviors that are not seen in other areas of dissemination, but it does not necessitate uniqueness. Therefore, it is initially important to examine the ways in which this format repeats patterns in other spheres of scientific dissemination, and the ways in which it reinvents these patterns or creates new ones.

To construct a framework for this comparison process, we draw upon multiple streams of dissemination research in attempting to cover the ambiguous position of facilitators

and community engagement. This includes novice reasoning about scientific information in structured settings (Sandoval, 2003; Schank & Ranney, 1995), and informal settings (Zimmermann, Bisanz & Bisanz, 1998), lay advocacy and policy involvement (Epstein, 1996; Margolis, 1996), scientists reasoning among themselves (Latour, 1987), and interactions between lay people and experts (Lemke, 1990). Using this framework, we examine a community engagement program providing parents and professional caregivers with information about new psychological and neuroscientific research on early childhood learning and development.

Sites & Program Description

The program that we have been following, *The First Teacher Project* (*FTP*), is part of a larger initiative started in the city of Chandler, AZ in 2002, *The Steps to Learning Initiative* (*StL*). Funded by an Early Learning Opportunities Act Grant from the U.S. Department of Education, *StL* was created to educate the community about the importance of early literacy and learning, develop stronger links among service providers working with children and families in the Chandler community, create a comprehensive network of early childhood programs, and make information and programs more affordable and accessible. The grant was secured and is overseen through the Mayor's Literacy Task Force, and administered by the Chandler Public Library. Other partners include the Chandler Unified School District and the Chandler (East Valley) Regional Hospital.

Chandler is one of the fastest growing cities in Arizona, with a large traditionally underserved population. In the 2000 Census, Chandler had one of the largest Latino populations in the state, ranging from 25% to over 50%, depending on neighborhood (Morrison Institute, 2001). It is also an economically diverse city, home to Intel and Motorola, but also to a federally-designated Enterprise Zone. Eighty percent of Zone residents are Latinos and 68% of households are monolingual Spanish. Seventy percent of students qualify for free or reduced-price lunch, 50% of families earn less than $5,000/year, and almost 50% of adults lack a high school diploma.

The *FTP* component of *StL* focuses on children's development from ages 0 to 3. The program focuses on sensory development, bonding and attachment, cognitive skills such as categorization and language, and the value of play and pretense. The information provided can be used to identify developmental delays, sensory deficits, and other problems early on, as well as providing parents of mainstream children with new perspectives on their children's learning and development. *FTP* involves disseminating a significant amount of scientific information, much of it relatively new even to scientists in the relevant fields. Topics include neural pruning, synaptic formation, plasticity, limbic and cortical functions, biological and psychological aspects of temperament and language acquisition.

The *FTP* initiative is coordinated by a full-time outreach coordinator. A group of 12 paid community professionals (eight educators, three librarians, and the outreach coordinator) receive forty-five hours of training, and conduct mock workshops before beginning to facilitate parent workshops in their schools and libraries. Facilitators receive continuing education on a monthly basis, and have committed to a tenure of at least 18 months. *StL* is currently looking for ways to fund and support the program beyond this 18 month timeframe.

Activity, Artifacts, Dissemination & Education

In addition to setting up a content framework, we need also to construct an epistemological and methodological framework for the analysis. We do so in a hierarchical fashion.

At the highest level, we have chosen to adopt an activity theory perspective. In activity theory, the unit of analysis is continually developing activities—events, transactions, practices—and the analysis is organized around objects that motivate, guide, and give meaning to activity. Objects have both physical and semiotic properties, and affect human interactions with their environment, as tools for physical and mental activity. Because of activity theory's emphasis on social factors and the interaction between agents and objects, it is useful for capturing the process of scientific dissemination, the practices of which depend heavily on tools and networks of social interaction.

In identifying objects that organize events and transactions of importance, we use Latour's concept of artifact (Latour, 1987). It is a fairly broad conceptualization of artifact, in which artifacts are physical entities that have been given meaning by human beings through utilization and construction.

Using this artifact-oriented approach to examine the dissemination of scientific information, specifically in the context of educational dissemination, and compare our findings to the existing research in other areas of scientific dissemination. Based on this analysis, we find that the scientific content is altered by organizational goals, available materials, etc; that it is important to distinguish explicit, tacit, and incidental features of artifacts; and that the distinction between the "scientific content" of the artifact and elements added during these alterations is often not identified by facilitators and parents.

Method

In this study, we take the perspective that the *FTP* can be conceived of as an activity system with the primary purpose of knowledge management and community dissemination. Drawing upon concepts from activity theory, social networks theory, and distributed cognition, we describe and analyze the development and consequences of stakeholders' understandings, goals, values, artifacts, actions, and organizational dynamics. We collected data over the full 18 month existence of the program, using observation, video recordings, field notes, focus groups and stimulated recall interviews, artifact collection, and surveys.

Extended observation and videography was conducted throughout the life of the program; all training sessions and most of the parent workshops were observed and/or recorded, and most of the ongoing monthly meetings have

been observed. In addition, several meetings of the grant oversight committee, the Mayor's Literacy Task Force, have been attended by at least one of the authors. Field notes provide in depth descriptions of activities, settings, and interpersonal dynamics, were the only means of establishing a record of events where we were not granted permission for videography, or when it was not appropriate to record a particular event.

Focus groups with the facilitators took the form of discussions that allowed us to collect information about their perspective, and the meaning they attached to particular artifacts and events. Video-elicited and artifact-elicited interviews are used to obtain an in-depth perspective of the local meanings teacher create in relation to key perceptions, goals, experiences, actions, and elements of this program.

Artifact collection and documentation refers to the process of gathering/recording objects and conceptual symbols. Artifacts are objects that have both material and conceptual characteristics and that have been transformed through the history of this program. This category includes curriculum binders, slides, handouts, props, toys, logos, memos, announcements, electronic newsletters, websites, acronyms, jargon, and definitions.

Content area questionnaires are used to assess teachers' knowledge of infant brain development before training began, and at intervals after training. These assessments include fact-based, open-ended, and problem-solving items. A separate motivation survey was designed to address affect and efficacy in relation to distinct aspects of participation; training, instruction, curriculum materials, trainers, and programmatic characteristics.

Results & Discussion

The *First Teacher Project* is best described as an activity system configured into a dynamic network of information management. This network relies upon the interconnection of different levels of cognitive mediation (e.g., object, social, organizational). Our analysis is primarily based on the study of how these mediations become embodied into the conceptual to material continuum of artifacts. We use Collins et al (2002) hierarchy of mediating artifacts to categorize *what, how, why,* and *where-to* artifacts. The *what category* refers to artifacts that serve as a means to achieving an object (e.g., using chart paper to write down parent questions). *How* artifacts contribute to understanding how to achieve purposes or goals (e.g., using a case study to demonstrate how routines help babies). *Why* artifacts motivates achievement of the goal (e.g., presenting statistics of neglect and abuse linked to academic achievement to encourage parent-child bonding). *Where-to* artifacts motivate the evolution of all activity elements (e.g., identifying a pocket population that was not targeted and re-defining main project goals).

Artifact analysis is primarily an in depth description of the history and meaning of tools and signs that evidence intentionality and activity of agents within this network of information management. Artifact analysis is a process of analytic induction that focuses on how artifacts evidence actions that occur in specific settings and in connection to specific meanings. We use Erickson's (1990) five methods of evidentiary inadequacy to determine the degree to which we have a) adequate amounts of evidence, b) adequate variety of evidence, c) trustworthy evidence, d) adequate disconfirming evidence, and e) adequate discrepant case analysis.

An example of an artifact is the brochure community professionals put together to attract participants to the parent workshop. At one point, this brochure may represent everything target parents know about the project. However, parents are unaware of the history of this artifact, how the printed language reflects interpretations of science, how explicit goals of the workshop relate to assumptions about needs in this community, or how this workshop expects to influence parenting. The brochure is a byproduct that reflects negotiated goals, program priorities, perceptions of the target population, and a way to sum up the essential components of a newly developed expertise. The final draft of the brochure is edited by the project coordinator after asking community professionals to develop drafts, after discussing these drafts during taskforce meetings, and after receiving approval from all stakeholders. In this way, the development of a simple communication product is informative of the way this project is represented to the larger target population, the role of distributed cognition and distributed practice, and the protocols and the organizational structure necessary to develop this double-sided page. And the workshop brochure is just the entry point to the vast world of artifacts that are part of this BBE curriculum. As the parent arrives to the actual workshop he/she will be exposed to graphs, binders, slides, toys, props, sounds, video-clips, case studies, analogies, metaphors, acronyms, jargon, and abstract ideas.

It is important, too, to recognize that artifacts are not necessarily bounded physically, but by the role they play in a network of activity. We address this in our analysis by examining agent-artifact units, i.e. units comprised of an artifact and the agent who is currently making use of the artifact. Thus, a brochure handed to a parent by their child's teacher is a different agent-artifact unit than a brochure taken from a stand at the door of a library.

Content transformations

The main goal of the *FTP* is to translate neuroscience into recommended practices that will improve parenting and normal child development outcomes. Research techniques and directions, however, often do not directly support this goal. Much of the neuroscientific research available, however, has been conducted using deficit models, and highly constrained tasks and environments. Therefore, application to normal developmental practices is rarely an explicit element of the scientific report. Thus, when a report in a journal is read by a curriculum developer, the developer-report unit is a different entity than the scientist-report unit, and the report is used for different purpose (developing parenting recommendations vs. informing peers of experimental results), establishes credibility in different ways (appearance in a prestigious journal vs. surviving the actual peer-review), and becomes a symbol for establishing

authority rather than a document containing information to be examined.

The effects of this transformation are several. One way in which this is done is through broad generalizations into maxims that would be difficult to find objectionable. Statements that encourage parents to provide a stimulating but not overwhelming environment, to not neglect their children emotionally or physically, to create a loving and protective environment. The training providers believe that by taking these unobjectionable messages and pairing them with laboratory research that is tenuously connected, they will make these messages more persuasive by making them more authoritative and making them appear to be based in "science."

Another approach is to take deficit model findings and transform them into "best practice" recommendations. The logic, roughly, is that if the absence of certain elements has a deleterious effect, then parents should be encouraged to make sure these elements are present. While this is not always faulty logic, it can at times produce the implication that since less is bad, more is better, and that greater amounts of play, visual stimulation, exposure to human faces, and so on, will have a beneficial effect beyond that which normal caregiving would provide. Research studies based on abnormal case studies produce dramatic research findings on how neurological disorders, neglect and abuse can adversely affect brain development. However, this program is not designed to target parents of children with major disabilities, but to target the general population. In this way, research findings from deficit models are discussed outside of their context, and derived applications may involve unwarranted alterations of the science content. For example, facilitators are taught that physiological and psychological traumatic events can chronically elevate an individual's cortisol levels, which in turn may result in the destruction of neurons or a reduction of synaptic connections. Children who have high levels of cortisol in response to trauma have been shown to experience more developmental delays (Gunnar, 1996). A key artifact here is a video-clip interview of neuroscientist who explains how cortisol levels show how the brain responds to stress levels. In the context of a parent workshop or facilitator training, this functions not so much as a way to deliver information but to prove the curriculum's scientific backing. That sustained high levels of cortisol can cause delays does not imply either that transient elevation from minor stresses will cause problems, nor does it imply that extremely low stress will facilitate development. With community facilitators there is a tendency to blur two distinctions: the difference between stress and trauma, and the distinction between temporary and permanent changes in cortisol levels.

The content may also be transformed because of the physical constraints imposed as artifacts are paired with new agents. An example is the inclusion of infant massage experiments in the curriculum. A meta-analysis conducted by the Cochrane Review found the evidence to be weak, though in the direction of supporting the use of massage with infants receiving neo-natal intensive care (Vickers, Ohlsson, Lacy, & Horsley, 2004). Findings regarding its use in other areas appear likewise ambiguous.

Those experiments supporting infant massage as beneficial are incorporated into popular books (e.g., Field, 2000) by clinicians and researchers that wish to make their case with the public. These are then taken by curriculum developers and integrated with specific *how-to* activities and instructions that guide parents into giving leg, foot, arm and hand, face and head massages to infants. *How-to* activities often have not been equally researched, though this distinction is not made in the materials given to facilitators during training. A focal artifact here is a written description created by the curriculum developers where they describe that, ideally, parents are to take into consideration the age of the child to determine the type, duration and frequency of massaging a baby: A massage for a newborn baby should be limited to 3-5 minutes, while a month old baby can receive a 10 minute massage. In addition, it is said that parents should be attentive to determine individual differences in responding and tolerating touch. When the community professionals attended their training, these written instructions are verbally described by the trainers who also modeled concepts by using realistic baby dolls. Participants practice the massage on these dolls, which then become part of their representation of infant massage.

However, the specifics of the infant massage curriculum are not covered in the same fashion by facilitators as they bring this information to parents. The infant massage demonstration and hands-on activity is time-consuming, so the facilitators do not have the same opportunity to emphasize this topic as do the developers. Moreover, they have only one doll, making even demonstrating to a group somewhat difficult. The facilitators rely on slides showing bullet points that summarize the main ideas on how touch enhances bonding. Moreover, we observed that discussions often wandered onto interesting but misleading tangents, such as formal training in infant massage therapy. The superficial overview in the context of such a discussion is misinterpreted by some parents as a need to seek a special training. There is not a deliberate plan to distort information, but the way information is presented has an unexpected effect.

Artifacts & Expectations

During the initial intensive period of training, *NDI* structured the content of its curriculum around two acronyms *STEPS* for security, touch, eyes, play, and sound and *ABC*'s of learning for attention, bonding, and communication. These acronyms were developed as a way to organize the curriculum, and resulted from feedback of prior *FTP* programs that had been implemented in different communities. Those prior programs received a similar training with very little organizing structure, and limited curriculum materials. *FTP* trainees received a binder with five major divisions that corresponded to the *STEPS* acronym, their training was structured around these topics, and each of the *STEPS* concepts was discussed as relating to the *ABC*'s. In addition, *NDI* developed a wide array of materials that included slides, power-point presentations, video-clips and activity sets called the brain boxes. In turn, the facilitators structured their first parent workshop series as five meetings, each of which reflected the *STEPS*

structure. The different sites reported that five workshops proved to have a high turnover, and *StL* decided along with *NDI* to structure the curriculum around 3 meetings. Then, the curriculum started to be reorganized around the *ABC*'s as the guiding acronym, with the *STEPS* concepts subsumed. These acronyms are artifacts that reflect which concepts are central to *NDI*, how those concepts help organize activity, and the affordances those acronyms have. These acronyms are just one way of organizing knowledge that links neuroscience with parenting and child development, and they in fact seem to be useful in organizing workshops and discussions. Our preliminary findings suggest that these acronyms act as a paradigm through which experts, trainees, and parents think about and recall infant brain development information. For example, during a focus group activity we asked facilitators to write down which are the most compelling ideas they take from this training, and most participants referred to the acronyms. Other ways in which this acronyms influence information management and distributed cognition is that they implicitly convey the idea that these categories are all encompassing, and that scientific information is stable. We discuss these ideas further in the next section.

Artifacts may set up expectations because of their appearance or immediately perceived function. If they are improperly designed, or if the design is misinterpreted, problems can occur. To illustrate this point we refer to the script NDI puts together for facilitators to guide their presentation as they conduct the parent workshops. The script is text that corresponds to a particular slide and elaborates the main ideas represented. The parent workshop is usually structured around a series of slides, and the series of slides are connected through an overarching curriculum concept (e.g., security, eyes, touch). Most facilitators plan their parent workshops by reviewing this script, and they often refer back to the script as they conduct their presentations. This way of implementing the workshop is efficient in conveying many concepts to the parents who are part of the audience, and it also creates consistency and a good point of reference across facilitators. On the other side, the script winds up dictating most of what is said during the parent workshop. The script sets the tone for the presentation of information slide after slide, with facilitators either rephrasing or reading off the script. As a result, it is not infrequent to observe that the workshop is run as a forty-slides presentation with very few questions asked. Therefore, the script drives the workshop, leaving a small amount of time for unscripted events, which is taken up making introductions, allowing for breaks, doing take-home activities, and checking out materials.

Furthermore, the script and slides seem to endorse the perception that these curriculum materials are a self-contained representations of brain-research; sufficient to achieve the main goals of the workshop. Facilitators do not feel the need to continue exploring the science beyond this point. During continuous support meetings, NDI has emphasized the importance of speaking more explicitly about the specific brain research facts and language. Facilitators try to adjust by using the language that is part of the script, but do not go beyond this information.

As facilitators become more experienced in conducting the parent workshop, however, they take greater ownership of the content. They rely less on the script and make use of personal examples that have been effective in the past. Still, when parents ask questions that relate to more specific details of the research, facilitators have difficulty addressing those questions. A parent asked how scientists know that children see blurry at first and they see faces very clearly around three months. Even though an explanation of techniques used to determine babies' responses is provided for facilitators, they only seem reliably aware of the information presented through the script. They have a very hard time addressing those issues if they feel the question must have a right answer that is lying somewhere in a library. Facilitators are often more successful if they can find examples that relate to their own personal experiences as caregivers or teachers. They are capable of finding connections that are relevant and that help illustrate the main points, but when questions are asked about the scientific content that cannot be grounded in case from their experience, facilitators quickly face difficulties.

Conclusions

The unique characteristics of community based programs for the dissemination of scientific information include a rapid training turnaround and the opportunities to document how science concepts are transformed through actions, objects, social interaction, and organizational elements.

Community-based programs for the dissemination of science are complex activity systems that manage information in ways that reflect elements such as organizational knowledge, learning, and culture. In this particular case, the *FTP* program based high-level goals by presenting them as truisms that are difficult to challenge (e.g., parents should create a loving environment), while tacit low-level goals go underdetermined (e.g., research based on deficit models is applied to the general population). As a result, neuroscience is translated in ways that bypass issues of ecological validity.

Artifacts with flawed designs, or artifacts that are misinterpreted are likely to create problems that can go unidentified. Scripts that are meant to guide facilitators end up dictating the pace of the parent workshops in ways that limit parents' active participation, and in ways that communicate to facilitators that these materials are a finished-all-inclusive product. Finally, the development of artifacts such as communications (e.g., newsletters, brochures) give insight into how goals are proposed, negotiated, and enacted. This analysis also illustrates how the entire system works as a network that manages information.

Acknowledgments

This study is funded by an ASU/Spencer Fellowship to the first author, and an NSF CAREER grant to the second author.

References

Epstein, S. (1996) Impure Science; AIDS, activism, and the politics of knowledge. Berkeley, CA: University of California Press.

Fields, T. (2000) Touch Therapy. NY: Harcourt-Brace.

Gunnar, M. (1996). Quality of care and the buffering of stress physiology: Its potential in protecting the developing human brain. University of Minnesota Institute of Child Development.

Latour, B. (1987). Science in action. Cambridge, MA: Harvard University Press.

Margolis, H. (1996). Dealing with Risk: Why the Public and the Experts Disagree on Environmental Issues. Chicago, IL: University of Chicago Press.

Minkler, M. & Wallerstein, N. (2002). Community-based participatory research for health. San Francisco, CA: Jossey-Bass.

Ranney, M. & Schank, P. (1995). Protocol modeling, textual analysis, the bifurcation/bootstrapping method, and Convince Me: Computer-based techniques for studying beliefs and their revision. Behavior Research Methods, Instruments, and Computers, 27, 239-243.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. Journal of the Learning Sciences, 12, 5-51.

Vickers, A., Ohlsson, A., Lacy, J.B., Horsley, A. (2004) Massage for promoting growth and development of preterm and/or low birth-weight infants (Cochrane Review). In: The Cochrane Library, Issue 1. Chichester, UK: John Wiley & Sons, Ltd.

Zimmerman, C., Bisanz, G. L., & Bisanz, J. (1998). Everyday scientific literacy: Do students use information about the social context and methods of research to evaluate news briefs about science? The Alberta Journal of Educational Research , 44**,** 188-207.

The Influence of the Tutee in Learning by Peer Tutoring

Rod D. Roscoe (roscoe@pitt.edu) **Michelene T. H. Chi (chi@pitt.edu)**

Learning Research and Development Center, 3939 O'Hara Street
Department of Psychology, University of Pittsburgh
Pittsburgh, PA 15260 USA

Abstract

Previous research has demonstrated that students can learn by tutoring other students. Tutors are thought to learn because they generate instructional explanations and monitor their own understanding while teaching. We analyzed verbal data from tutorial sessions to explore how the *tutees* influence this process. We found that tutors were primarily responsible for introducing topics, but the tutees stimulated more thorough discussions of topics. We also found that tutee questions influenced tutor explanations and metacognition. Tutor responses to "deep" questions were more likely to contain inferences and self-monitoring than responses to "shallow" questions. In sum, tutees had a significant and positive influence on the tutors' learning activities and opportunities.

Introduction

Peer tutoring and cross-age tutoring are popular and cost-efficient educational interventions in which students provide instruction for other students. One reason for the widespread use of these interventions is their effectiveness – with training, students seem quite capable of successfully teaching each other and younger pupils (e.g. Cohen, Kulik, & Kulik, 1982; Greenwood, Carta, & Hall, 1988). Another reason for the popularity of peer and cross-age tutoring programs is the robust finding that the tutors also benefit academically from the teaching experience (e.g. Allen & Feldman, 1973; Annis, 1983; Cloward, 1967; Cohen et al., 1982; Greenwood et al., 1998; Morgan & Toy, 1970; Rekrut, 1992). Based on such findings, some researchers have advocated reciprocal tutoring programs in which the participating students take turns being the tutor and tutee. In general, these programs are educationally effective (e.g. Fantuzzo, King, & Heller, 1992; Fantuzzo et al., 1989; Fuchs et al., 1997; King, Staffieri, & Adelgais, 1998; Palincsar & Brown, 1984).

Why do students learn by tutoring? Some evidence suggests that tutors learn by generating instructional explanations, which facilitates integration and organization of knowledge. For example, Coleman, Brown, & Rivkin (1997) found that when students were told to teach a peer by explaining, they learned better than students told to teach by summarizing and better than students who did not teach. Similarly, Fuchs et al. (1997) showed that training students to give each other conceptually-rich explanations during reciprocal tutoring was more effective than classroom instruction and reciprocal tutoring without such explanations. Additional evidence indicates that tutoring may also encourage students to engage in metacognitive self-monitoring, which helps learners to detect and repair missing knowledge and misconceptions. For example, King et al. (1998) trained reciprocal tutors to give quality explanations and to ask each other questions that stimulated critical thinking and self-monitoring. They found that these explaining and metacognitive activities resulted in better learning than explaining activities alone. Explaining and self-monitoring have also been shown to improve learning in solo studying (e.g. Chi, 2000; Chi, deLeeuw, Chiu, & LaVancher, 1994) and collaborative learning (e.g. Coleman, 1998; Webb, Troper, & Fall, 1995), which further highlights the efficacy of these activities.

In this paper, we explore the hypothesis that tutees influence the learning activities of the tutors in important ways. In other words, tutors might be able to learn by explaining and self-monitoring, but tutees may affect how and whether these activities occur. One way that tutees may guide the tutorial session is by choosing which topics are discussed and in how much detail, thus creating or limiting opportunities to think about the underlying ideas. Another powerful way in which tutees may influence the learning activities of the tutor is through the kinds of questions they ask. As described above, King (e.g. King, 1994; King et al., 1998) has shown that when students construct and ask each other questions based on high-level question stems (i.e. questions prompting for comparisons, justifications, causes-and-effects, evaluations, etc.), they produce better explanations and learn more effectively. Coleman (1998) has demonstrated very similar findings in collaborative learning settings with students using high-level explanation prompts. Research on naturalistic tutoring has shown that tutees do occasionally ask "deep" questions in tutoring sessions, although the majority of questions are "shallow" (Graesser & Person, 1994). These deep questions, although they may be rare, should stimulate deeper responses.

In order to address these hypotheses about the influence of the tutee on tutor learning, we analyzed tutor learning in a non-reciprocal and naturalistic (i.e. little or no training) tutoring context. This design allowed us to be more sensitive to the benefits and processes of tutoring. In reciprocal tutoring, by definition, students learn from both teaching and being taught, and thus it is almost impossible to assess the specific contribution of tutoring activities to learning in these settings. Similarly, it is possible that when tutoring programs are highly structured (i.e. training on when and how to explain, ask and answer questions, etc.), important aspects of spontaneous tutoring behaviors that positively or negatively impact learning may be obscured.

Method

Background

In a larger study, we compared learning by self-explaining to learning by explaining-to-others. Overall, we found that self-explaining was superior to explaining-to-others on measures of both deep and shallow learning. Self-explaining also seemed to more naturally foster productive learning activities. However, the focus of the current analyses is on the learning outcomes and activities associated with providing instruction for other students.

Conditions

The data we analyze here was obtained from two tutoring conditions. In one condition, a student who had read and studied a text about the human eye and retina (the tutor) taught this information to another undergraduate (the tutee) in a face-to-face setting. In a second condition, a student who had read and studied the human visual system test (the tutor) produced a videotaped explanatory lesson that could be later used by a different student to learn the material (an "anticipated" tutee). The face-to-face tutoring condition can be conceptualized as an "instructional dialogue" whereas the videotape condition can be thought of as an "instructional monologue." The participants received no formal training for the tutoring task. The tutors were simply instructed to explain the text information by "going beyond what the text says." Students in the instructional dialogue condition were encouraged to try to answer the tutees' questions.

Participants

Twenty-four college undergraduate students participated in the instructional dialogue (n = 7 tutor/tutee pairs) and instructional monologue conditions (n = 10 tutors) of the original study. In order to ensure that all participants had low prior knowledge about the learning domain (the human eye and retina), students who had taken certain biology, physiology, and neuroscience courses were not eligible to participate. Participants were paid for their time.

Materials

Human Visual System Text All tutors initially read and studied a short text describing the structure and functions of the human eye and retina. The text was divided into topic-based sections, with each topic presented on a separate page. These topics included both familiar, everyday concepts (e.g. the pupil) and unfamiliar, technical ideas (e.g. refractive properties of the vitreous humor), thereby providing ample opportunities to make connections with prior knowledge and explore new ideas. However, the text itself provided few examples or analogies. The text was accompanied by a labeled cross-section diagram of the whole eye and a schematic diagram of the retina. Prior research has shown that the availability of diagrams can support and stimulate effective explaining (Ainsworth & Loizou, 2003).

Learning Assessments Learning outcomes were assessed using two written measures. For the Definition Test, students provided definitions of key terms. For the Question Test, students responded to short-answer questions testing recall, integration, and application of information. The Definition Test can be viewed as a measure of the students' shallow learning, and the Question Test can be considered a measure of deeper learning. Both measures were scored by tabulating the number of correct and relevant ideas produced.

Procedure

The study was divided into two sessions in order to facilitate recruitment and scheduling of participants. In the first session, the tutors read and studied the text for 30 minutes and then completed both learning assessments (tutor pre-test). It should be noted that the tutors studied the text without foreknowledge of their future teaching task. The purpose of this design was to bypass complications due to preparation-to-teach effects (Bargh & Schul, 1980; Renkl, 1995). The tutees also completed both learning assessments in this phase, but did not have the opportunity to read about the visual system (tutee pre-test). In the second session, the tutors either taught an actual tutee or produced a videotaped lesson (30 minutes duration). Afterwards, the tutors and tutees completed the learning assessments again (post-test).

Coding of Tutor Activities

The tutorial sessions of the dialogue and monologue conditions were transcribed and segmented according to changes in the topic of discussion. These segments formed the boundaries of episodes, which were categorized by the type of learning activity that occurred. Several different activities were observed and are briefly described below.

Summary In "basic" summaries, the tutor paraphrased the current contents of the text without elaborating on the text ideas. In "elaborated" summaries, the tutor paraphrased the text, but also provided additional information or inferences not contained in the text. Neither type of summary was significantly correlated with learning outcomes.

Review In "basic" reviews, the tutor reviewed previously discussed information without elaboration. In "elaborated" reviews, the tutor reviewed previously covered material, but also provided new information and inferences. Elaborated reviews were highly metacognitive (i.e. students monitored themselves for understanding and accuracy) and positively correlated with learning outcomes.

Sense-Making In sense-making episodes, the tutor generated inferences and integrated text concepts in order to address a perceived misconception or one's own curiosity. Sense-making episodes were highly metacognitive (i.e. students monitored themselves for understanding and accuracy) and positively correlated with learning outcomes.

Analyses and Results

Tutor and Tutee Learning

Our results indicated that the two tutoring conditions were not equally effective for learning (Table 1). Tutors in the instructional dialogue condition performed better than tutors in the instructional monologue condition on post-test measures of shallow learning (Definition Test) and deeper learning (Question Test), although only the Definition Test difference was statistically significant after controlling for pre-test differences, $F(1,14) = 9.22$, $p = .009$.

In order to establish that the dialogue tutors were effective instructors, we compared the tutors' final scores to their tutee's final scores. For both tests, the tutees performed almost as well as their tutors, suggesting that the tutors were mostly successful in teaching their pupils (Table 1). Neither difference was significant. Although the tutees learned somewhat less than the tutors, it is still quite impressive given that the tutees were exposed to the material only once (the tutoring session) and never read the text.

Table 1: Mean Definition Test and Question Test scores.

Measure	Monologue Tutors	Dialogue Tutors	Dialogue Tutees
Definition Test	21.3	33.0	27.5
Question Test	20.2	28.0	25.4

Spontaneous and Elicited Tutor Activities

These learning outcome differences were paralleled by the extent to which the tutors engaged in episodes of integrative and metacognitive activity (Table 2). Overall, the dialogue tutors produced more elaborated review and sense-making episodes than monologue tutors, $F(1,14) = 5.47$, $p = .035$ and $F(1,15) = 16.22$, $p = .001$, respectively. No other differences were significant.

Table 2: Overall mean frequency of episodes.

Episode Category	Monologue Tutors	Dialogue Tutors
Summary		
Basic	10.5	13.0
Elaborated	4.4	5.4
Review		
Basic	3.1	6.3
Elaborated	0.2	1.7
Sense-making	0.4	3.7

In order to examine this finding more closely, we further distinguished between activities that the tutors self-initiated and activities that were elicited by the tutee. An episode was coded as "tutee-initiated" if the tutee selected the topic or asked a question leading the tutor to engage in some activity. All other episodes were categorized as "tutor-initiated". All of the monologue tutors' activities were counted as tutor-initiated because no tutee was present.

The pattern of episode frequencies (Table 3) suggests that tutors in both conditions preferred to summarize the text, while tutees in the dialogue condition elicited most of the reviewing activities. Direct comparisons of the mean frequencies of tutor and tutee-initiated activities confirmed this impression. Dialogue tutors initiated significantly more basic and elaborated summaries than dialogue tutees; $F(1,12) = 8.2$, $p < .05$ and $F(1,12) = 8.3$, $p < .05$, respectively. However, the tutees initiated significantly more basic and elaborated reviews; $F(1,12) = 7.5$, $p < .05$ and $F(1,12) = 5.3$, $p < .05$, respectively.

Table 3: Mean frequency of tutor-initiated and tutee-initiated episodes.

Episode Category	Monologue Tutor-Initiated	Dialogue Tutor-Initiated	Dialogue Tutee-Initiated
Summary			
Basic	10.5	9.7	3.3
Elaborated	4.4	4.7	0.7
Review			
Basic	3.1	1.3	5.3
Elaborated	0.2	0.2	1.5
Sense-Making	0.4	2.3	1.4

The critical difference between the monologue tutors' activities and the tutor-initiated activities of the dialogue tutors was in the occurrence of sense-making episodes; $F(1,15) = 4.5$, $p < .05$. No other difference was significant. Tutors engaged in sense-making when they realized that they had a flawed or incomplete understanding of some concept and needed to revise their own knowledge. Thus, in addition to eliciting productive reviewing of the material, the tutees seem to also directly and indirectly facilitate the tutors' recognition and repair of their own misconceptions. Perhaps the tutee's misunderstandings and questions served as a signal to the tutor that the tutor's explanations were incorrect or unclear, and this realization spurred the tutor to engage in sense-making in order to understand the material better and to be a more effective teacher.

In sum, these results provide evidence that tutors in non-reciprocal tutoring settings, and with minimal training, can learn from generating instructional explanations and self-monitoring. However, when tutors provided instruction to an actual tutee, they learned and explained more effectively. Thus, it appeared the tutees did in fact contribute to the tutors' learning activities in meaningful ways. In the next sections, we explore two hypothesized mechanisms for this influence, topic selection and tutee questions.

Topic Coverage

One way that tutees may guide the tutorial session is by choosing which topics are covered and how much time is spent on those topics. Topics that receive more thorough consideration should be better learned. To examine the coverage of topics in the tutoring sessions, each episode was coded by whether it contained a novel topic (i.e. topic was

introduced in that episode) or whether it contained a continuation of a previous topic. A continuation episode could contain a review or elaboration of the topic, and thus represents a deeper or more thorough discussion (Table 4).

Overall, we observed a clear pattern in which the tutors were primarily responsible for introducing new topics in the tutoring session (76% of novel episodes were tutor-initiated), whereas tutees stimulated much of the subsequent discussion of topics (61% of continuation episodes were tutee-initiated). This pattern was statistically significant; $\chi^2(1,N=349 \text{ episodes}) = 50.0$, $p < .001$, and indicates that tutees directly influenced opportunities for tutors to delve more deeply into the text information by selecting topics for review or elaboration.

Table 4: Introduction and continuation of topics of discussion by tutors and tutees.

Topic Selector	Novel Topic	Continued Topic
Tutor	134 (76%)	67 (39%)
Tutee	42 (24%)	106 (61%)
Totals	176	173

Tutee Questions and Tutor Responses

Another important mechanism by which tutees might influence the learning activities of the tutor is through asking questions. By asking deeper questions, tutees may stimulate a more enriched discussion and higher quality tutor explanations, which should facilitate learning.

Because the episodes used in previous analyses could contain multiple tutee questions, we re-segmented the dialogue tutoring protocol data using "question-response exchanges" as the unit of analysis. A "question" was defined as an interrogative statement in which the tutee requested information (or verification of information). For the purposes of this paper, we excluded questions that were not directly relevant to the content (i.e. questions about task procedures or off-topic issues were not counted). A "response" was defined as any information or feedback (or lack thereof) provided by the tutor in answer to the question.

Tutee questions were then labeled as either "shallow" or "deep." A deep question was one that either required the tutor to generate an inference or contained a tutee-generated inference that the tutor had to evaluate. A shallow question was one that did not contain or require any information beyond the text contents. Tutor responses to these questions were similarly coded as "shallow" or "deep," depending on whether they contained inferences or novel elaborations of the text. Tutor responses were further classified as being "metacognitive" or "non-metacognitive," based on whether they contained self-monitoring statements (a statement such as "I don't know that" or "This is easy to remember").

Out of a total of 240 content-relevant questions asked across the seven dialogue tutoring pairs, 37% (88 questions) were classified as deep and 63% (152 questions) were shallow. Our results indicated that shallow questions were much more likely to receive a shallow response, but deep questions were equally likely to elicit a deep or shallow response (Table 5). In other words, deep questions were more likely to receive a deep response (41%) than were shallow questions (14%). It was fairly rare for a question to be ignored (receive no response). The overall pattern was significant; $\chi^2(2,N=240 \text{ questions}) = 26.1$, $p<.001$.

Table 5: Tutee questions and subsequent shallow or deep tutor responses.

Question Depth	No Response	Shallow Response	Deep Response	Totals
Shallow	15 (10%)	116 (76%)	21 (14%)	152
Deep	12 (14%)	40 (46%)	36 (41%)	88

Analyses of self-monitoring in tutor responses to tutee questions showed a similar pattern (Table 6). Shallow tutee questions tended to elicit non-metacognitive responses. However, the tutees' deep questions elicited metacognitive responses from the tutors about half the time. This pattern was significant; $\chi^2(1,N=240 \text{ questions}) = 20.8$, $p<.001$.

Table 6: Tutee questions and subsequent metacognitive or non-metacognitive tutor responses.

Question Depth	Metacognitive Response	Non-Metacognitive Response	Totals
Shallow	32 (21%)	120 (79%)	152
Deep	42 (48%)	45 (52%)	88

In order to confirm that tutor responses to deep questions were both deep and metacognitive (rather than one or the other), we cross-tabulated tutors' shallow versus deep and metacognitive versus non-metacognitive responses (Table 7). This analysis generally confirmed that deep, inferential responses were more likely to contain self-monitoring statements. Shallow responses were more likely to be non-metacognitive. This pattern was significant; $\chi^2(1,N=212 \text{ responses}) = 43.3$, $p<.001$.

Table 7: Tutors' deep and metacognitive responses

Response Type	Metacognitive Response	Non-Metacognitive Response	Totals
Shallow	29 (19%)	127 (81%)	156
Deep	37 (66%)	19 (34%)	56

In summary, the nature of the tutees' questions had an substantial impact on the subsequent integrative and metacognitive activities of the tutors. When tutees asked shallow questions, the tutors responses were frequently shallow and non-metacognitive. However, when tutees asked deep questions that contained or required an inference, the tutors were more likely to respond with a deep and metacognitive response.

Examples of Question-Response Exchanges

The following excerpts demonstrate how these processes occurred in a tutorial session. In the first example, a tutor and tutee are discussing the blind spot in the retina. The tutor summarizes background information and the tutee follows up with a deep question that leads the tutor to generate a novel analogy. The text provided only a structural description of the blind spot with no analogies.

Tutor: This is the blind spot [points to diagram]. You can't see anything there because that's where the optic nerve leaves the eye. So there aren't receptors right there. (paraphrase)
Tutee: Okay, wait. The blind spot is where all the nerves are located? (shallow question)
Tutor: Yeah. Like, that's where all of the optic nerves come together. They go all around and that's where they all pull together and go back to the eye. Or back to the brain. So right there, there aren't any receptors. (shallow, text-based response)
Tutee: So how does that affect your vision? (deep question)
Tutor: If something comes in and your lens refracts it to that point then you don't see it. (new inference)
Tutee: Oh, okay.
Tutor: So, it's just like when you're driving and there's that little spot in the mirror where you just won't see the person behind you. It's like that, except for the eyes. (deep response; novel analogy)

In the second example, a tutor and tutee are talking about the relationship between the iris and the pupil. The tutee's deep question causes the tutor to engage in sense-making activity, drawing on her prior knowledge in order to visualize and better understand these eye components. The text only discussed how the iris/pupil regulates the amount of light that can enter the eye, but did not describe how the iris reacts to light.

Tutor: The iris is the colored part of your eye. And it can expand or contract radially or circularly (paraphrase).
Tutee: What's radially? Like outward? (shallow question)
Tutor: Um. It explains that on the next page [skims text]. Yeah. That's outward. And when the radial muscles contract, the pupil gets larger. (shallow, text-based response)
Tutee: Okay. So, pretty much… contract is to make it smaller. So wouldn't the iris get smaller? (deep question)
Tutor: *Oh*. That makes so much sense now. Yeah. Like when your iris gets smaller, your pupil gets bigger. Like when someone's coming out of dark room or they get surprised. Your pupil gets really big and your iris gets really small. (new inference; draws on prior knowledge to visualize)
Tutee: Mm hmm.

Missed Opportunities for the Tutors

It is important to note that the mapping between tutee question quality and tutor response quality was far from perfect. About half of the tutees' deep questions failed to elicit a deep, metacognitive response from the tutor.

There are several potential explanations for this problem. One explanation is that the tutee's deep question contained an obvious inference and the tutor did not feel it was necessary to elaborate. Another explanation is that the tutor evaded the question because he or she did not have the requisite knowledge to answer it. A third reason might be that the tutor did not recognize the depth of the tutee's question. Chi, Siler, & Jeong (in press), have shown that even adult, non-peer, tutors often fail to diagnose a tutee's understanding. In all cases, the tutors miss out on an chance to build on their existing knowledge, fill knowledge gaps, or remediate errors – to learn, in other words.

The following excerpt provides an example of one of these missed opportunities. In this example, a tutor and tutee are discussing light refraction and the role of the cornea and lens in that process. The tutee asks two deep questions about the function of the cornea. Unfortunately, the tutor cuts this potentially productive exchange short rather than attempting to repair his knowledge gap.

Tutor: I'm going to talk about refraction, which is bending of the light. Most of it is done with the cornea [points to diagram]. But there's additional light bending done through the pupil. Or through the lens, I mean. And this is changed by altering the thickness of the lens. (paraphrase)
Tutee: The cornea doesn't change at all? (deep question)
Tutor: The cornea just stays the same. (new inference)
Tutee: Okay. Then how is it responsible for 70% of the focusing power? (deep question)
Tutor: I don't know. It doesn't say. (expresses ignorance and misses opportunity to repair this knowledge gap)

Conclusion

Previous research has established that students benefit academically from teaching other students. These learning outcomes have most often been attributed to the tutors' generation of instructional explanations and metacognitive self-monitoring while teaching. However, these mechanisms have been relatively understudied outside of reciprocal tutoring settings, which confound the benefits and processes of tutoring and being tutored. The analyses presented in this paper provide some converging evidence from a non-reciprocal tutoring setting that students learn by teaching due to explaining and self-monitoring activities. In addition, these behaviors were unstructured, indicating that tutors can learn even without a great deal of support and training (although well-structured interventions probably support more efficient and consistent learning behaviors).

Our findings show that the tutees played a very important role in shaping the learning activities and learning opportunities of the tutors. Although tutors paraphrased the text and introduced many of the topics discussed in the tutoring sessions, tutees stimulated much of the reviewing activity in which topics were covered more thoroughly. Tutees also directly and indirectly facilitated sense-making activities in which the tutors became aware of their own misconceptions and then attempted to repair them. These

elaborated reviewing and sense-making activities were likely guided by the kinds of questions that the tutees asked. Shallow questions tended to receive shallow and non-metacognitive responses from the tutors. However, deep questions asked by the tutees provided an important (if not always consistent) impetus for integrating ideas, generating inferences, and self-monitoring. More research is needed to understand how and why "missed opportunities" occur.

Acknowledgments

Funding for this research was provided by a grant awarded to Rod Roscoe by the University of Pittsburgh, FAS Office of Graduate Studies, and in part by the National Science Foundation, Grant Number NSF (LIS): 9720359, to the Center for Interdisciplinary Research on Constructive Learning Environments (CIRCLE, www.pitt.edu/~circle).

The authors would like to thank Marguerite Roy, Robert G. M. Hausmann, and several anonymous reviewers for their feedback and advice.

References

Ainsworth, S. & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*, 669-681.

Allen, V. L. & Feldman, R. S. (1973). Learning through tutoring: Low-achieving children as tutors. *Journal of Experimental Education, 42*(1), 1-5.

Annis, L. F. (1983). The processes and effects of peer tutoring. *Human Learning, 2*, 39-47.

Bargh, J. A. & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology, 72*(5), 593-604.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology*. Mahwah, NJ: Erlbaum.

Chi, M. T. H., deLeeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanation improves understanding. *Cognitive Science, 18*, 439-477.

Chi, M. T. H., Siler, S. A., & Jeong, H. (in press). Can tutors monitor students' understanding accurate? To appear in *Cognition and Instruction*.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Cloward, R. D. (1967). Studies in tutoring. *Journal of Experimental Education, 36*(1), 14-25.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237-248.

Coleman, E. B. (1998). Using explanatory knowledge during collaborative problem-solving in science. *Journal of the Learning Sciences, 7*(3), 387-427.

Coleman, E. B., Brown, A. L., & Rivkin, I. D. (1997). The effect of instructional explanations on learning from scientific texts. *Journal of the Learning Sciences, 6*(4), 347-365.

Fantuzzo, J. W., King, J. A., & Heller, L. R. (1992). Effects of reciprocal peer tutoring on mathematics and school adjustment: A component analysis. *Journal of Educational Psychology, 84*, 331-339.

Fantuzzo, J. W., Riggio, R. E., Connelly, S., & Dimeff, L. A. (1989). Effects of reciprocal peer tutoring on academic achievement and psychological adjustment: A component analysis. *Journal of Educational Psychology, 81*(2), 173-177.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Phillips, N. B., Karns, K., & Dutka, S. (1997). Enhancing students' helping behavior during peer-mediated instruction with conceptual mathematics explanations. *Elementary School Journal, 97*(3), 223-249.

Graesser, A. C. & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104-137.

Greenwood, C. R., Carta, J. J., & Hall, R. V. (1988). The use of peer tutoring strategies in classroom management and educational instruction. *School Psychology Review, 17*(2), 258-275.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*(2), 338-368.

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring interaction to scaffold peer learning. *Journal of Educational Psychology, 90*(1), 134-152.

Morgan, R. F. & Toy, T. B. (1970). Learning by teaching: A student-to-student compensatory tutoring program in a rural school system and its relevance to the educational cooperative. *Psychological Record, 20*, 159-169.

Palincsar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

Rekrut, M. D. (1992). Teaching to learn: Cross-age tutoring to enhance strategy instruction. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA, April.

Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction, 5*, 21-36.

Webb, N., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87*(3), 406-423.

A Brief Introduction to the Guidance Theory of Representation

Gregg Rosenberg (ghrosenb@ai.uga.edu)
Artificial Intelligence Center, University of Georgia
Athens, GA 30602 USA

Michael L. Anderson (anderson@cs.umd.edu)
Institute for Advanced Computer Studies, University of Maryland
College Park, MD 20742 USA

Abstract

Recent trends in the philosophy of mind and cognitive science can be fruitfully characterized as part of the ongoing attempt to come to grips with the very idea of *homo sapiens*—an intelligent, evolved, biological agent—and its signature contribution is the emergence of a philosophical anthropology which, *contra* Descartes and his thinking thing, instead puts doing at the center of human being. Applying this agency-oriented line of thinking to the problem of representation, this paper introduces the guidance theory, according to which the content and intentionality of representations can be accounted for in terms of the way they provide guidance for action. We offer a brief account of the motivation for the theory, and a formal characterization.

Introduction and Background

Recent trends in the philosophy of mind and cognitive science can be fruitfully characterized as part of the ongoing attempt to come to grips with the very idea of *homo sapiens*—an intelligent, evolved, biological agent—and its signature contribution is the emergence of a philosophical anthropology which, *contra* Descartes and his thinking thing, instead puts doing at the center of human being. Work that falls under this broad umbrella includes accounts of human cognition which stress embodiment and environmental situatedness (Anderson, 2003; *forthcoming*-a; Ballard et al., 1997; Clancey, 1997; Clark, 1995; Varela, Thompson, & Rosch, 1991), pragmatic and evolutionary accounts of human knowledge and culture (Barkow, Cosmides, & Tooby, 1992; Guignon, 1983; Hacking, 1983; Munz, 1993; O'Donovan-Anderson, 1997; Rescher, 1990) and action-oriented accounts of perception (Aloimonos, 1992; Ballard, 1991; Gibson, 1966; Milner & Goodale, 1995; O'Regan & Noë, 2001), to name only a few categories, and a few of the many works in each. The current essay introduces the results of our effort to build a theory of representation on the basis of the same kind of agency-oriented approach. It is *only* an introduction, and many difficult issues will have to be treated briefly, or not at all. The interested reader is encouraged to consult the fuller treatment given in (Rosenberg & Anderson, *forthcoming*).

A representation is something that stands in for, is in some sense *about*, something else. How is one thing ever *about* another? To answer this question is usually to analyze this relation of aboutness—the intentionality of a representation—in terms of some other, presumably more basic relation. For instance, a typical causal theory of representation might hold that a given representation **R** is about **E** just in case it has a certain specified set of causal relations to **E**, for instance, that perceiving an instance of **E** will cause one to represent with **R** (Fodor, 1981; 1987). Likewise an information-content approach might hold that a given representation is about that object from which the information it contains in fact derived (Dretske, 1981; 1986; 1988). Conceptual role theories, on the other hand, try to analyze meaning in terms of the role played by the concept in inferential and other conceptual/cognitive processes: roughly speaking, the representation **R** is about **E** just in case it is used to make warranted inferences about **E** (Harman, 1982; 1987). Naturally, there are also theories that try to combine these two approaches, producing the so-called "two-factor" accounts (Block, 1986; Loar, 1981; Lycan, 1984). There is no need, nor is this the place, to rehearse the standard critiques of these various theories (but see Anderson, *forthcoming*-b). However, by way of situating and introducing our own account of representational content, let us say that we find the various causal approaches too *input focused*, meaning they give too much importance to the ways in which the environment affects the organism to endow its states with representational meaning, and while the conceptual role theories seem to us a step in the right direction in that they draw attention to the importance of cognitive actions taken by the subject with its representations, *none* of the theories outlined above give sufficient weight to the full range of what a subject *does* with its representations.

In contrast, we ask first not what a representation *is*, but what it *does* for the representing agent, and what the agent does with it; what is a representation *for*? Our contention is essentially that representations are what representations do, and that what a representation *does* is provide guidance for action. Whatever the details of its instantiation or structure, whatever its physical, informational, or inferential features (and these are quite various across different representing systems), what makes a given item *representational* is its role in providing guidance to the cognitive agent for taking actions with respect to the represented object. In our view, each of those other special features a given representing token might possess—e.g. co-variance with,

openness to the causal influence of, or resemblance to its object—correspond to one of the range of strategies that our various representation-forming and representation-consuming systems have evolved to solve the biologically fundamental problem of providing autonomous organisms with guidance for action.

On the guidance theory action is fundamentally intentional: it is first and last a directed engagement with the world. Our basic claim is that representations come into existence and derive their content from their role supporting the basic intentionality of action. The fact that subjects take action with respect to things is what confers content on representations; it is how representations reach outside the organism and touch things in the world. The guidance theory presumes, then, that the intentionality of representation can be grounded in the intentionality of action.

A Formal Account of the Guidance Theory

Let us say that a token *provides guidance* to a subject by making its features available to the subject's motor systems and rational control processes for use in making discriminating choices between possible actions or possible ways of executing actions. Below we introduce the foundations of the guidance theory in terms of a set of propositions, which together characterize the most central features of the theory.

(1) An ***entity*** is anything that can be represented: a property, a concrete particular, an aspect of a thing, a state of affairs, a number, etc.

(2) A ***subject*** is any representation-consuming cognitive engine. To be a representation consumer, it must be capable of interacting in the world in a rational, goal directed way due at least partly to guidance it receives from tokens within its cognitive systems.

(3) A ***circumstance*** is a circumstance of the subject. A circumstance consists in the subject's internal states, including the subject's bodily changes, registrations, representations, expectations, priorities, values, options for action, homeostatic self-evaluations, procedural knowledge, motor schemas and also the subject's immediate environment.

(4) A subject ***standardly uses*** tokens (of a type) to provide guidance with respect to an entity **E** in a given (type of) circumstance **C** if, and only if, the subject has an enduring conscious preference or conditioned reflex to use the tokens (i.e., members of the type) to provide guidance with respect to **E** when in circumstance **C**.

(5) An *action* can be a motor process or a cognitive process. This yields two clauses in the definition of action:

> **(5.1)** In the case of a motor process, a motor process is an *action* if, and only if, it is activated under control of perceptual/cognitive feedback processes capable of effectively modulating or bringing about changes in the organism or in the world
>
> **(5.2)** In the case of cognitive processes, a cognitive process is an *action* if, and only if, it is a mental process under intentional control whose results contribute to circumstances (as defined above) used to direct motor processes. A cognitive process is under intentional control if the working of that cognitive process is subject to modification by processes of attention, short-term memory, valuation, assent and dissent, practiced learning, and consciously administered self-criticism and praise.

As mentioned already, the fact that subjects take action with respect to things is what confers content on representations; it is how representations reach outside the organism and touch things in the world. The central importance of the intentionality of action means that it is vital to correctly understand—without regress—what it is for an action to be taken with respect to something.

(6) An action is taken ***with respect to an entity*** **E** if, and only if,

> **(6.1)** The action is a motor program, **E** is the focus of the intended change or efforts at control in the world; or
>
> **(6.2)** The action is a motor program and an assumption of information about **E** is a motivating reason that the given action, rather than some alternative non-**E** involving action, was undertaken; or
>
> **(6.3)** The action is a cognitive process undertaken to discover or confirm facts, to modify values, or to decide between alternative actions, and an assumption of information about **E** is necessary if the process as a whole is to provide guidance for the subject's motor actions.

This definition uses three further terms—*motivating reason*, *focus*, and *assumption of information*—that present the potential for regress and require further discussion.

Motivating Reason

For an account of motivating reason, we hold only that any analysis must be such that it would be applicable to goal-directed behavior of entities that do not have representations at all. For example, it must be of a piece with how we would identify the motivating reasons for why a plant turns toward the sunlight. The plant's behavior is goal-directed behavior even if it is not action in the sense defined above, and the motivating reason for the behavior is to maximize the amount of sunlight available for photosynthesis. Because the plant does not have representations, a correct account of motivating reason cannot appeal to representational content.

We also distinguish motivating reasons from applications of causal force. A child may go to bed early on Christmas Eve to encourage Santa Claus to bring presents, and this may be the child's motivating reason, even though Santa Claus is not capable of applying causal force on the child's mind. A hungry wolf may look for prey and its motivating reason may be a future state of satiety, even if the cause of its behavior is a present internal state. Any account of motivating reasons must allow for motivating reasons that are non-representational facts and entities, even for agents that possess representations.

At its heart, the concept of a motivating reason is deeply tied to concepts of rational interpretation like the one found in Daniel Dennett's description of the intentional stance (Dennett, 1987). We take no position here on the basis of, or constraints on, any specific standards of rational interpretation.

Focus

As it is used above, the idea of an action's focus is intended to express a functionalist concept. When a subject is performing an action it places itself into a potential feedback loop with its environment. Its purpose is to monitor the result of the action and to plan adjustments to its course of action.

(7) The ***focus*** of an action is the ultimate entity being monitored through the feedback channels taken to provide indications of its status.

A subject may monitor the focus directly, or indirectly by monitoring the status of some entity being used as an indicator of facts about the focus. Because indicators are made part of an extended guidance control system, indications about the focus will cause in the subject beliefs, decisions or equivalent states about further appropriate actions or perhaps that action may cease. When the focus is monitored through an indicator the subject may have an indirect causal connection to the focus or even no causal connection at all. An example of an indirect causal connection to a focus would be an engineer monitoring a gauge that is itself monitoring engine pressure. Examples of focii to which there is no causal connection are things like the time of day or a mathematical operation on numbers. To monitor the first we might monitor an indicator like a clock face and to monitor the second we might monitor a progression of numerals manipulated according to established rules. In both of these cases the focus of the action is something that is not present and to which the subject is not even indirectly causally connected, but which can be monitored nevertheless, despite the lack of causal causation, by establishing a connection to something else that can be manipulated to vary systematically with facts about the focus.

Identifying the focus of an action in a given case requires establishing the facts about what the subject is monitoring in its circumstances, and understanding these facts in terms of the subject's motivating reasons.

Assumption of Information

An assumption of information is to be cashed out in terms of facts about the actual operation of the representing system (or subject). Beginning with an example will make the concept easier to grasp. Imagine a computer processing a user's command to print a document. To do this, the computer must determine to which printer it should send its own commands. To guide this action, the computer reads several character strings contained on its hard disk, one identifying the printer and others with other information about the printer. These strings guide it regarding where it should send its print commands and what protocol it should use to communicate with the printer. From the perspective of the guidance theory, here is the key fact: these character strings represent what they do both because of the circumstances in which the computer is reading them and also because of the *assumptions* built into those circumstances. The computer processes the strings *as if* they conveyed information about the printer to which it sends its commands and which communication protocol it should use. There is no regress involved in claiming it makes this assumption, because the assumption itself is not a matter of having representational content. There is no representation inside the computer with the content: *I assume that this string has information about the printer*. Even more strongly, its ability to make an *assumption* of information does not require that the computer actually *possesses* information, nor that it ever did.[1] In the case described, the character string the computer accesses could have been placed on the disk via the output of a random number generator and by coincidence be effective in directing it to the printer. Even were that to be true, the string still would be providing guidance and the computer would still be making an assumption that the string contained information about the correct printer. Therefore, the ability to make an assumption of information does not require an ability to have or obtain information.

Rather, the assumption of information about the printer is a matter of *know-how* that is built into the architecture of the computer: how it accesses representations, in what circumstances it accesses them, how it reads and interprets their structure, what actions it initiates and monitors upon accessing them, how those actions cause it to interact with the world, and so forth. We can provide a candidate analysis of this *know-how*. To do this, we first need to define, for any given token, the class of actions it supports. The class of actions a token **T** supports is relative to the kinds of circumstances **C** where the system is prepared to use the token for guidance. It consists of all the actions the system can initiate or modulate in **C** due to its processing of **T**. Let us label this class of supported actions $\mathbf{A}_{\mathbf{supp}}$.

(8) An action **A** is a member of the class of actions, $\mathbf{A}_{\mathbf{supp}}$, supported by a token **T** used by a subject **S** in circumstances **C** if, and only if, **S** in **C** would use **T** for guidance regarding the initiation or manner of execution of **A**.

We should think of the actions in $\mathbf{A}_{\mathbf{supp}}$ as focus-neutral descriptions of an action in need of association with a focus in particular initiations. So, for example, if in some circumstances a system is prepared to use a token for guidance in running, the action *running* is the focus neutral description. If the specific initiation of this action occurs when the focus of the action is a bear, the focus-neutral action "running" is initiated as the focus-specific action "running away from a bear." Actions obtain a focus in the way discussed above.

Furthermore, since subjects do not initiate actions at random, for each action in $\mathbf{A}_{\mathbf{supp}}$, there will be a (possibly

[1] This, assuming that possessing information depends on causal history and connection, which may not be the case.

very large but) finite set of circumstances capable of triggering the initiation of the action. We can call this set of triggering circumstances $\mathbf{A_{circ}}$. The number of triples $\mathbf{<A \in A_{supp}, C \in A_{circ}, Focus>}$ representing supported actions **A** initiated in circumstances **C** with focus **Focus** provides a class of counterfactual *action scenarios*, $\mathbf{A_{scene}}$, in which the token **T** provides guidance for a subject. These are the action scenarios in which **T** *participates*.

Most actions are complex, both in the sense that they have many different specific features that must be managed (e.g., the trajectory and velocity of a running motion), and in the sense that they almost always require initiating smaller or tangential actions involving entities besides its focus if they are to succeed in affecting their intended change or control (e.g., jumping over the branch on the ground while running from the bear). Because of the complexity of action, subjects needing to execute an action will almost always use representations other than the tokens representing the focus of the action. In fact, activation of these further tokens is necessary to fill out the circumstances in which all the tokens are used.

These other active representations will fall into several categories: conscious representations with foci of their own serving the larger action program; unconscious but potentially conscious representations supporting the interpretation of the circumstances and manner in which the action is executed; and sub-conscious representations that can never be conscious but that provide support for basic perception, adjusting bodily movements, and triggering emotion. We should construe the entities towards which the supporting tokens provide guidance as sub-foci in sub-actions lying under the umbrella of the main action. Therefore, these further tokens, the ones that support the guidance for the main action within a given $\mathbf{A_{scene}}$, have functional roles determined by their potential relationships to their own foci within the circumstances **C** of $\mathbf{A_{scene}}$.

Relative to these action scenarios, the guidance theory supposes that in each $\mathbf{A_{scene}}$ where an active token succeeds in having reference[2] the token can be mapped to an entity through its functional role under the rational constraints associated with assigning motivating reasons to their sub-actions. This supposition is justified because, in providing guidance, a token will make features of itself available to the subject, which the subject can use to differentially control its actions with respect to an entity which is a focus or sub-focus of a given action.

The know-how involved in an assumption of information, then, is a question of the way that the subject's decoders and action mechanisms process and/or respond to representations (i.e., how it accesses representations, in what circumstances it accesses them, how it reads and interprets their structure, what actions it initiates and monitors upon accessing them, how those actions cause it to interact with the world, and so forth) given the subject's capabilities, needs, environment, and cognitive architecture. The general idea is that assumptions of information consist in non-representational facts about how the subject *works*, not in further representational facts about, or representations used by, the subject. Although this account is clearly preliminary, it does at least show how the idea of an assumption of information can be interpreted, and used as part of the machinery involved in determining the content of a representation, without initiating a vicious regress or involving circular appeals to representational content.

This brings us, finally, to the cumulative definition of representation. On the guidance theory, *representation* is simply tracking in the sense defined below:

(9) A token **T** *tracks* an entity **E** for a subject **S** in token circumstances **C** if, and only if, **T** is standardly used to provide guidance to **S** for taking action with respect to **E** in **C**.

(10) A token **T** *represents* an entity **E** for a subject **S** in token circumstances **C** if, and only if, **T** tracks **E** for **S** in **C**.

By linking representation to guidance in this way, the guidance theory distributes responsibility for the existence of representational content across a representational token (the representation) and an interpretative decoding mechanism (the decoder) integrated with a subject's action-determining processes. The effect of distributing responsibility is to introduce new degrees of freedom regarding the exact physical or informational requirements for something to be a representation, as the requirements on the representation will depend on the capabilities of the decoder and the circumstances in which it is used. In general, the demands on each part of the coupled system vary inversely with the demands on the other. A representation that is highly structured and closely coupled with what it represents needs a less sophisticated decoding mechanism, while a very sophisticated (or very rigid and simple) decoding mechanism may embody (or presume) so much implicit domain knowledge that it can get by with very sparse representations.

Representation and Misrepresentation

One of the most important problems that any theory of representation must solve is the problem of normativity: representations are assessable for accuracy, and therefore they can be in error. To be complete, the guidance theory must account for this feature of representations. Because the guidance theory is an action-based theory of representation, the natural thing to do is to base error on the failure of action and the way that a representation's guidance contributes to that failure. The intuitive idea, then, is that a representation is in error if it provided guidance to an action that failed in its intent, and it failed partly or wholly because of the guidance provided by that representation. This intuitive idea can be formalized as follows:

(11) An action *fails in its intent* if, and only if,

(11.1) It is a motor action and the intended change is not achieved or the intended process is not brought under control; or

[2] The concept of error will be defined formally in the next section.

(11.2) It is a cognitive process and it (a) confirms a representation that is in error[3]; or (b) disconfirms a representation that is not in error; or (c) modifies a value in a way that the subject later regrets; or (d) recommends a course of action that fails.

(12) An action **A** fails in its intent *because of **R*** if, and only if, (a) **A** failed; and (b) **A** was taken with respect to an entity **E**; and (c) **R** provided guidance for **A** w.r.t. **E**; and (d) **R** has feature **F**; and (e) **R** with **F** represents that **E** has property **P**; and (f) **A** failed because **E** was not **P**. Note that the term "represents" in clause (e) is to be read in light of the current theory of representation.

(13) A token representation **R** *is in error* for subject **S** and action **A** in token circumstances **C** if, and only if, **A** would fail because of **R** if taken by **S** in **C**.

The representation may be said to be in error for **S** *simpliciter* if and only if the class of actions for which **R** provides guidance in **S**'s circumstances **C** is dominated by actions that would fail because of **R**.

Comparison to Related Work

The guidance theory, broadly speaking, takes both a naturalistic and a functional perspective on representation. It is motivated by the same fundamental insight regarding the epistemic importance of action and interaction as gave rise to the theory of interactive representation (Bickhard, 1993; 1999). However, we offer a significantly different development and formalization of this shared insight. For instance, Bickhard's analysis relies heavily on control theory, cashes out representational content in terms of 'environmental interactive properties', and assumes some version of process ontology. The guidance theory, while compatible with these possibilities, does not require them. Still, the relative advantages of these two analyses remain largely to be determined. While there are many other naturalistic theories of representation on offer, very few adopt the functional perspective in as thoroughgoing a way as we do. For instance, Dretske (1986; 1988) adopts the functional perspective largely as a post-hoc fix to what remains an information-content approach to representation, so as to be better able to account for *mis*information. In contrast, Ruth Millikan does take the functional perspective as the starting point for her theory of representation, and the guidance theory thus bears the most resemblance to hers (Millikan, 1984; 1993). Thus, although the current article is meant only as a concise introduction to the guidance theory, and is not the place for any detailed comparisons with rival theories, it is nevertheless worthwhile to say a few words about Millikan's theory in particular.

The resemblance between the guidance theory and Millikan's own biologically inspired theory is strongest when she writes things like: "Cognitive systems are designed by evolution to make abstract pictures of the organism's environment and to be guided by these pictures in the production of appropriate actions." (Millikan, 1993:11) However, the impression of similarity fades quickly as the details are examined. For while we agree on this very general characterization of cognitive systems, we differ as to the core point: that mental representations must be pictures and, even when they are pictures, we differ as to what makes such "abstract pictures" *representations*.

There are three main components to this very basic disagreement. First, on our view, a given mental token is a representation just in case it is standardly used by a given organism to guide its behavior with respect to the intended object; Millikan, in contrast, suggests that it is only a representation if it is the result of (or consumed by) a properly functioning system, performing the function it was selected to perform: "It is not the facts about how the system *does* operate that make it a representing system and determine what it represents. Rather, it is the facts about what it would be doing if it were operating according to biological norms." (Millikan, 1993:10-11)

Second, and deeply related to the first, Millikan relies heavily on the notion of such a "proper function" to explain the possibility of representational error (a representation is in error when the relevant representation-producing or representation-consuming system is not functioning according to biological norms). In contrast, our theory allows for the possibility that a system serving some function other than that for which it was selected, or mal-functioning in some very lucky way, could, in its use of mental tokens, be *representing* just in case (roughly speaking) the mental tokens in question were being used to (successfully) guide the agent's actions with respect to the indicated objects. Rather than analyze representational error in terms of mal- or non-standardly-functioning systems, we cash it out in terms of failure of action. Although we think representational systems *did* evolve, and attention to their evolutionary history can help us understand how and why they function as they do, we believe a system can sometimes competently perform a function, including representing, for which it was not selected, and in these cases its unusual provenance should be no barrier to recognizing this fact.

Third (and finally), whereas Millikan's view of behavior and action revolves around the function or purpose of the organism or its parts (a movement by the organism is only a behavior of that organism if it can (or perhaps must) be understood in terms of the organism's proper function or biological purposes), our own definition of action includes motor and cognitive processes effected for a broader range of motivating reasons. Although some element of teleology is apparently necessary to ground the idea of a motivating reason for acting, it is not clear to us that this must necessarily be accounted for in terms of natural selection. It could be the teleology of the subject itself, understood as having a subjective purpose like maintaining its homeostatic condition, pursuing hedonic value, or maintaining adherence to a moral, political, or aesthetic principle. A more detailed discussion relating the guidance theory to some alternative theories, including Millikan's, can be found in (Anderson, *forthcoming*-b).

[3] This clause in the definition is an embedded recursion, not a circularity.

Conclusion

The guidance theory is an action-focused theory of representation according to which content is derived from the role a representational vehicle plays in guiding a subject's actions with respect to other things. What qualifies an element of experience as a representation is, strictly speaking, only that the element of experience be capable of providing a subject with guidance for its actions with respect to entities. To be capable of providing guidance an element of experience only needs to have features useful for exploitation by the subject's action-producing mechanisms.

In the full formalization, we show that the guidance theory can account for various problem cases of representational content such as abstract, fictional and non-existent objects (Rosenberg & Anderson, *forthcoming*). Twin-Earth and swampman are discussed in (Anderson, *forthcoming*-b). Future work will consider the evolutionary development of representation in more detail, and the implications of the guidance theory for the correspondence theory of truth, for scientific realism, and for consciousness and phenomenal content (Rosenberg, 2004).

References

Aloimonos, Y. E. (1992). Purposive active vision. *CVGIP: Image Understanding*, *56*, 840-50.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, *149*(1), 91-130.

Anderson, M. L. (forthcoming-a). Cognitive science and epistemic openness. *Phenomenology and the Cognitive Sciences*.

Anderson, M. L. (forthcoming-b). Representation, evolution and embodiment. In: D. Smith (Ed.), *Evolutionary Biology and the Central Problems of Cognitive Science*, special issue of *Theoria et Historia Scientarum*, *9* (2005/1).

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral & Brain Sciences*, *20*(4), 723-767.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, *48*, 57-86.

Barkow, J. H., Cosmides, L. & Tooby, J. (Eds.) (1992). *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York, NY: Oxford University Press.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, *5*, 285-333.

Bickhard, M.H. (1999). Interaction and representation. *Theory & Psychology*, *9*(4), 435-458.

Block, N. (1986). Advertisement for a semantics for psychology. In P. French, T. Uehling & H. Wettstein, (Eds.), *Midwest Studies in Philosophy X*, 615-678. Minneapolis: University of Minnesota Press.

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge: Cambridge University Press.

Clark, A. (1995). *Being There*. Cambridge, MA: MIT Press.

Dennett, D.C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA, MIT Press.

Dretske, Fred (1986). Misrepresentation. In R. Bogdan, (Ed.), *Belief: Form, Content, and Function*. New York: Oxford University Press.

Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.

Fodor, Jerry A. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Guignon, C. (1983). *Heidegger and the Problem of Knowledge*. Indianapolis, IN: Hackett Publishers.

Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.

Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*, 242-56.

Harman, G. (1987). (Nonsolopsistic) conceptual role semantics. In E. LePore (Ed.), *Semantics of Natural Language*. New York: Academic Press.

Loar, B. (1981). *Mind and meaning*. London: Cambridge University Press.

Lycan, W. (1984). *Logical form in natural language*. Cambridge, MA: MIT Press.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press.

Milner, A. D. & Goodale, M. A. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Munz, P. (1993). *Philosophical Darwinism: On the Origin of Knowledge by Means of Natural Selection*. London: Routledge.

O'Donovan-Anderson, M. (1997). *Content and Comportment: On Embodiment and the Epistemic Availability of the World*. Lanham, MD: Rowman and Littlefield.

O'Regan, K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24* (5), 883-917.

Rescher, N. (1990). *A Useful Inheritance: Evolutionary Aspects of the Theory of Knowledge*. Lanham, MD: Rowman.

Rosenberg, G. (2004). *A Place For Consciousness: The Theory of Natural Individuals*. Oxford: Oxford University Press.

Rosenberg G. & Anderson M.L. (forthcoming). Content and action: The guidance theory of representation.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Educational Effects of Reflection on Problem Solving Processes: A Case of Information Seeking on the Web

Hitomi Saito (hsaito@auecc.aichi-edu.ac.jp)
Programs in Education for Information Sciences, Faculty of Education, Aichi University of Education
Kariya, 448-8542, JAPAN

Kazuhisa Miwa (miwa@cog.human.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University
Nagoya, 464-8601, JAPAN

Abstract

In this study, we design a learning environment that supports reflective activities for information seeking on the Web and evaluate its educational effects. The features of this design are: (1) to visualize the learners' search processes as described, based on a cognitive schema, (2) to support two types of reflective activities, such as "reflection-in-action" and "reflection-on-action," and (3) to facilitate reflective activities by comparing their own search processes to other learners' search processes. We have conducted an experiment to investigate the effects of our design. The experimental results confirm that (1) the participants' search performance in the instructional group supported by our instructional design improved effectively than in the control group, (2) they changed their ideas about important activities when seeking information on the Web, and (3) they activated their search cycles more than the control group did.

Introduction

In the field of learning science, many researchers have investigated metacognitive activities that facilitate learners' problem solving and deep understanding (Lin & Lehman, 2001). Metacognition is generally referred to as knowledge and activities to monitor, control, and manipulate individual cognitive processes (Brown et al., 1983). Several studies have shown that experts or good learners practice metacognitive strategies more actively than novices or poor learners (Chi et al., 1989; Ertmer, Newby, and MacDougal, 1996; Leinhardt & Young, 1996). Additionally, based on the findings from these studies, various systems or instructional designs that support learners' metacognitive activities have been developed, and their educational effects have been examined (Aleven & Koedinger, 2002; Hershkowitz & Schwarz, 1999).

Metacognitive activities to monitor and control individual cognitive processes are fostered by various activities connected with cognitive efforts, such as self-explanation, self-regulation, and reflection. We focus on reflective activities within these metacognitive activities. Reflection is defined as a cognitive activity for monitoring, evaluating, and modifying one's thinking and process (Lin, Kinzer, & Secules 1999). In this study, based on the standpoint that metacognitive activities help students learn with greater understanding, we examine effective methods for supporting reflective activities.

Lin et al. (1999) proposed that there are at least two levels of reflection in learning: reflection on a product and its value and reflection on a process by which the product was created. They suggested that supports reflection on a process is more important because the process is less explicit than the product for learners. Moreover, they identified a process display as one of the scaffolds that supports reflection on the processes. A process display shows learners explicitly what they are doing to solve a task or learn a concept. This method allows learners to observe and analyze their own problem-solving processes and evaluate the effectiveness of their learning. For example, Geometry Tutor, which was designed by Anderson, Boyle, & Reiser (1985) to help students learn geometry, displays learners' geometric reasoning processes as a proof graph that consists of tree diagrams of their own solution paths between the "given" and "goal" states of problem-solving. Schauble, Raghavan, & Glaser (1993) also developed the Discovery and Reflection Notation (DARN) system, which shows students a graphical trace notation to support students' reflection on their scientific reasoning with computer-based laboratories. Although many studies have developed systems that provide students with learning processes, the educational effects of reflection on the problem-solving processes are not clear. It is also necessary to examine how we should show learners their problem-solving processes and how learners should reflect on their problem-solving processes. In this study, we design a learning environment that supports learners' reflection on problem-solving processes when seeking information on the Web and evaluate its educational effects.

First, in order to show learners their problem-solving processes, we have developed a feedback system for search processes that provides learners with their own information-seeking processes, which are described based on a cognitive schema. In problem-solving studies, a cognitive schema has been widely used to describe human problem-solving processes. We use such a cognitive schema to visualize learner's problem-solving processes and provide them with learners. We then investigate whether a cognitive schema can be applied as a cognitive tool in learning science. We will explain our system and the cognitive schema in the next chapter. Second, in order to help learners reflect on their problem-solving processes more effectively, we focus on two types of reflective activities that are referred to as "reflection-in-

action" and "reflection-on-action," proposed by Schön (1987). Schön categorized reflection as "reflection-in-action" and "reflection-on-action" from the viewpoint of a context and time. The former refers to monitoring ongoing learning activities, while the latter means revisiting and monitoring critical events in one's own learning experiences after learning activities. Schön suggested that these two types of reflection are imperative factors for learning in any field with the purpose of effective learning transfer. In this study, we investigate an educational design to support these two types of reflective activities.

A Search Process Feedback System

We constructed a feedback system for search processes that supports learners' reflections on their problem-solving processes when seeking information on the Web. This system supports learners' reflection on their own search processes by (1) providing visual support for their search processes, and (2) prompting searchers to reflect on their search processes.

A Search-Process Describing Schema

The system describes learners' information-seeking processes on the Web based on a schema for describing search processes, and allows these processes to be shown in real time. The search-process description schema was proposed to analyze searchers' processes for seeking information on the Web (Saito & Miwa, 2002). This schema was constructed based on the Problem Behavior Graph (PBG), proposed by Newell & Simon (1972), which is well known as one of the most fundamental schema for describing the subjects' problem-solving processes.

Usually, we begin the search with a search engine when we want to find something on the Web. Following that, we consider keywords and search queries to input to a search engine, and browse the results of a search or each Web page. In this schema, a phase in which keywords and search queries are considered is defined as a search in the Keyword space, while a phase in which information on the Web, such as the results of a search and Web pages, is searched is defined as a search in the Web space. Furthermore, the Web space is subdivided into the Result-of-Search space and the Web-Page space. Figure 1 shows a sample description of the search-process description schema.

The searchers' processes are described as transitions of nodes and operators through these three search spaces. A node represents a searcher's behavioral state, and each node's components differ from space to space. In the Keyword space, a node consists of a serial number and search queries, a node in the Result-of-Search space consists of a serial number, search queries, and the number of search results page, and a node in the Web-Page space includes a serial number and the depth of links. An operator shows an operation to the node. The following six operators are defined in this schema:

Search: searching with a search engine

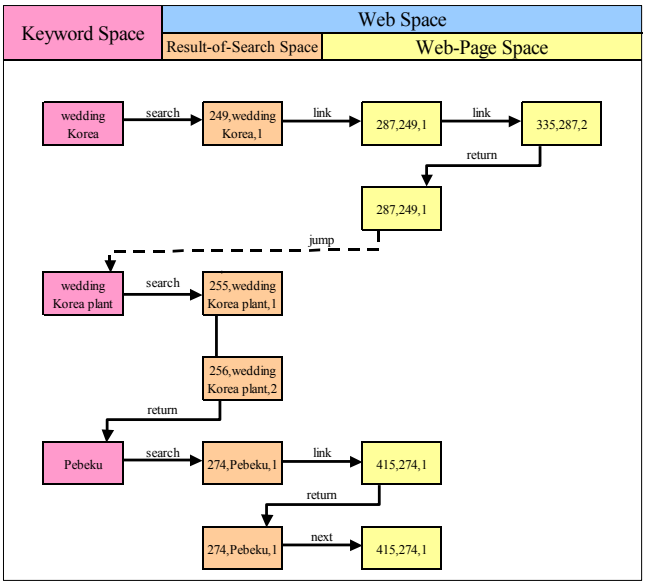


Figure 1: Sample description of the search-process description schema.

Link: going to a page connected with a link

Next: going forward to the next page after having gone backward

Return: going backward to the last page just visited

Jump: revisiting a page

Browse: browsing search results just obtained

Prompting

The system prompts questions to help learners reflect on their own search processes presented by the system. When the system prompts a question, learners are required to answer the question while referring to the learners' own search processes. Table 1 shows each type of question presented by the system. The following three types of questions were used: (a) questions on the Keyword space, (b) on the Result-of-Search space, and (c) on the Web-Page space.

Experiment

We have devised an instructional design that includes the search process feedback system as a core part of the design and two types of reflection. In this section, we conducted an experiment to evaluate how support for reflection affects learners' problem-solving processes and their search performance.

Participants

Thirty-eight university freshmen participated in our experiment as a part of a class. The participants were divided randomly into two groups. One group (the instructional group) was supported based on our instructional design, whereas the other (the control group) was

Table 1: Each type of prompt presented by the system.

Types of Prompts	Questions
Keyword Space	What kinds of keywords did you use, or how did you combine these keywords ?
Result-of-Search Space	How many results of search pages did you browse per search ?
Web-Page Space	How many links did you click on per page ?

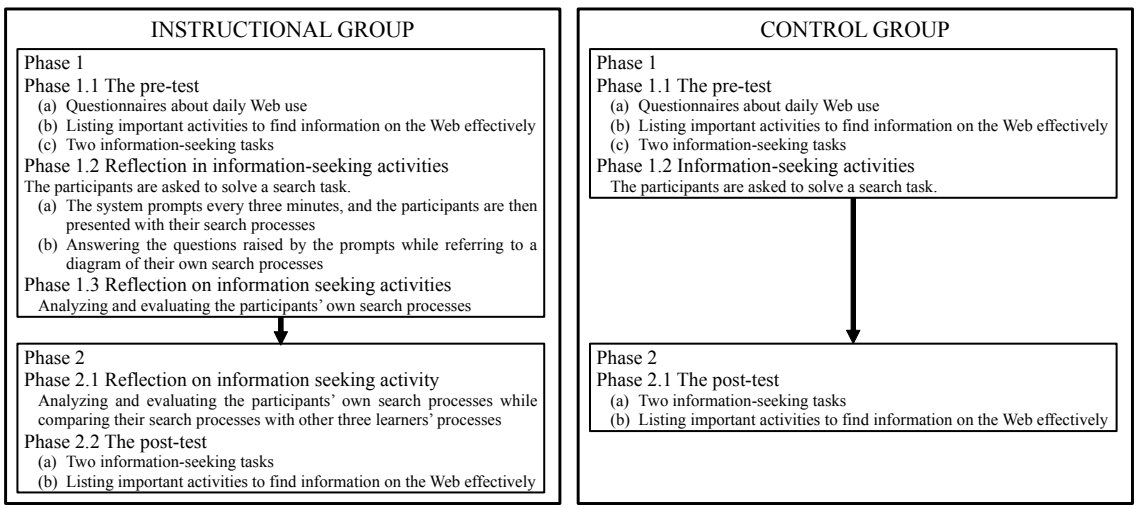


Figure 2: Summary of the experimental procedure.

not supported. The instructional group comprised 19 participants, as did the control group. We examined the participants' experiences of using the Web. The average time consumed per day was 26.5 minutes for the instructional group and 33.3 minutes for the control group. There was no significant difference between the two groups ($t(37) = .879, n.s.$).

The experiment consisted of two phases, which were separated by an interval of at least one day. Figure 2 shows a summary of the experimental procedure. In the following, we explain the experimental procedures.

Pre- and Post-tests

We conducted the pre- and post-tests to confirm whether the participants' search performance and their ideas about information seeking on the Web improve through their reflective activities. Each test consisted of (1) listing at least five important activities to find information on the Web effectively, and (2) solving two information-seeking tasks to measure the participants' search performance. In the information-seeking tasks, the participants were asked to find target information within ten minutes for each task, using a normal Web browser, where none of the participants were provided with their search processes. The tasks were counterbalanced between the participants.

The instructional group

Phase 1.2 Reflection in information-seeking activities In Phase 1.2, the participants in the instructional group experienced "reflection-in-action," wherein the participants reflect on their own search processes while seeking information on the Web. Following the pre-test, we explained to them the experimental task and how to use the system. Next, they were asked to solve a search task using the system. The search task lasted for about 20 minutes, and the participants in the instructional group were shown a prompt every three minutes then presented with their search processes described by the system. They considered the questions raised by the prompts while referring to a diagram of their own search processes, and entered their answers to the answer sheet.

Phase 1.3 Reflection on information-seeking activities In Phase 1.3, the participants in the instructional group experienced "reflection-on-action." After the search task, the participants reflected on their own search activities, analyzing and evaluating their own search processes for twenty minutes as instructed by an experimenter. First, they analyzed their search processes based on the perspective of a search among the three spaces (the Keyword space, the Result-of-Search space, and the Web-Page space) while referring to their own search processes. Second, they considered the advantages and disadvantages of their search processes and how to improve those disadvantages. Following that, they filled in their answer sheets with their ideas.

Phase 2.1 Reflection on information-seeking activities In Phase 2.1, the participants in the instructional group also experienced "reflection-on-action." In contrast to Phase 1.3, the participants reflected on their search activities through comparing their own search

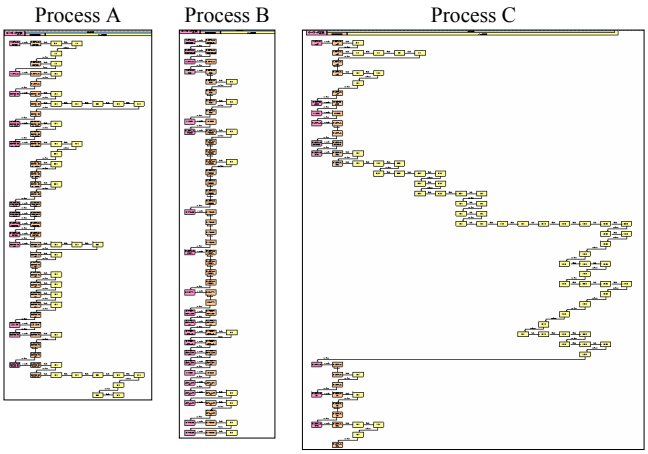


Figure 3: The three processes presented to the participants in the instructional group.

processes with the other three learners' processes that had been selected from the control group by one of the authors.

The presented three processes are shown in Figure 3. Process A is a process by a participant who found a correct answer. One feature of this process is that the balance of searching each space is relatively well coordinated (balanced search). Process B and Process C are processes of participants who could not find a correct answer. In contrast to Process A, these processes tend to cling to a search of one or two of the three spaces. The participant followings Process B hardly searched the Web-Page space at all. He or she repeatedly shuttled between searching in the Keyword space and the Result-of-Search space (breadth-first search).

The participant followings Process C searched the Web-Page space in great detail (depth-first search). The instructional group was provided with these three processes plus information on whether each participant found the correct answer. Then, they analyzed and evaluated their own search activities while comparing their own search processes to the three typical processes, just as in Phase 1.3.

The control group

The participants in the control group engaged in the pre- and post-tests and the search task in Phase 1.2. In Phase 1.2, the participants in the control group solved the search task without receiving the prompts and the presentation of their own search processes.

Effectiveness of the instructional design

In this section, we evaluate the effects of our instructional design based on the experimental results. We compare changes from the pre- to post-tests in the instructional group with those in the control group based on the following three points: (1) the participants' search performance, (2) their ideas about important activities in information-seeking on the Web, and (3) their search processes.

Three out of thirty-eight participants were eliminated because one did not understand the experimental instruction and the others did not participate in Phase 2. Therefore, we analyzed the results of the 35 participants: 17 participants from the instructional group and 18 participants from the control group.

Search Performance

The scores of the search tasks in the pre- and post-tests were estimated to determine whether the participants could locate Web pages containing the target information. The participants' performances in the pre- and post-tests are shown in Table 2. Each score (0, 1, and 2) shows the number of tasks in which the participants could find a correct answer, and each frequency in each cell of this table show the number of the participants getting each score. We compared the number of participants who increased their scores from the pre-test to post-test with the number of participants who did not.

From the result of the chi-square test, Groups (the instructional/control groups) × Performances (improving/not improving), we found that the number of participants who improved their search performance from the pre- to post-tests significantly differed for the two groups ($\chi^2(1) = 4.13, p < .05$). This result indicates that the participants, who engaged in reflective activities supported by our instructional design, improved their search performance more effectively.

Table 2: Participants' performances in pre- and post-tests.

(a) Instructional group

		Post Test			
		0	1	2	Sum
Pre Test	0	9	7	0	16
	1	1	0	0	1
	2	0	0	0	0
	Sum	10	7	0	17

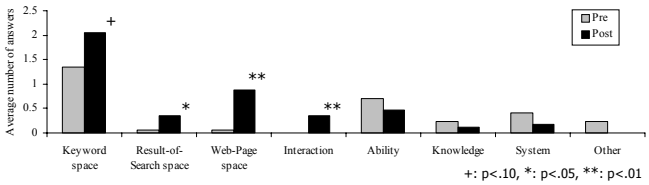
(b) Control group

		Post Test			
		0	1	2	Sum
Pre Test	0	11	1	0	12
	1	2	2	0	4
	2	0	0	0	0
	Sum	13	3	0	16

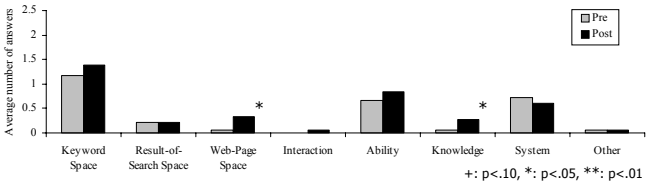
Important activities in information seeking on the Web

In the pre- and post-tests, the participants were asked to propose five activities that they considered important in information-seeking on the Web. The participants' answers in each test were categorized into the following eight types.

Keyword space: activities with search in the Keyword space



(a) Instructional group



(b) Control group

Figure 4: Average number of answers in each category in the pre- and post-tests.

Results-of-Search space: activities with search in the Results-of-Search space

Web-Page space: activities with search in the Web-Page space

Interaction: activities with transitions among multiple spaces

Ability: necessities of abilities and attitudes

Knowledge: knowledge required in information seeking on the Web

System: functions of a search system, such as a search engine

Figure 4 shows the average number of items in each category in the pre- and post-tests. In the instructional group, paired t-tests indicated significant differences in the increase of the number of items in "Results-of-Search space" ($t(16) = 2.582, p < .05$), "Web-Page space" ($t(16) = 3.846, p < .01$), "Interaction" among spaces ($t(16) = 2.954, p < .01$) and a slight difference in the increase of the number of items in "Keyword space"($t(16) = 2.073, p < .10$).

The items above were related to the search processes on which the participants reflected. On the other hand, in the control group, paired t-tests indicated significant differences in the increase of the number of items in "Web-Page space" ($t(17) = 2.557, p < .05$) and "Knowledge" ($t(17) = 2.204, p < .05$). These results indicate that the participants who reflected on their search processes in the instructional group acquired different notions as important concepts for the Web search than those in the control group; in particular, they realized their own search activities more profoundly.

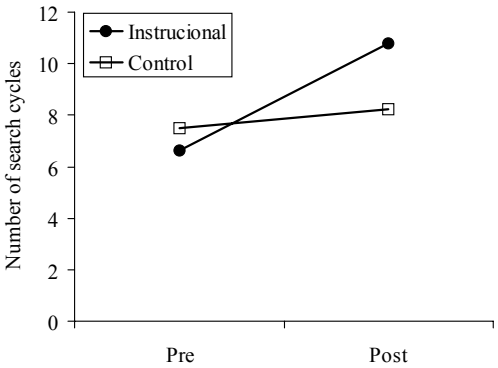


Figure 5: Average number of search cycles in the pre- and post-tests.

Search processes

Finally, we discuss whether the participants' processes improved with our instructional design by comparing the pre- and post-tests in each group. In this study, we consider learners' information seeking processes as a cycle of search in the Keyword space and the Web space. This approach, where problem solving is considered to be a search for multiple spaces, has been widely approved in the studies on scientific discovery and creative processes. These studies have suggested that target activities are developed while repeating the cycle of searching multiple spaces. Therefore, we focused on the cycle of searching multiple spaces. We defined one search cycle as "a set of transitions from the Keyword space to the Web-Page space." We counted the number of search cycles in each task, and Figure 5 shows the average number of search cycles in each group.

The number of search cycles was analyzed in a two-way mixed ANOVA with the group (the instructional/control) as a between-subjects factor and the test (pre-test/post-tests) as a within-subjects variable. There was a significant main effect of the test ($F(1, 33) = 6.37, p < .01$), indicating that the number of cycles increased from the pre-test to the post-test. The Group $\times$ Test interaction was also found to show a trend toward significance ($F(1, 33) = 3.07, p < .10$), which indicates that the participants in the instructional group more effectively increased the number of cycles than did in the control group. These results prove that the participants in the instructional group searched two spaces more actively in the post-test than in the pre-test.

Discussions and Conclusions

In this study, we proposed an instructional design that supports reflective activities by presenting learners' problem solving processes in information seeking on the Web and evaluated its educational effects. We conducted an experiment to evaluate the effects of our design. Experimental results revealed that the participants' search performance in the instructional group improved more effectively than in the control group. Additionally, their ideas about important activities in information-seeking

Table 3: Multiple scaffolds in our instructional design.

	Reflection in Action	Reflection on Action
Process Display	◯	◯
Process Prompt	◯	×
Process Models	×	◯
Reflective social discourse	×	×

on the Web and that their search processes also changed from the pre-test to the post-test in comparison with the control group. These results indicate that our design helps learners improve their search performances and acquire search skills.

Finally, we discuss scaffolds in our instructional design. In this study, we focused on the process display, pointed out by Lin et al. (1999) to support learners' reflection on their problem-solving processes. Furthermore, they also proposed the following three scaffolds for reflective thinking:

Process prompts: prompting students' attention to specific aspects of processes while learning is in action

Process models: modeling of experts' thinking processes that are usually tacit so that students can compare and contrast with their own process in action

Reflective social discourse: creating community-based discourse to provide multiple perspectives and feedback that can be used for reflection

Lin et al. (1999) suggested that it is important to incorporate all four scaffolds when developing designs because each method supports a different aspect of reflective thinking. We designed a learning environment in which learners could experience two types of reflection, such as "reflection-in-action" and "reflection-on-action", providing multiple methods for scaffolds referred by Lin et al. (1999) to support learners' reflective activities. Table 3 summarized types and methods of scaffolds in our design. In this paper, we empirically verified the effectiveness of combining these multiple methods for supporting reflective thinking.

Additionally, experimental results also imply that a cognitive schema is useful for not only analyzing human cognitive processes, but also supporting learning activities. However, we need to conduct further investigations on how each component in our educational design, such as a cognitive schema, "reflection-in-action," and "reflection-on-action," and above scaffolds, affects the learners' improvements.

References

Aleven, V. & Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2):147–179.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, 228:456–462.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In Flavell, J. & Markman, E., editors, *Cognitive Development*. John Wiley and Sons, New York. Handbook of child psychology: Vol. III.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: how students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.

Ertmer, P. A., Newby, T. J., & MacDougal, M. (1996). Students' responses and approaches to case-based instruction: the role of reflective self-regulation. *American Educational Research Journal*, 33(3):719–752.

Hershkowitz, R. & Schwarz, B. (1999). Reflective processes in a mathematics classroom with a rich learning environment. *Cognition and Instruction*, 17(1):65–91.

Leinhardt, G. & Young, K. M. (1996). Two texts, three readers: distance and expertise in reading history. *Cognition and Instruction*, 14(4):441–486.

Lin, X., Hmelo, C., Kinzer, C. K., & Secules, T. J. (1999). Designing technology to support reflection. *Educational Technology Research and Development*, 47(3):43–62.

Lin, X. D. & Lehman, J. (2001). Designing metacognitive activities. *Educational Technology Research and Development*, 49(2):23–40.

Newell, A. & Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, N.J.

Saito, H. & Miwa, K. (2002). Discovery process on the www: Analysis based on a theory of scientific discovery. In *Proceedings of the 5th International Conference on Discovery Science (DS 2002)*, LNCS 2534, (pp. 449–456).

Schauble, L., Raghavan, K., & Glaser, R. (1993). The discovery and reflection notation: A graphical trace for supporting selfregulation in computer-based laboratories. In Lajoie, S. & Derry, S., editors, *Computers as Cognitive Tools*. Lawrence Erlbaum Associates, Hillsdale, N.J.

Schön, D. A. (1987). *Educating the Reflective Practitioner*. Heath, Boston.

Modeling Effects of Age in Complex Tasks: A Case Study in Driving

Dario D. Salvucci **Alex K. Chavez** **Frank J. Lee**
(salvucci@cs.drexel.edu) (achavez@drexel.edu) (fjl@cs.drexel.edu)
Department of Computer Science, Drexel University
3141 Chestnut St., Philadelphia, PA 19104

Abstract

While computational cognitive modeling has made great strides in addressing complex dynamic tasks, the modeling of individual differences in complex tasks remains a largely unexplored area of research. In this paper we present a straightforward approach to modeling individual differences, specifically age-related cognitive differences, in complex tasks, and illustrate the application of this approach in the domain of driving. We borrow ideas from rigorous work in the EPIC cognitive architecture (Meyer et al., 2001) and extend them to the ACT-R architecture (Anderson et al., in press) and a recently-developed ACT-R driver model (Salvucci, Boer, & Liu, 2001) to model the effects of age on driver behavior. We describe two validation studies that demonstrate how this approach accounts for two important age-related effects on driver performance, namely effects on lateral stability and brake response during both normal driving and driving while performing a secondary task.

Introduction

Computational architectures and cognitive modeling have in recent years begun to account for increasingly complex and dynamic tasks, in domains such as piloting combat aircrafts (Jones et al., 1999) and controlling air traffic (Lee & Anderson, 2001). While such models have captured many aspects of human cognition and performance in these tasks, one aspect of complex tasks, namely individual differences, remains a largely unexplored area of research. The modeling community has seen several rigorous studies of individual differences in the context of cognitive architectures, perhaps most notably the work of Meyer et al. (2001) in the EPIC cognitive architecture (Meyer & Kieras, 1997) and that of Lovett, Daily, and Reder (2000) in the ACT-R architecture (Anderson et al., in press). However, due to their emphasis on specific sources of individual differences, these studies focused on relatively short laboratory tasks in controlled environments rather than more complex continuous tasks in dynamic environments.

Our goal in this paper is to generalize ideas from existing work on individual differences in simpler tasks to account for individual differences in complex dynamic tasks. We illustrate our approach in the domain of driving, a complex task that people perform on daily basis. There now exist several so-called "integrated driver models" (e.g., Aasman, 1995; Levison & Cramer, 1995) that attempt to combine the lower-level aspects of driving (e.g., steering control) with the higher-level aspects of the task (e.g., decision making, navigational planning). In particular, Salvucci, Boer, and Liu (2001) have developed and refined an ACT-R driver model that predicts many aspects of driver control, situational awareness, and decision making during common highway driving. However, to date, no integrated models of driving, including the ACT-R driver model, have accounted rigorously for any individual differences in driver behavior and performance.

This paper builds on previous work by presenting an account of individual differences, specifically age-related differences, in the complex task of driving. Not surprisingly, age plays a significant role in driver differences, often couched in broad terms as differences between younger drivers (roughly 20-30 years of age) and older drivers (roughly 60-70 years of age). Our approach borrows recent results of Meyer et al. (2001), who explored models of age-related individual differences in the context of the EPIC cognitive architecture. Age effects on driving offer a particularly interesting challenge to computational cognitive modeling: on the one hand, some studies have found that older drivers exhibit performance equal to that of younger drivers for certain combinations of driving and/or secondary tasks; on the other hand, other studies have found that older drivers sometimes experience extremely reduced performance, particularly in the presence of secondary tasks (e.g., using a cell phone). Thus, the effects of age are far from trivial and must be taken in the fuller context of both the complex behavior necessary for driving and also the complex interaction between the driver and the "artifact" (i.e., vehicle, road, etc.) through which the driver's behavior is externalized.

In the next section of the paper, we describe our basic approach and its instantiation in the ACT-R cognitive architecture. We then present two modeling studies that validate our approach for complementary tasks and aspects of behavior, namely drivers' ability to maintain lateral stability on the road and drivers' ability to respond (i.e., brake) to sudden external stimuli. While our work in this paper emphasizes driver behavior and the ACT-R cognitive architecture, the fundamental ideas generalize well to other complex task domains and other modeling frameworks. Thus, our ultimate goal is to explore the interaction between basic individual differences and their downstream effects on performance in complex dynamic task environments.

Modeling Age Effects in Driving

The various types of age-related differences that might arise in driving can be categorized broadly in terms of "hardware" and "software" differences (Meyer et al., 2001). Hardware differences arise from fundamental changes to the human system — for instance, a slowdown in cognitive processing, visual processing, or motor movement. Software differences arise in modifications or differences in the strategies used to accomplish tasks — for instance, intentionally slowing down and backing away from a lead vehicle when talking on the phone. In this paper, we focus on hardware differences, specifically differences in cognitive processing. While there is no doubt that both hardware and software differences play a role in effects on driver performance, we wish to explore to what extent modeling of basic hardware differences can account for critical effects on performance found in recent driver studies.

The ACT-R Cognitive Architecture

The ACT-R cognitive architecture (specifically version 5.0: Anderson et al., in press; see also Anderson & Lebiere, 1998) is a production-system cognitive architecture based on two types of knowledge stores, declarative and procedural. Declarative knowledge embodies "chunks" of symbolic information including factual ('3+4=7'), perceptual ('car 10 m in front'), and goal-related ('driving to the grocery store') information. Procedural knowledge operates through condition-action "production rules" that evaluate the current state of declarative knowledge (e.g., 'if my goal is to pass the lead car') and enact changes on memory and/or the environment accordingly (e.g., 'check that there is sufficient room in the left lane'). Each production rule firing (instantiation and execution) requires 50 ms of cognitive "effort" time, in addition to any time needed to wait for conditions to be met, such as the completion of a memory retrieval. Overall, the ACT-R architecture has a number of built-in functions that enable human-like behavior (e.g., interaction of memory and perceptual-motor processes) as well as built-in limitations on behavior (e.g., forgetting declarative chunks after a period of inactivity). An in-depth discussion of the architecture is beyond the scope of this paper; interested readers may wish to consult Anderson et al. (in press) for more information.

ACT-R and Age Effects

To model age effects, specifically hardware-related effects on cognitive processing, we base our approach on recent work by Meyer et al. (2001). Meyer et al. found that one of the most robust differences between younger and older people arose in the speed of cognitive processing. In particular, they found that, in the context of their EPIC architecture (Meyer & Kieras, 1997), the time for a production-rule firing increases from 50 ms for a younger person to 56.5 ms for an older person — a 13% increase. They offer several pieces of evidence to back their claim. First, for the initial claim of a 50 ms firing time, they argue that this value has a neurological correlate in the average period between zero crossings in the brain's alpha rhythm for younger adults, which has a positive relationship with mean simple response time (see Callaway & Yeager, 1960; Surwillo, 1963; and Woodruff, 1975, as cited by Meyer et al.). For older adults, they argue that mean zero-crossing periods for alpha rhythms is about 10-15% higher for subjects with an age close to 70 when compared to young adults; older subjects' mean simple response times also show a 10-15% increase (see Cerella, 1985; Somberg & Salthouse, 1982, as cited by Meyer et al.). These data lead the authors to conclude that the mean cognitive processor time increases by 13% for older adults and that this is a robust finding independent of task.

The work of Meyer et al. has a straightforward interpretation in the ACT-R architecture. ACT-R, like EPIC, uses a 50 ms cycle time for production rules. To model an older person, we simply incorporate the same cycle-time increase as Meyer et al. — namely, we increase the cycle time for production rules (called "effort" times in ACT-R) by 13%. As we will show, this change impacts performance in non-trivial ways: instead of a 13% impact on performance across measures, the change produces no effects for some measures and large effects for others depending on the emergent interactions between model and task.

In this paper, we focus in particular on the effects of cognitive cycle time and ignore potential changes in the timing of perceptual and motor processes. Meyer et al. also explored how perceptual and motor processes are affected by age; however, the mapping of their results to the ACT-R architecture is not as straightforward as the mapping for cognitive cycle time, and thus we leave this for future work. Nevertheless, we demonstrate in this paper that at least some significant aspects of age-related individual differences in driving can be successfully accounted for simply by incorporating basic differences in cognitive processing.

Driving and Age Effects

Modeling the effects of age on driver performances centers on our use of the ACT-R integrated driver model (Salvucci, Boer, & Liu, 2001). The driver model, as mentioned, incorporates both the lower-level aspects of vehicle control with the higher-level aspects of driver situational awareness and decision making. The model can navigate a variety of highway environments, the most common being a multi-lane highway with automated traffic and realistic vehicle dynamics, as pictured in Figure 1. While driving, the model interacts with the simulated environment through a virtual steering wheel and pedals, producing behavioral protocols completely analogous to those of human drivers in the simulator — recording, for example, steering and pedal depression over time along with eye movements to visual regions. The model has been validated with respect to various aspects of basic driver behavior, such as curve negotiation and lane changing (e.g., Salvucci, Boer, & Liu, 2001), and also with respect to effects of secondary tasks on performance (e.g., Salvucci, 2001).

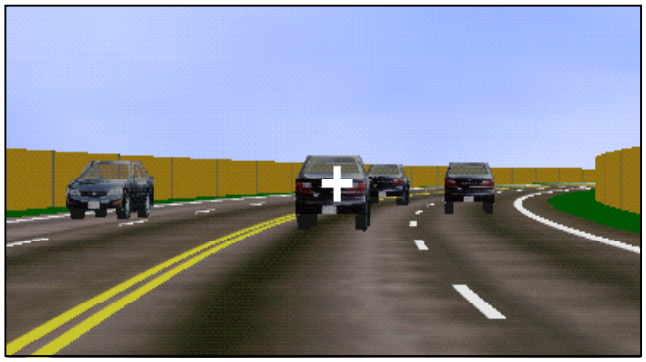


Figure 1: Sample driving simulation environment.

To model the effects of age on driver behavior, we incorporate the 13% cycle-time increase into the driver model. The increase affects all production rules across the model — most importantly, slowing down the iterating control cycle that handles the updates for steering and speed control. As mentioned, this has non-trivial downstream effects on performance rather than a simple 13% effect across measures of performance. The next two studies demonstrate how such a small change at the "hardware" level can result in very interesting emergent behavioral predictions.

Study 1: Age Effects on Lateral Stability

Our first validation study addresses effects of age on drivers' ability to maintain lateral stability — that is, side-to-side stability as measured by lateral velocity. Reed & Green (1999) compared the performance of younger and older human drivers in a simulator and on the road while executing a secondary task (dialing phone numbers). We focus our analysis on their simulator data, comparing the performance of their drivers to the predictions of the driver model in a simulator with the same task.

Human Data

In Reed and Green's (1999) study, drivers navigated a simulated straight road at a constant speed of roughly 60 mph and were occasionally cued verbally to perform a secondary task, namely dialing an 11-digit phone number (including '1' and an area code: e.g., '1-215-555-1212'). On cue, drivers picked up the phone, dialed the 11-digit number presented on a card located at the center console, and pressed a "Call" button to initiate the call. The driver then received a voice confirmation that the number was dialed correctly, and finally the driver pressed "End" to end the call. Reed and Green collected data from a total of twelve drivers, six of whom were older than 60 years of age, six of whom were between the ages of 20 and 30. They measured lateral stability as the mean lateral (side-to-side) velocity of the vehicle both during the secondary task and during normal driving.

Model Simulations

The model for the Reed and Green task was derived in a straightforward manner. The ACT-R driver model was integrated with a task environment analogous to that in the Reed and Green study — that is, a simulated one-lane straight road. The one difference in the model's task environment from that of Reed and Green arose in speed control: because the model has no speedometer with which to monitor speed, the model was given a lead vehicle driving at a constant speed, which it used to monitor its own speed. The model was then extended to include a model for the secondary task of phone dialing. This secondary-task model derived directly from a similar previous model of dialing (Salvucci et al., 2004) specified in the ACT-Simple framework (Salvucci & Lee, 2003), which is essentially a shorthand notation for standard ACT-R production rules. The model, shown in Table 1, differed from the previous model only in that it dialed the prefix "1" before the area code and phone number and that it looked at the cue card for the number rather than recalling it from memory. The "pop" marking in the table denotes commands after which the dialing model passed control to the driving task. Because the control characteristics of the Reed and Green simulator (e.g., steering force feedback) differed from those of the simulator used to validate the original driver model, three parameters[1] of the model that control overall steering were adjusted to produce the best fit in the results below. However, it should be noted that the model immediately produced the desired qualitative fit — this estimation only improved the quantitative fit. The younger and older driver models differed only in the 13% increase in cognitive cycle time for the older driver model. The model data reported below represents roughly 4-5 minutes of driving in which the model performed eight secondary-task trials with a 20 s delay between task trials.

Table 1: Secondary-task model for Study 1.

```
(move-hand device pop)
(think pop)
(press-button key1 pop)
(look-at device pop)
(think pop)
(press-button key2)
(press-button key1)
(press-button key5 pop)
(look-at device pop)
(think pop)
(press-button key8)
(press-button key6)
(press-button key7 pop)
(look-at device pop)
(think pop)
(press-button key5)
(press-button key3)
(press-button key0)
(press-button key9 pop)
(think pop)
(press-button send pop)
```

[1] $k_{far} = 13,\ k_{near} = 5.6,\ k_I = 1$

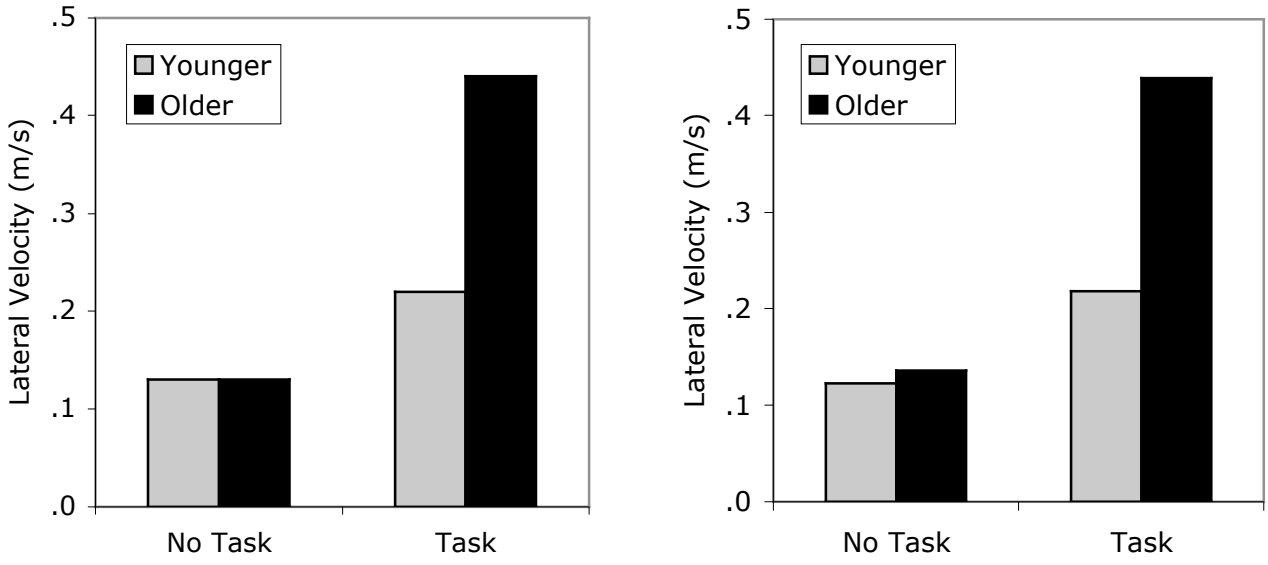


Figure 2: Lateral velocity, (a) human drivers (Reed & Green, 1999) and (b) model predictions.

Results

Figure 2(a) shows the lateral-velocity results taken from the human drivers in the Reed and Green study. In the No-Task condition, the younger and older drivers performed equally and there were no effects of age. In the Task condition, while the performance of both younger and older drivers degraded significantly, the performance of the older drivers was affected far more dramatically, with older drivers exhibiting a mean lateral velocity of .44 m/s and younger drivers a mean lateral velocity of .22 m/s in this condition.

Figure 2(b) shows the models' predictions for the same conditions, R=.99. The models, like human drivers, exhibit no age effect in the No-Task condition. Here the 13% cycle-time increase is not large enough to affect the downstream performance with respect to lateral velocity, effectively adding only tens of milliseconds to the overall control-cycle time: while the younger model updates control every 200 ms, the older model updates every 226 ms, and the extra 26 ms does not affect overall steering performance. However, also like human drivers, the model exhibits differential effects of age and task on performance. The younger model exhibits reduced performance because of less time devoted to control, as we have observed in previous studies (e.g., Salvucci, 2001). The older model exhibits an even greater degradation because of occasional, somewhat severe steering corrections: in situations where the younger model may not update control for, say, 1 s, the 13% increase for the older model would exceed 100 ms — enough time for the vehicle to travel roughly 2.7 m at the given speed and move significantly off-center in the lane. The model, seeing the large offset from lane center, performs a hard steering correction and generates a large lateral velocity. In fact, the younger model also experiences such corrections; however, the corrections are both more frequent and more severe for the older model.

Study 2: Age Effects on Brake Response

Our second validation study addresses effects of age on drivers' ability to respond to sudden external events via braking. Hancock *et al.* (2003) ran an empirical study on younger and older drivers to investigate differential effects of cell-phone distraction on braking performance. We now examine their task and results and show how the same driver model can account for this very different aspect of driver behavior and performance.

Human Data

In Hancock *et al.*'s (2003) study, drivers drove down a test track at approximately 25 mph toward an intersection with a stoplight. During some trials, the driver was cued by tone to perform a secondary task: they looked at a digit on mounted screen and pressed a key to indicate whether or not this digit corresponded to the first digit of a previously-memorized number. Also during these trials, the stoplight turned red 0.5-1 s after the onset of the secondary task, causing the driver to brake in response. During other trials, the driver only responded to the red stoplight without a secondary task. Hancock et al. collected data from 36 drivers — 19 between the ages of 25 and 36, and 16 between the ages of 55 and 65. Overall, they measured brake response time with and without the task as the time delay between the onset of the red stoplight and the initial depression of the brake.

Model Simulations

To model the Hancock et al. task, we took the model from the Reed and Green task and modified only the task components of the model. The driver model does not currently have the ability to encode and monitor stoplights, and thus we modified the environment such that the lead

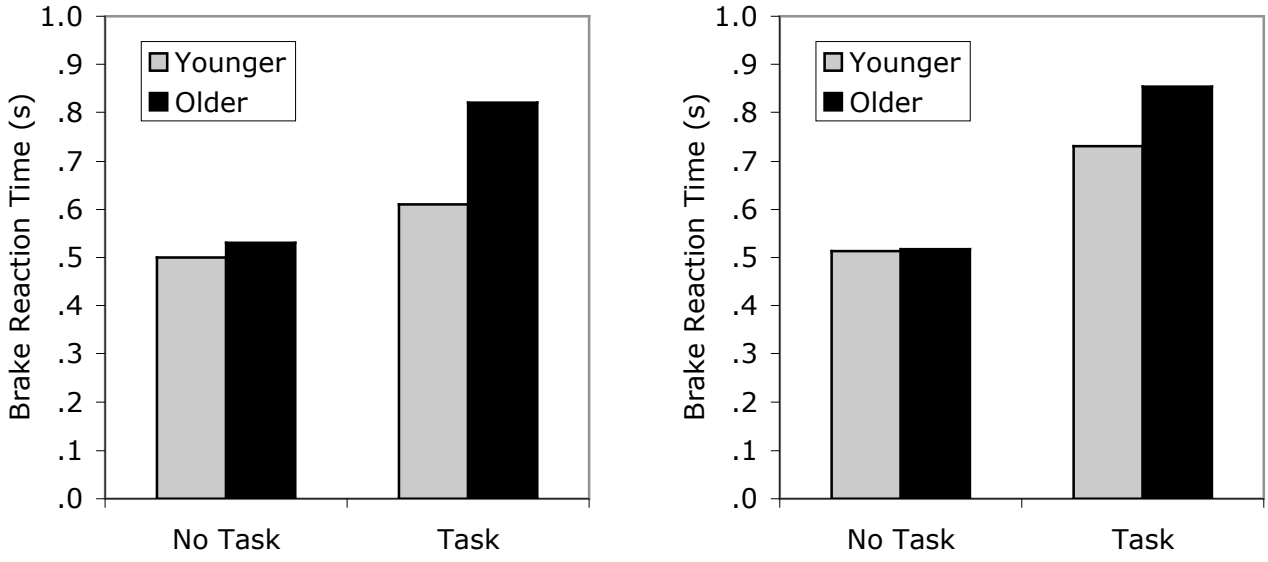


Figure 3: Brake reaction time, (a) human drivers (Hancock et al., 2003), and (b) model predictions.

vehicle's brake lights would turn red 0.5-1 s after the onset of the secondary task, keeping the basic temporal structure of the Hancock et al. task. The secondary-task model was derived by modifying the Reed and Green task model to type only 1 keypress as opposed to 11; the task model is shown in Table 2. All parameter values were taken directly from the model in Study 1. However, we had to modify one braking-related parameter because of the nature of this task: the standard driver model requires 500 ms to move its foot from accelerator to brake, but because of the emergency nature of this task, we instead used a time of 310 ms — recently reported by Lee et al. (2002) as the minimum time for this movement — and eliminated motor preparation time due to the drivers' pre-preparation of the movement as they approached the intersection. Again, the younger and older driver models differed only in the 13% increase in cognitive cycle time for the older driver model. The model data below includes roughly 5 minutes of driving in which the model performed 16 secondary-task trials with a 10 s delay between task trials.

Table 2: Secondary-task model for Study 2.

```
(move-hand device pop)
(look-at device pop)
(think pop)
(press-button key5 pop)
```

Results

Figure 3(a) shows the results for the human drivers. We see a similar pattern emerge in this study as we saw in Study 1. First, younger and older drivers showed no significant difference in the No-Task condition. Second, the secondary task significantly degraded the performance of both groups in the Task condition. Third, the task has a greater effect on the older drivers than the younger drivers. We should note that, although the graphs for Studies 1 and 2 are visually similar, they show very different aspects of behavior; the similarity is rather surprising given that one study examines lateral stability while the other emphasizes response time for longitudinal (braking) behavior.

Figure 3(b) shows the results for the model simulations, R=.94. As in Study 1, the models nicely account for the human drivers' behavior. The younger and older models show equivalent braking response times in the No-Task condition. Again, the slightly longer control update cycle for the older model is not enough to produce a significant effect. At the same time, the models show large effects of task in the Task condition, and the older model shows a significantly larger effect than the younger model. As in Study 1, the 13% cycle-time increase for the secondary task model increases pauses in control by 100 ms or more, thus producing an effect of roughly this size between the younger and older models. One might expect that the age effect for brake response might be heavily tied to differences in motor-movement speeds (from accelerator to brake), and given that the interaction effect in the human data is slightly larger than that for the model, this could indeed be one factor. Nevertheless, these results show that differences in cognitive processing are also a major component of this interaction and accounts for critical aspects of the human data.

General Discussion

In this paper we present a straightforward method of accounting for age-related differences in driver performance, focusing on "hardware" differences in cognitive processing time. While the idea of slowing processing time by 13%

for older people seems simple enough, it should be noted that the resulting predictions are far from trivial. Indeed, one might at first expect the slowdown to result in analogous performance decrements — for instance, a 13% degradation in lateral velocity and braking response. However, for complex dynamic tasks, this situation is much more complicated: the model's behavior is filtered through both the perceptual-motor processes and the vehicle dynamics, resulting in predictions that can only be generated and tested through "embodied" cognitive models that interact directly with realistic task environments. The two validation studies show that the ACT-R driver model, in the context of such a realistic environment, successfully accounts for these complex interactions in the driving domain, namely for both the lack of effects (in the No-Task condition) and larger-than-expected effects (in the Task condition) for lateral and longitudinal measures.

This work also illustrates one of the important advantages to working in the context of a cognitive architecture — namely, the sharing and re-use of ideas and model implementations within an architecture and even across different architectures. Not only does the work of Meyer et al. (2001) have large implications for their own EPIC architecture (Meyer & Kieras, 1997), the work translates well to other architectures such as ACT-R. In addition, this type of foundational work has immediate implications for all models developed in the architectures; for instance, other ACT-R models of complex dynamic tasks (or any tasks general) could incorporate the 13% cycle-time increase to immediately derive age-related predictions, enabling comparison to human data for a host of new measures. Such work would nicely complement recent work on other aspects of individual differences, such as differences in working memory (Lovett, Daily, & Reder, 2000). These studies of individual differences bring to light the predictive power inherent in cognitive architectures and help to make further strides toward Newell's (1990) vision of more "unified theories of cognition."

Acknowledgments

This work was supported in part by Office of Naval Research grant #N00014-03-1-0036 and National Science Foundation Grant #IIS-0133083 to the first author.

References

Aasman, J. (1995). *Modelling driver behaviour in Soar*. Leidschendam, The Netherlands: KPN Research.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (in press). An integrated theory of the mind. *Psychological Review*.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Erlbaum.

Callaway, E. & Yeager, C. L. (1960). Relationship between reaction time and electroencephalographic alpha base. *Science, 132,* 1765-1766.

Cerella J. (1985). Information processing rates in the elderly. *Psychological Bulletin, 98,*67-83.

Hancock, P.A., Lesch, M., Simmons, L. (2003). The distraction effects of phone use during a crucial driving maneuver. *Accident Analysis and Prevention, 35,* 501-514.

Jones, R. M., Laird, J. E., Nielsen P. E., Coulter, K., Kenny, P., & Koss, F. (1999). Automated intelligent pilots for combat flight simulation. *AI Magazine*, *20*, 27-42.

Lee, F. J., & Anderson, J. R. (2001). Does learning of a complex task have to be complex? A study in learning decomposition. *Cognitive Psychology*, *42*, 267-316.

Lee, J. D., McGehee D. V., Brown T. L., Reyes, M. L. (2002). Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high fidelity driving simulator. *Human Factors*, *44*, 314-334.

Levison, W. H., & Cramer, N. L. (1995). Description of the integrated driver model (Tech. Rep. No. FHWA-RD-94-092). McLean, VA: Federal Highway Administration.

Lovett., M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research*, 1, 99–118.

Meyer, D. E., Glass, J. M., Mueller, S. T., Seymour, T. L., & Kieras, D. E. (2001). Executive-process interactive control: A unified computational theory for answering twenty questions (and more) about cognitive ageing. *European Journal of Cognitive Psychology*, *13*, 123-164.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review*, *104*, 3-65.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Reed, M. P & Green, P. A. (1999). Comparison of driving performance on-road and in a low-cost driving simulator using a concurrent telephone dialing task. *Ergonomics, 42*, 1015-1037.

Salvucci, D. D. (2001). Predicting the effects of in-car interface use on driver performance: An integrated model approach. *International Journal of Human-Computer Studies*, *55*, 85-107.

Salvucci, D. D., Boer, E. R., & Liu, A. (2001). Toward an integrated model of driver behavior in a cognitive architecture. *Transportation Research Record*, 1779.

Salvucci, D.D., John, B.E., Prevas, K., & Centgraf, P. (2004). Interfaces on the road: Rapid evaluation of in-vehicle devices. To appear in HCIC 2004.

Salvucci, D. D., & Lee, F. J. (2003). Simple cognitive modeling in a complex cognitive architecture. In *Human Factors in Computing Systems: CHI 2003 Conference Proceedings* (pp. 265-272). New York: ACM Press.

Somberg, B. L. & Salthouse, T. A. (1982). Divided attention abilities in young and old adults. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 651-663.

Surwillo, W. W. (1963). The relation of simple response times to brain wave frequencies and the effects of age. *Electroencephalography and Clinical Neurophysiology, 15*, 105-114.

Woodruff, D. S. (1975). Relationships among EEG alpha frequency, reaction time, and age: A biofeedback study. *Psychophysiology, 12*, 673-681.

An Artificial Life Approach to the Study of Basic Emotions

Matthias Scheutz (mscheutz@cse.nd.edu)
Department of Computer Science and Engineering
Notre Dame, ND 46556 USA

Abstract

We propose a methodological framework for the study of emotional control based on extensive computer simulations with arti cial agents implementing emotional control mechanisms and demonstrate the methodology with simulations experiments in an arti cial environment. Speci cally, a biologically plausible schema-based model of basic forms of fear and anger is proposed and tested with respect to a variety of parameter ranges.

Introduction

Emotions are an integrative part of our mentality. At the level of the functional architecture they serve several crucial roles, from fast perceptions of threats, to focusing and redirecting attention, to in uencing memory storage and retrieval, to social regulation through expression and perception of emotions, and many more (Derryberry & Tucker, 1994; Fredrickson, 1998; Bless, Schwarz, & Wieland, 1996; Schwarz, 1990; Blaney, 1986; Kahneman, Wakker, & Sarin, 1997; Clore, Gasper, & Conway, 2001; Frijda, 2000; Cosmides & Tooby, 2000). Several circuits have been hypothesized to be involved in emotional processing in mammalian brains, yet only a few computational models (mostly of fear mechanisms) have been proposed and implemented in an effort to test theoretical predictions about emotion processes and mechanisms. Moreover, these models are limited to very specific processes (e.g., Pavlovian fear conditioning) and do not specify other parts of an architecture that are required for a complete, functional control system (e.g., homeostatic control mechanisms, various forms of perceptual processing, action selection mechanisms, etc.). Hence, they leave out and cannot address many other emotional states that essentially depend on additional processing components (e.g., such as social emotional states that depend on the expression and perception of emotions).

One way to study the effects of emotional control circuits for individual agents as well as groups of agents is to conduct simulations with artificial agents that are controlled by architectures that define emotion models. Such simulation studies have the advantage that the role of emotions and the consequences of emotional disturbances can be analyzed at several different levels at the same time: the *mechanistic level* of the implementation of the model (e.g., a neuronal level), the *individualistic level* (e.g., the control loops between emotion circuits and the agent body), and the *social level* (e.g., the effects of emotional signaling for the well-being or functioning of a group).

In this paper we will (1) propose a methodological framework for the study of emotional control based on extensive computer simulations with artificial agents implementing emotional control mechanisms and (2) demonstrate the methodology with simulations experiments in an artificial environment. Specifically, a biologically plausible schema-based model of basic forms of fear and anger is proposed and tested with respect to a variety of parameter ranges. The results show where emotional control is successful and better than non-emotional strategies, but also where it fails.

Background on Computational Models of Emotions

While several suggestions about the neural and functional organization of emotional circuits exist in the literature, there are currently only a few proposals for computational models that implement and test them. The existing computational models can be categorized into two main classes, based on whether they are aimed at explaining low-level neurological structures and mechanisms, or whether they are intended to model higher-level emotional processes. The low-level models can further be divided into general processing models of brain mechanisms and specific emotion models of particular brain structures.

The most extensively developed low-level models among the first kind are Grossberg's *CogEM* models (e.g., (Grossberg & Schmajuk, 1987)), which are models of learning cognitive, emotional, and motor properties. *CogEM* models can account for several effects in Pavlovian fear conditioning (e.g., secondary conditioning or attentional blocking), but have not been directly applied to empirical data (e.g., data from fear conditioning studies with rats).

Another class of low-level neural models is targeted specifically at modeling the amygdala, which performs several functions in emotion processing (LeDoux, 1996; Rolls, 1995). The lateral amygdala, for example, has been shown to exhibit associative plasticity during fear learning (Blair, Tinkelman, Moita, & LeDoux, 2003) and a preliminary computational model of associative learning in the amygdala has been developed and tested in three associative learning tasks (Balkenius, 2000).

Moreover, recent evidence from studies with rats suggests that the amygdala, in particular, the frontotemporal amygdala, which is taken to integrate sensory information, encodes hedonic values of an unconditioned stimulus as part of the fear memory (Fanselow & Gale, 2003). LeDoux and colleagues have hypothesized a dual pathway model of emotional processing in the amygdala, which they tested in auditory fear conditioning studies (Armony, Servan-Schreiber, Cohen, & LeDoux, 1995). These models have been also used in simulated lesion studies and successfully compared to data from actual lesion studies with rats.

While most research on emotional modeling in low-level models is focused on Pavlovian conditioning and targeted at neural structures and processing mechanisms, higher-level models of emotions are intended to capture the processing sequence involved in emotion processes and are typically concerned with a wider range of emotions. While all low-level models are neural network models, higher-level models comprise both connectionist and symbolic approaches.

An example of a high-level connectionist approach is the ITERA model (Nerb & Sperba, 2001), which is intended to study how media information about environmental problems in uences cognition, emotion, and behavior. Facts, input types, emotions, and behavioral intentions are all represented in terms of individual neural units that are connected via excitatory and inhibitory links and compete for activation.

Most attempts to model emotions at higher levels, especially in artificial intelligence research, are however based on symbolic architectures (e.g., Soar (Newell, 1990) or ACT (Anderson, 1993)). They typically focus on the OCC model (Ortony, Clore, & Collins, 1988), which hypothesizes prototypical "update rules" for changes in emotional state that can be directly implemented in rule-based systems (e.g., (Marsella & Gratch, 2002).

What is common to all the above emotion models is that they have been implemented and tested in isolation from any *body model*. Consequently, it is difficult if not impossible to investigate crucial aspects of emotion processing that need a body for control and thus go beyond functional properties (like the effects of Pavlovian conditioning), which can be tested in stand-alone models (e.g., by applying a stimulus and measuring the output).

While some attempts have been made to implement connectionist emotion models on robots, where different emotions types are represented as connectionist units that compete for activation, which in turn cause the robot to exhibit a particular behavior (e.g., (Michaud & Audet, 2001; Breazeal, 2002; Arkin, Fujita, Takagi, & Hasegawa, 2003)), these architectures do not attempt to model any specific psychological or neurobiological theory of emotions (e.g., in an effort to verify or falsify its predictions). Rather, they are mainly concerned with the applicability of a particular control mechanism from an engineering perspective. Moreover, these models typically lack a systematic evaluation of their performance (an exception is (Breazeal, 2002)). Finally, no experiments have been performed with these robotic architectures to investigate the effects of "emotional malfunctioning".

Probably the most significant restriction of current efforts to model emotions is that they have not been extended to multi-agent environments. Yet, social aspects of emotions (such as signaling emotional states through facial expressions, prosody, gestures, etc.) and the resultant effects at the group level cannot be studied in a single, isolated agent. Rather, multiple interacting agents with emotional control systems are required, especially for arguments about the adaptive role of emotions (e.g., (Cosmides & Tooby, 2000)). To our knowledge only one project (Dulk, Heerebout, & Phaf, 2003) uses an artificial life simulation to study some evolutionary aspects related to emotional processing, specifically, the evolutionary justification for LeDoux's dual-route fear processing proposal (LeDoux, 1996). However, the employed neural network does not and is not intended to implement emotions or model emotional circuits. And while the employed neural network suggests some interesting conclusions about the circumstances under which dual processing routes might be beneficial, it does not capture emotional circuits, and is, therefore, silent about emotional phenomena.

Simulations of Emotional Agents

Over the last few years we have developed an agent-based simulation environment SWAGES to investigate different agent architectures and architectural mechanisms. In particular, two main roles of emotions in agent control systems have been studied in extensive simulations in an effort to evaluate the utility of emotional control (compared to other non-emotional control strategies): *the role of emotions for individual agents* (e.g., the selection of actions) and *the role of emotion for social groups* (e.g., in con icts with conspecifics and individuals from other species).

Results from simulation experiments with agents performing foraging tasks, for example, show that *action selection* based on emotional states can be very effective in the competition for resources in hostile multi-agent environments (e.g., (Scheutz, 2001) and that motivational "hunger" and "thirst" states as well as emotional "fear" and "anger" states are likely to evolve in a variety of competitive multi-agent environments (Scheutz & Sloman, 2001)).

In general, we found that agents with emotional control mechanisms performed much better in a variety of foraging and survival tasks in environments with little to no structure than agents with much more sophisticated cognitive control systems if the "cost of deliberation" is taken into account (e.g., (Scheutz & Schermerhorn, 2002)).

On the social side, we found that expressing emotions and being able to react to emotional expressions of others can have a beneficial regulatory effect in social groups and lead to superior con ict resolution strategies (e.g., (Scheutz & Schermerhorn, 2004)).

In all these studies, we construed emotions as con-

trol processes that initiate, interrupt, suppress, reprioritize, or in general modify behavior or behavioral dispositions. Emotions are implemented in terms of control components (typically, in neural networks) that are connected in appropriate ways to sensors and effectors of agent body models. The underlying assumption is that the level of control components is appropriate for analyzing and understanding the functional organization of emotion mechanisms. In the following we brie y outline our architectural approach to the study of emotions and present some experimental results.

Basic Motivations and Emotions as Control Processes

Motivations may be considered *desire-like states* in that they in uence and bias an agent's behavioral dispositions in such a way as to contribute to the realization of a desired change in the environment and/or agent. We use the term "basic motivations" to refer to motivations that have little to no cognitive involvement and are primarily linked to "basic needs" of an agent (e.g., to maintain a certain energy level). For some of these, the familiar term "drive" is appropriate, namely if the agent is driven in a mostly reactive way to act so as to eliminate the disparity between a desired and an actual state that was the cause for the motivation. For example, a state of an agent's control system qualifies as a "hunger" state if it is caused by lack of energy and results in food-seeking behavior (McFarland, 1981).

It is possible to use control components, whose outputs control gain values of motor controllers, to implement the kind of control system that will be able to instantiate basic motivations. For example, "hunger" could be instantiated by a proportional controller P (Özbay, 2000) such that input to P comes from an internal sensor S that measures the current energy level. P compares a desired equilibrium energy level (i.e. set point), e_{des}, to the actual energy level e_{act} and scales the difference by a gain factor g_e: $P = g_e \cdot (e_{des} \quad e_{act})$. The output then is a measure of the urgency with which the system requires energy. Hence, the intensity of basic motivations is modeled by the magnitude of the control circuits' outputs that can in turn modulate behavior.

Emotions may also be considered to be desire-like in that they in uence and bias an agent's behavior. Again, we use "basic emotion" to refer to states with little or no cognitive involvement. For our purposes, we distinguish *basic emotions* from *basic motivations* in that basic emotions need not be related to a perceived difference between an actual and a desired state. Furthermore, basic emotions themselves can be states that the agent does or does not desire whereas basic motivations are directed towards or away from what the agent desires. "Fear", for example, in and of itself is an undesirable state of an agent in that it is indicative of danger. As such, it causes the agent to behave in such a way as to be prepared for or avoid danger. Hence, while "fear" can be also motivational in the sense that it may move the agent away from the cause of fear it is also emotional as it itself is not a desired state. A fear state with no clearly discernible danger present, which causes an agent to be more cautious and alert, may itself not instantiate a motivational state that is connected to a particular goal such as running away from a particular threat.

"Fear", as discussed above, can be instantiated by a controller C, which integrates over time the frequency of occurrence of fear triggering conditions. Input to C comes from an internal sensor S that is activated by a fear triggering condition. C integrates these inputs over time and outputs a signal that corresponds to the intensity of "fear" and modulates behavior to be more alert and ready for sudden activity. A neural control circuit implementing an appropriate response characteristic (similar to that given by $g(t) = e^{-t}$ to a unit impulse, which is generated by the sensor or the perceptual system detecting a dangerous stimulus), could use an *interactive activation and competition* (IAC) unit (McClelland & Rumelhart, 1988), whose change in activation is given by $\Delta act = S \cdot g_S \cdot act + decay \cdot act$, where act is the current activation level of the control system, g_S is the gain for the sensor input and $decay$ is the discount value for past activations.

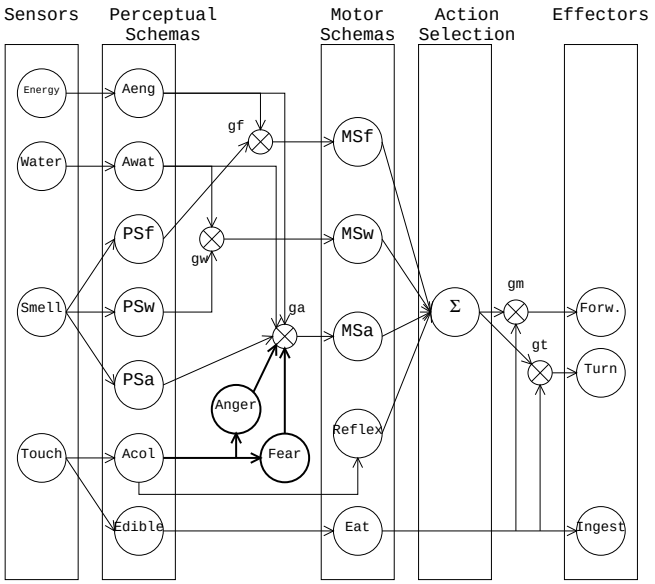


Figure 1: The schema-based architecture for the simulated emotional agents (see the text for details).

To be able to instantiate a fear state, the above controller needs to control the agent's effectors in a way that the positive output from the controller can in uence and bias the agent's behavior towards avoiding or attempting to avoid dangerous objects. As such, the intensity with which the agent avoids or attempts to avoid these objects depends on the magnitude of the output of the controller: the agent's behavior is modulated by its level of fear.

A Schema-Based Agent Architecture

Using the above control elements to implement basic motivations and emotions, we have compared the performance of agents with mechanisms to implement *fear*

and *anger* to that of agents without these mechanisms in a hostile multi-agent environment, where agents need to forage for resources in order to survive and procreate. The employed architecture is a biologically plausible schema-based architecture (Arbib, 1992) for both agent kinds, which allows the agents to forage for food and water. In this architecture, the behavior of an agent depends at any given time on the relative contributions from a variety of motor schemas. While non-emotional agents have fixed behavioral dispositions to deal with competitors for resources, emotional agents use their emotional control circuits to adapt their behaviors based on past encounters.

Figure 1 shows the architecture for the emotional agents (their emotional subsystem is an implementation of the higher-level functional organization of the basic mammalian "fear/anger system" in the terms of the above suggested control units, e.g., (Berkowitz, 2003)). Schemas are depicted by large circles where the names indicated their function.[1] Small crossed circles indicate gains of schemas (i.e., behavioral dispositions) that are taken as architectural parameters to be varied in the experiments: the degree to which an agent is attracted to food (g_f), to water (g_w), and to other agents (g_a). The bold-face circles labeled "Fear" and "Anger" represent the "fear schema" and "anger schema", respectively. They are only present in the architecture of emotional agents. Both emotion schemas are connected to an "alarm schema" (*Acol*), which is triggered if an agent touches other agents. This mechanism changes the agent's propensity to fight other agents or to ee: the higher the output of a controller, the more stronger the behavioral disposition (i.e., to fight for anger, or ee for fear).

More formally, let $Ent = \{f, w, a\}$ be an index set of the three types of objects in the simulation environment: *food*, *water*, and *agents*. For each object type in Ent, a force vector F_i is computed, which is the sum, scaled by $1/|v|^2$, of all vectors v from the agent to the objects of type i within the respective sensory range, where $|v|$ is the length of vector v. These *perceptual schemas* are mapped into motor space by the transformation function $T(x) = \sum_{i \in Ent} g_i \cdot F_i(x)$, where the g_i are the respective gain values of the perceptual schemes. The gain values simply scale the effect of sensory input, providing a means by which to prioritize certain inputs (e.g., if food is especially important, its gain value could be higher than the other gain values, so that sensing food has a greater impact on the direction chosen than sensing other entities).

All feedback controllers are implemented in a feed-forward three-layer interactive activation and competition neural network (with three input units *in*, three hidden units *hid*, and three output units *out*). The input units receive their activations (via appropriate scaling functions) from the $Water$ (in_w) and $Energy$ level sensors (in_f) via the perceptual *Awat* and *Aeng* schemas as well as from the $Touch$ sensor via the *Acol* schema (in_a), respectively.

The output units are connected to the gain values in the motor scheme via individual scaling functions $f_i(x) = x \cdot c_i + b_i$ (where b_i is the *base gain value* and c_i the scaling factor for the activation of out_i).

The activation value $act_i(t)$ of an IAC unit i at time t is defined by

$$act_i(t) = \begin{cases} (max \quad act_i(t \quad 1)) \cdot net_i(t) \quad decay, \\ \qquad\qquad\qquad\qquad net_i(t) \geq 0 \\ (act_i(t \quad 1) \quad min) \cdot net_i(t) \quad decay, \\ \qquad\qquad\qquad\qquad net_i(t) < 0 \end{cases}$$

where min and max are the minimum and maximum activation level, respectively, *decay* is a decay factor defined by $d \cdot (act_i(t) \quad rest)$ (where d is a constant), $rest$ the rest level, and $net_i(t)$ the weighted sum of all inputs to unit i at time t.

The choice of IAC units over standard perceptrons is based on their update rule, which is particularly suited to implement important temporal features of some emotional states in that it (1) takes into account the *previous activation* (hence, can be used to implement "inner states"), and (2) incorporates a *decay term* to raise or lower the activation to a predetermined *base level*.

Non-emotional agents have a constant g_a gain (i.e., their $c_i = 0$), hence their behavioral dispositions towards other agents are fixed. Emotional agents, on the other hand, can adapt their behavior dispositions, i.e., their g_a gain, by virtue of the feedback controllers implemented in the neural net (their $c_i \neq 0$). Depending on whether g_a is positive or negative, they can implement basic "anger" or "fear" states (as argued in (Scheutz, 2001)).

The Utility of Anger and the Limits of Fear

We report results from two classes of experiments studying the role of emotions in foraging and survival tasks.[2] In the first class, the gain g_a is set to a negative value for both agent kinds, thus making them disposed to avoid other agents. For the second class, g_a is positive for both agent kinds, thus making them disposed to be aggressive towards other agents. Performance was measured in terms of the number of surviving agents after 10000 simulation cycles averaged over 40 runs with random initial conditions. The upper and lower parts of Figure 2 show the results from both classes of experiments for both agent kinds for two architectural variations: agent gain and water gain (i.e., 25 sets of 40 experimental runs each). All runs started with 10 agents of each of the two kinds placed at random location in the environment together with 20 randomly placed food and 20 randomly placed water items; new food and water items are generated on every 4 and 6 cycles in random locations, respectively.

While emotional agents in the first set have a performance peak (of 23.625) that is slightly higher than that of non-emotional agents (of 23.35), the difference is not

[1] For space reasons we cannot describe all the details of the architecture here.

[2] For more details about simulation setup and simulation parameters see (Scheutz, 2001).

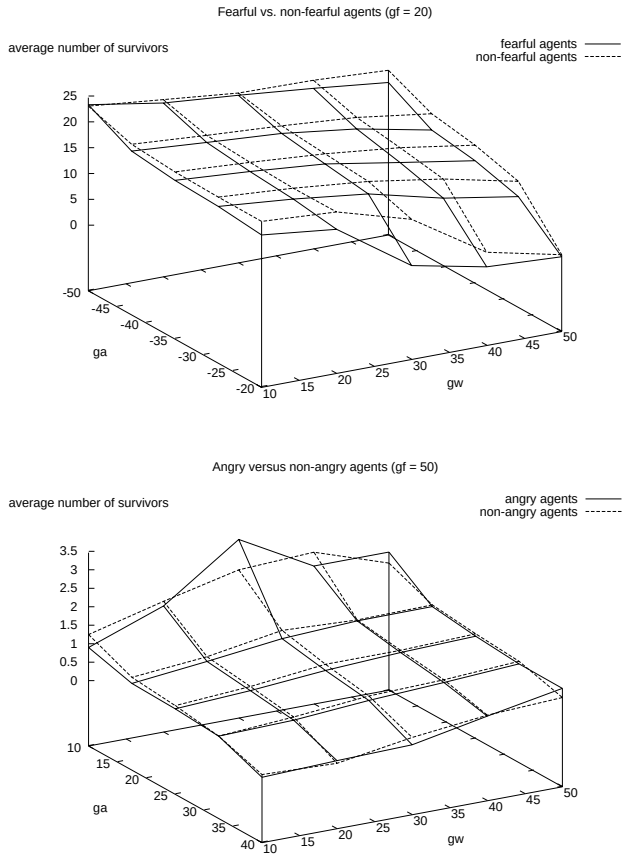


Figure 2: The performance space of the emotional vs. and non-emotional agents (fearful top, angry bottom) based on variations along two architectural dimensions.

significant (t-test, $p > 0.1$). Consequently, being fearful in addition to having the behavioral disposition of avoiding other agents does not increase the overall performance, it may in fact reduce it for some settings of the gain values (e.g., $g_a = \ \ 20$ and $g_w = 30$). For emotional agents in the second class of experiments, however, we find a marginally significant global maximum at $g_a = 10$ and $g_w = 30$. Consequently, in the kinds of environments studied, "anger" does sometimes prove useful for survival.

Discussion

The results reported here are only a very small part of a large set of experiments, in which up to five architectural dimensions were varied in an effort to determine the circumstances in which emotional control is beneficial and where it might be detrimental. The methodology on which they are based consists of a four step process: (1) emotion concepts are analyzed and defined in terms of architectural capacities of agent architectures (Sloman, 2002). (2) Agent architectures with particular emotional control mechanisms (as defined in (1)) are constructed for a given task, for which also a performance measure is defined. (3) Simulations experiments are carried out with the so-defined emotional agents and their performance is determined for a predetermined set of architectural and environmental variations. The outcome then is a *performance space* that corresponds to the varied parameters. The last two steps are repeated with agents implementing non-emotional (or, in general, other) architectures. (4) All resulting performance spaces are then compared with respect to the agents' *performance-cost tradeoffs*, i.e., their performance taken relative to the (computational) cost necessary to maintain and run the instantiated architecture (in the reported experiments the cost for both architectures was taken to be the same). The last point is crucial as it may well be that emotional agents do not perform better than non-emotional ones on a given task in absolute terms, but that they do much better in relative terms, i.e., with fewer resources (which is usually believed to be the case by emotion researchers). Especially from an evolutionary perspective relative performance is the relevant measure.

We believe that the proposed methodology to experiment with agent architectures in an artificial life environment cannot only form the basis for a thorough comparison of the different emotion models that can otherwise not be studied easily (e.g., social emotions and their role in the control of agents), but can also inform emotion researchers interested in clinical aspects of emotions by performing simulated *lesion studies*, where parameters of functional agents are modified or components of the architecture are removed. This, in turn, might help us isolate not only the functional roles of emotions in the control of creatures, but also the ways in which emotional control can fail and how it might be possible to reestablish normal functioning in dysfunctional systems.

References

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Arbib, M. (1992). Schema theory. In S. Shapiro (Ed.), *The handbook of brain theory and neural networks* (pp. 830–833). MIT Press.

Arkin, R., Fujita, M., Takagi, T., & Hasegawa, R. (2003, March). An ethological and emotional basis for human-robot interaction. *Robotics and Autonomous Systems*, *42*, 3-4.

Armony, J., Servan-Schreiber, D., Cohen, J., & LeDoux, J. (1995). An anatomically constrained neural network model of fear conditioning. *Behavioral Neurosciences*, *109*(2), 256–257.

Balkenius, J. M. C. (2000). A computational model of emotional learning in the amygdala. *Cybernetics and Systems*, *32*(6), 611–636.

Berkowitz, L. (2003). Affect, aggression, and antisocial behavior. In (Davidson, Scherer, & Goldsmith, 2003) (pp. 804–823).

Blair, H., Tinkelman, A., Moita, M., & LeDoux, J. (2003). Associative plasticity in neurons of the lateral amygdala during auditory fear condition-

ing. *Annals of the New York Academy of Sciences*, *985*, 485–487.

Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin*, *99*(2), 229–246.

Bless, H., Schwarz, N., & Wieland, R. (1996). Mood and the impact of category membership and individuating information. *European Journal of Social Psychology*, *26*, 935-959.

Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.

Clore, G., Gasper, K., & Conway, H. (2001). Affect as information. In J. Forgas (Ed.), *Handbook of affect and social cognition* (p. 121-144). Mahwah, NJ: Erlbaum.

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 91–115). NY: Guilford.

Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (Eds.). (2003). *Handbook of affective sciences*. New York: Oxford University Press.

Derryberry, D., & Tucker, D. (1994). Motivating the focus of attention. In P. Neidenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influence in perception and attention* (pp. 67–96). San Diego, CA: Academic Press.

Dulk, P. den, Heerebout, B., & Phaf, R. (2003). A computational study into the evolution of dual-route dynamics for affective processing. *Journal of Cognitive Neuroscience*, *15*(2), 194–208.

Fanselow, M. S., & Gale, G. D. (2003). The amygdala, fear, and memory. *Annals of the New York Academy of Sciences*, *985*, 125–134.

Fredrickson, B. (1998). What good are positive emotions? *Review of General Psychology*, *2*, 300–319.

Frijda, N. H. (2000). The psychologists' point of view. In (Lewis & Haviland-Jones, 2000) (pp. 59–74).

Grossberg, S., & Schmajuk, N. (1987). Neural dynamics of attentionally-modulated pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology*, *15*, 195–240.

Kahneman, D., Wakker, P., & Sarin, R. (1997). Back to bentham? explorations of experienced utility. *Quarterly Journal of Economics*, *112*, 375-405.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

Lewis, M., & Haviland-Jones, J. M. (Eds.). (2000). *Handbook of emotions* (2nd ed.). New York: The Guilford Press.

Marsella, S., & Gratch, J. (2002, May). Modeling the in uence of emotion on belief for virtual training simulations. In *Proceedings of the 11th conference on computer-generated forces and behavior representation.* Orlando, FL.

McClelland, J. L., & Rumelhart, D. E. (1988). *Parallel distributed processing* (Vol. 1 and 2). Cambridge: MIT Press.

McFarland, D. (1981). *The oxford companion to animal behavior*. Oxford: Oxford University Press.

Michaud, F., & Audet, J. (2001). Using motives and artificial emotion for long-term activity of an autonomous robot. In *5th autonomous agents conference* (pp. 188–189). Montreal, Quebec: ACM Press.

Nerb, J., & Sperba, H. (2001). Evaluation of environmental problems: A coherence model of cognition and emotion. *Cognition and Emotion*, *4*(15), 521–551.

Newell, A. (1990). *Uni ed theories of cognition*. Cambridge, MA: Harvard University Press.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of the emotions*. New York: Cambridge University Press.

Özbay, H. (2000). *Introduction to feedback control theory*. London: CRC Press.

Rolls, E. T. (1995). A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1091–1106). Cambridge, MA: MIT Press.

Scheutz, M. (2001). The evolution of simple affective states in multi-agent environments. In D. Cañamero (Ed.), *Proceedings of aaai fall symposium* (pp. 123–128). Falmouth, MA: AAAI Press.

Scheutz, M., & Schermerhorn, P. (2002). Steps towards a theory of possible trajectories from reactive to deliberative control systems. In R. Standish (Ed.), *Proceedings of the 8th conference of arti cial life.* MIT Press.

Scheutz, M., & Schermerhorn, P. (2004). The role of signaling action tendencies in con ict resolution. *Journal of Arti cial Societies and Social Simulation*, *7*(1).

Scheutz, M., & Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In N. Zhong, J. Liu, S. Ohsuga, & J. Bradshaw (Eds.), *Intelligent agent technology: Research and development* (pp. 200–209). New Jersey: World Scientific Publisher.

Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. In E. Higgins & R. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 2, p. 121-144). New York: Guilford Press.

Sloman, A. (2002). Architecture-based conceptions of mind. In *Proceedings 11th International Congress of Logic, Methodology and Philosophy of Science* (pp. 397–421). Dordrecht: Kluwer. ((Synthese Library Series))

Emergent Meaning in Affective Space: Conceptual and Spatial Congruence Produces Positive Evaluations

Simone Schnall (schnall@virginia.edu)
University of Virginia, Department of Psychology, 102 Gilmer Hall
Charlottesville, VA 22904 USA

Gerald L. Clore (gclore@virginia.edu)
University of Virginia, Department of Psychology, 102 Gilmer Hall
Charlottesville, VA 22904 USA

Abstract

Based on the theory of conceptual metaphor we investigated the evaluative consequences of a match (or mismatch) of different conceptual relations (good vs. bad; abstract vs. concrete) with their corresponding spatial relation (UP vs. DOWN). Good and bad words that were either abstract or concrete were presented in an up or down spatial location. Words for which the conceptual dimensions matched the spatial dimension were evaluated most favorably. When neither of the two conceptual dimensions matched the spatial dimension, ratings were not as favorable as when the dimensions did match, but were still significantly more favorable than when one conceptual category was matched with the spatial category (e.g., UP and abstract), while the other one was not (e.g., UP and bad). Results suggest that a metacognitive feeling of fluency can produce an additional layer of evaluative information that is independent of actual stimulus valence.

Background

A recent theory of conceptual structure proposes that bodily processes influence and constrain cognitive information processing, and that the resulting knowledge is structured in a largely metaphorical way (Gibbs, 1994; Lakoff & Johnson, 1980, 1999). According to this view, the body is a source of knowledge, and by means of conceptual metaphors, very basic "embodied" concepts are mapped onto more abstract concepts. For instance, the basic orientation of the human body in space (certain things are "up" or "down", relative to the body) is used when conceptualizing abstract categories, such as emotions, when metaphorically talking about "feeling up" or "down". Thus, metaphor, defined as "*understanding and experiencing one kind of thing in terms of another* (Lakoff & Johnson, 1999, p. 5, emphasis in original)," does not merely concern language usage. How we use metaphor to talk about things also has implications for how we act upon, and think about those things.

Central to the theory of conceptual metaphor is the notion of "image schema" (Johnson, 1987, 1999), which describes a pattern of perceptual experience that emerges from very basic bodily activities. For instance, the sensorimotor experience of moving with one's own body through space results in the image schema of VERTICALITY, or the understanding that we usually function in an upright position, with a clear up-down orientation. Indeed, spatial perceptions and spatial language are closely intertwined (Hayward & Tarr, 1995; Miller & Johnson-Laird, 1976; Richardson, Spivey, Barsalou, & McRae, 2003; Tolaas, 1991). Spatial metaphors derived from the concrete concept of VERTICALITY can be used to describe various abstract concepts:

> GOOD IS UP; BAD IS DOWN:
> Things going *downhill*; Feeling *down* in the dumps
> ABSTRACT IS UP; CONCRETE IS DOWN:
> *Higher-order* categories; *sub*-types

Thus, many image schemata are hypothesized to be derived from the basic experience of the body functioning in three-dimensional space: By definition, all human behavior takes place in space. As a consequence, the source domain of space provides a metaphor for multiple target domains (Lakoff & Johnson, 1980, 1999).

Indeed, evidence has been obtained supporting the notion that concrete spatial relations provide opportunities for mapping conceptual relations. For instance, time is often conceptualized as movement through space (Boroditsky, 2000; Boroditsky & Ramscar, 2002; Gentner & Imai, 1992; Gentner, Imai, & Boroditsky, 2002). Graphs are easier to understand when an increase in quantity is represented by an increase in slope, corresponding to the spatial metaphor of MORE IS UP (Gattis & Holyoak, 1996). Inferences about given premises are more accurate when the premises are mapped onto a spatial medium, compared to when they are not (Schnall & Gattis, 1998). Positive words are categorized faster when they are presented in an upward location, whereas negative words are categorized faster whey they are in a downward location (Meier & Robinson, 2004). Thus, when spatial relations can be mapped onto corresponding conceptual relations, cognitive operations are facilitated.

In the current study we investigated the confluence of spatial relations and conceptual relations. As noted earlier, spatial concepts, such as VERTICALITY, serve as the source domain for various target domains, such as goodness/badness, and concreteness/abstractness. Figure 1

describes the relationship between the source domain VERTICALITY and those two target domains. Both target domains have an implicit connection with the source domain. For example, the concept "love" is good and abstract, and both those conceptual categories are conceptualized as UP.

Table 1: Relationship between spatial and conceptual dimensions.

		CONCEPT	
		VALENCE	CONCRETENESS
SPACE	UP	Good	Abstract
	DOWN	Bad	Concrete

The goal of the current study was to investigate the effect of a match (or mismatch) of two different conceptual relations with their corresponding spatial relation. The spatial dimension we investigated was VERTICALITY (UP vs. DOWN), and the two conceptual dimensions were VALENCE (*good* vs. *bad*) and CONCRETENESS (*concrete* vs. *abstract*). The spatial dimension was varied by a simple perceptual manipulation: The stimulus word, which consisted of a good (or bad) word that was either abstract (or concrete) was placed either on top, or on the bottom of the page on which participants evaluated the word.

Table 2 describes the different ways in which the relations can either be matched or mismatched. *Congruent relations* are present when both relations are matched, such as when good, abstract concepts are UP (top panel of Table 2, denoted by "+ +"), but also when both relations are mismatched, such as when bad, concrete concepts are UP (denoted by "- -"). In the latter case neither of the conceptual dimensions matches the spatial dimension, therefore no conflict between the spatial and conceptual relations exists. In contrast, *incongruent relations* are present when only one of the two conceptual dimensions matches the spatial dimension, and the other one does not.

We expected that the extent to which the two conceptual dimensions were in accordance with the spatial dimension would influence the perceived valence of the stimulus words. Specifically, we predicted that when both relationships are matched, as is the case for the congruent "+ +" conditions, stimuli should be rated more favorably. Further, for the congruent "- -" conditions, stimuli should be perceived as less positive than in the congruent "+ +" conditions, but as more positive than in any of the incongruent "+ -" "- +" conditions. This hypothesis regarding the mapping of spatial and conceptual dimensions was tested by presenting participants with strongly positive and strongly negative words that were either abstract or concrete, and thus the content of the words crossed both conceptual dimensions of goodness/badness and abstractness/concreteness. Each word was presented either on top, or on the bottom of a piece of paper, and participants evaluated how good the word was.

Table 2: Congruent (" + +" "- -") and incongruent ("+ -" "- +") relations of spatial and conceptual dimensions.

Stimulus	SPACE	VALENCE Good = UP Bad = DOWN	CONCRETENESS Abstract = UP Concrete = DOWN
Good Abstract e.g., *talent*	UP	+	+
Good Concrete e.g., *palace*	UP	+	−
Bad Abstract e.g., *malice*	UP	−	+
Bad Concrete e.g., *bullet*	UP	−	−
Bad Concrete e.g., *blister*	DOWN	+	+
Bad Abstract e.g., *neglect*	DOWN	+	−
Good Concrete e.g., *circus*	DOWN	−	+
Good Abstract e.g., *passion*	DOWN	−	−

Method

Participants

Participants were 61 undergraduate students from the University of Virginia who received course credit.

Procedure

Participants filled out a survey as part of an experimental session. Instructions specified that the participant's task was to make a judgment about how good or how bad certain words were. It was emphasized to participants that they should go with their first intuition, and that judgments should be made according to what they personally thought, rather than what other people might think.

The word stimuli were presented in the following manner (see Figures 1 and 2). Each stimulus was printed on a separate sheet of paper measuring 4 ¼ by 5 ½ inches. A horizontal line was drawn in the middle of the paper to emphasize up and down locations. The stimulus word was printed either in the space on top of the line (in the upper half of the page), or below the line (in the lower half of the page). Each word was followed by a rating scale on which the participant evaluated the word from 1 (very good) to 7 (very bad). All stimuli were assembled into a booklet that presented the stimuli in a fixed random order. Half of the stimuli were strongly positive words, the other half were strongly negative words. These words were selected from a word list for which normative affective ratings have been established (Bradley & Lang, 1999), and were matched for word length and word frequency. For each valence, half of the words were abstract (e.g., "honor," "greed"), the other half were concrete (e.g., "bouquet," "thief"), thus resulting in eight different experimental conditions. Each stimulus was presented only once, and each participant received all conditions.

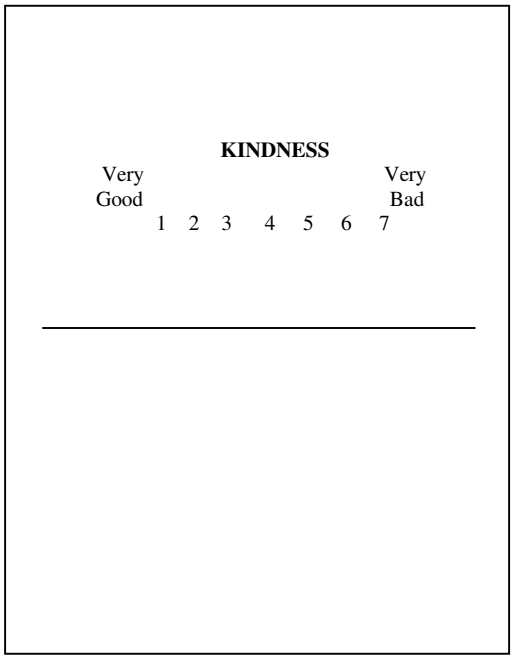


Figure 1: Spatial set-up of survey: "Good"abstract item in upward location.

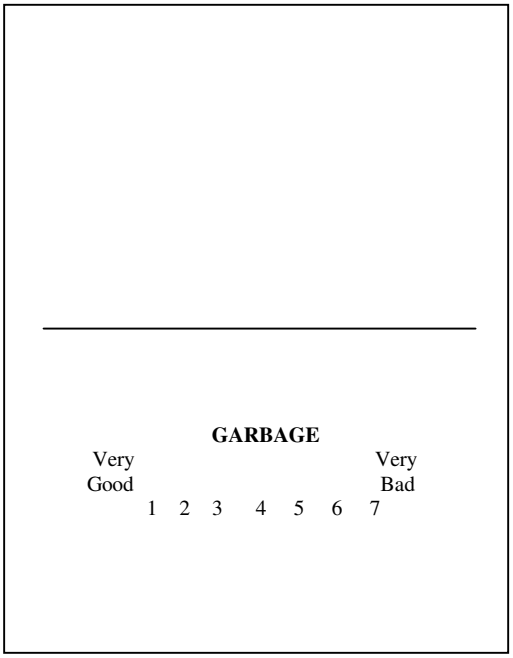


Figure 2: Spatial set-up of survey: "Bad"concrete item in downward location.

Table 3: List of word stimuli.

Good Words		Bad Words	
Abstract	Concrete	Abstract	Concrete
joy	toy	sin	fat
fun	kiss	scorn	tomb
wise	gift	greed	bomb
honor	jewel	upset	thief
brave	dinner	devil	bullet
talent	palace	deceit	prison
fantasy	circus	malice	poison
miracle	delight	misery	blister
passion	sunset	hatred	morgue
kindness	bouquet	failure	garbage
intimate	treasure	neglect	hostage
ambition	sunlight	jealousy	mosquito
affection	butterfly	ignorance	hurricane

Results

Because the valence of the words was strongly positive or strongly negative, and no interaction effect of valence and spatial position was expected (i.e., positive words were not expected to be rated as negative, or negative words as positive depending as a function of their spatial location), separate within-subjects ANOVAs were conducted for positive and negative items. For "good" words, the

interaction of spatial position and level of abstractness was significant, $F(1, 60) = 64.48$, $p < .0001$, with the highest mean for the congruent condition, namely abstract positive words presented in the up location ($M = 5.43$, $SD = .40$)(see Figure 3). Subsequent paired-samples t-tests showed that words in the congruent condition received significantly higher positive ratings than concrete positive words in the up location ($t(60) = -15.95$, $p < .0001$), abstract positive words in the down location ($t(60) = -2.18$, $p < .03$), and concrete positive words in the down location ($t(60) = -7.91$, $p < .0001$). Thus, when the perceptual dimension (UP) was matched with both conceptual dimensions (*good* and *abstract*), evaluations of the positive words became even more positive compared to when they were not.

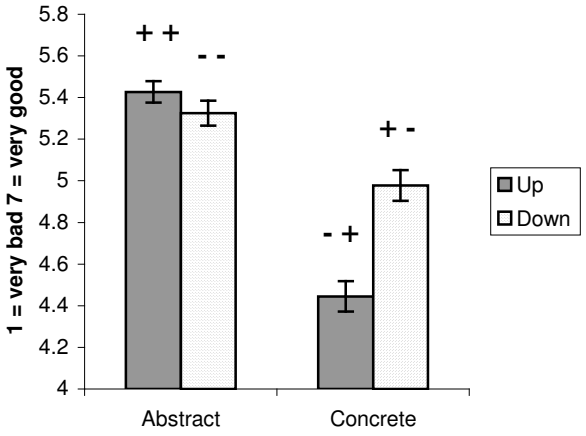


Figure 3: Mean ratings for "good" words.

For the "bad" words, there was also a significant interaction of space and level of abstractness, $F(1, 60) = 54.60$, $p < .0001$. Words in the congruent condition, that is, concrete negative words presented in the down location received the most *positive* ratings ($M = 1.39$, $SD = .39$), and differed significantly from abstract negative words in the down location ($t(60) = 11.06$, $p < .0001$), concrete negative words in the up location ($t(60) = 9.24$, $p < .0001$), and abstract negative words in the up location ($t(60) = 10.68$, $p < .0001$) (see Figure 4). Remarkably, this match between the perceptual dimension and its corresponding two conceptual dimensions did not result in making the negative words more extreme, and thus more negative; rather, as was the case for the positive words, it led to more positive ratings for the negative words.

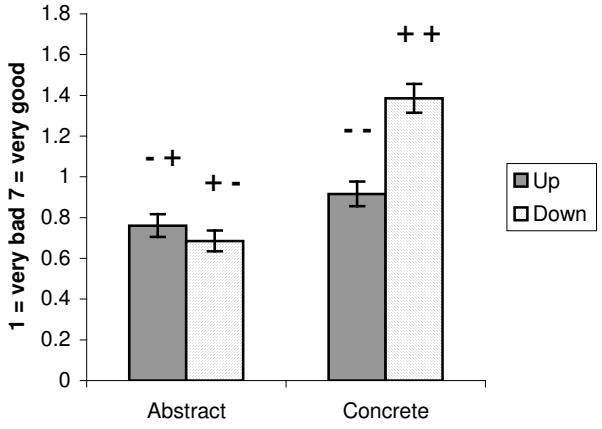


Figure 4: Mean ratings for "bad" words.

The analyses so far have dealt with the "+ +" match conditions of perceptual and conceptual dimension, denoted with two plus signs in Table 2. In addition, more positive ratings were also observed for the " - -" match conditions, denoted with two minus signs in Table 2, where the spatial dimension neither matched the valence, nor the level of abstractness of the words.

To compare specifically the congruent "+ +" with the congruent "- -" conditions, as well as with the incongruent "+ -" "- +" conditions of spatial dimension with the two conceptual dimensions, composite scores were computed. As predicted, the composite average rating for the two congruent "+ +" conditions ($M = 3.41$, $SD = .23$) was significantly more positive than the composite average rating for the two congruent "- -" conditions ($M = 3.12$, $SD = .26$), $t(60) = -8.49$, $p < .0001$ (see Figure 7). In addition, the composite average rating for the two congruent "- -" conditions was significantly more positive than the incongruent "+ -" "- +" conditions ($M = 2.72$, $SD = .23$), $t(60) = -9.83$, $p < .0001$.

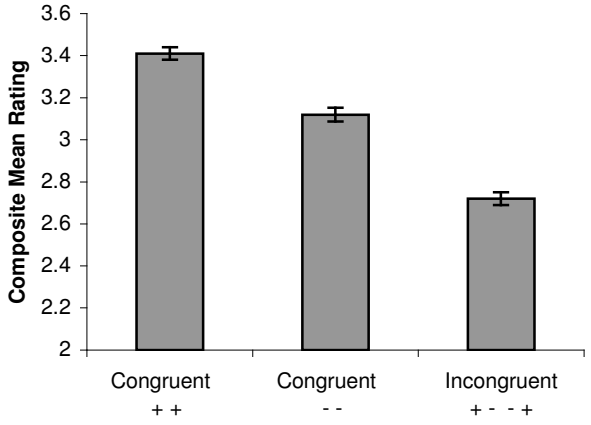


Figure 5: Composite ratings for congruent and incongruent matching conditions.

To summarize, abstract positive words presented in an up location, and concrete negative words presented in a down location were evaluated most favorably. Furthermore, when neither of the two conceptual dimensions matched the spatial dimension, that is when abstract good words were down, or when concrete bad words were up, ratings were not as favorable as when the dimensions did match, but were still significantly more favorable than when one conceptual category was matched with the spatial category (e.g., UP and abstract), while the other one was not (e.g., UP and bad).

Discussion

We found evidence for a connection between a spatial source domain and two conceptual target domains. A match between the source domain of VERTICALITY and the two corresponding target domains of goodness/badness and abstractness/concreteness resulted in more positive evaluations for affectively toned material, regardless of whether this material was positive or negative in valence. Abstract good things were rated as even better when they were presented on top of the page, but concrete bad things were also rated as better when they were presented on the bottom of the page.

The finding regarding the negative words might be considered surprising. A different outcome might have been that the meaning of bad things was intensified with congruent spatial and conceptual dimension, so that bad things became even worse. But this is not what we found. How can this somewhat counterintuitive finding be explained? One possibility is that if conceptual relations are indeed as inherently connected to spatial relations as our findings suggest, then people are more familiar with spatially represented conceptual structure. As a consequence of this familiarity, these mappings are experienced as more pleasant. Indeed, it has been well documented that the repeated presentation of a stimulus is sufficient to increase positive affect toward that stimulus, relative to a stimulus that has not been presented repeatedly. In a classic study originating the work on the so-called mere exposure effect, Zajonc (1968) presented Chinese-looking characters, nonsense words, or yearbook photographs for either 0, 2, 5, 10 or 25 times to participants. Participants subsequently rated how "good" or "bad" the meaning of the Chinese characters, or of the nonsense words was, and how much they liked the person shown in the photographs. For all three kinds of stimuli, participants' ratings became more positive with increased number of presentation. Many studies have since replicated and extended this basic effect, suggesting that the mere exposure effect is a very robust phenomenon (Bornstein, 1989).

Thus, it is possible that external conceptual organization that conforms with one's own representational structure is perceived as more familiar, and therefore, as more pleasing. Perhaps our results may be regarded a *representational mere exposure effect*, where the highest positive valence is assigned to those conceptual organizations that have the highest degree of familiarity. In this regard, it is instructive to review the explanations that have been put forward to explain the mere exposure effect.

Some have proposed that fluency of cognitive operations can explain why people like things better the more often they experience them. Fluency refers to properties of continuous information processing, such as the speed, or ease of processing (Jacoby, Kelley, & Dywan, 1989). These properties emerge as a feature of the *process*, rather than the *content* of cognitive functioning (Winkielman, Schwarz, Fazendeiro, & Reber, 2003). Generally, high fluency, that is, fast and effortless processing of information, signals positive states of the environment, and of one's cognitive processes. As a consequence, fluency can result in positive affect, as well as positive evaluations of target stimuli toward which fluency is experienced. Research on this effect has generated compelling evidence for the fluency hypothesis (Winkielman et al., 2003).

Further, studies involving affective evaluations in particular demonstrate an asymmetric effect, such that only positive evaluations, but not negative evaluations, are influenced by fluency manipulations, regardless of how questions concerning the ratings are worded. For instance, Reber et al. (1998) found that high fluency led to increased judgments of liking and decreased judgments of disliking. Similarly, Winkielman and Cacioppo (2001) instructed half of their participants to report positive affect, and half of the participants to indicate negative affect after a fluency manipulation. Only those reporting on positive affect showed increased positive affect when exposed to high fluency, whereas those reporting negative affect did not show such an effect. Our finding that congruence between spatial and conceptual dimensions led to increased positive ratings even for negative words is consistent with this documented asymmetric effect where only positive evaluations increase as a function of fluency, but not negative evaluations.

An additional finding in the present study was that not only words in the congruent "+ +" conditions, but also words in the congruent "- -" conditions were rated more positively than words in the incongruent "+ -" "- +" conditions. Other data are consistent with the present finding that sometimes two negatives combine to make a positive, so to speak. For instance, according to the *affective certainty* model (Tamir, Robinson and Clore, 2002), when personality traits match with current affective states, people experience facilitated performance on motivationally relevant cognitive tasks. Thus, people who are generally happy, and who found themselves in a happy mood, were more successful at processing affectively valenced information, but the same was true of people who are generally unhappy and found themselves in an unhappy mood, compared with people who are in conflict regarding their beliefs about themselves, and their actual experiences. Similarly, in the present study, more fluency, and therefore higher positive ratings was the result of a lack of representational conflict between the source domain and the

target domain, even if that meant that neither of the target domains could be mapped onto the source domain.

In conclusion, we found that "metaphorical" mappings between inherent spatial and conceptual relations can produce an additional layer of complexity, where the confluence of source domain and target domains has emergent affective properties: Both good and bad things are evaluated as more positive when an explicit spatial representation fits with implicit conceptual structure. In such situations, metacognitive processes involving perceived fluency provide information that goes well beyond representational content itself.

Acknowledgments

Support from NIH grants 1R03MH67580-01 (to S. S.) and 5R01MH050074-06 (to G. L. C.) is acknowledged. We thank Steven Cholewiak, Jeffrey Claiborne and Irina Komarovskaya for assistance in data collection, and Jeanine Stefanucci for helpful comments on an earlier draft of the manuscript.

References

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968-1987. *Psychological Bulletin, 106,* 265-289.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75,* 1-28.

Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science, 13,* 185-189.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English Words (ANEW): Stimuli, instruction manual and affective ratings.* Technical Report C-1, Gainesville, FL: Center for research in psychophysiology, University of Florida.

Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22,* 231-239.

Gentner, D., & Imai, M. (1992). Is the futures always ahead? Evidence for system mappings in understanding space-time metaphors. *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society (pp. 510-515).*Hillsdale, NJ: Erlbaum.

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space time metaphors. *Language and Cognitive Processes, 17,* 537-565.

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge: Cambridge University Press.

Hayward, W. G., & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition, 55,* 39-84.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving*. Hillsdale, NJ: Erlbaum.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason.* Chicago: University of Chicago Press.

Johnson, M. (1999). Embodied Reason. In G. Weiss & H. Fern Haber (Eds.), *Perspective on embodiment: The intersections of nature and culture.* London: Routledge.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought.* New York: Basic Books.

Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up: Associations between affect and vertical position. *Psychological Science, 15,* 243-247.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception.* Cambridge, MA: Harvard University Press.

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science, 27,* 767-780.

Schnall, S., & Gattis, M. (1998). Transitive inference by visual reasoning. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 929-934). Hillsdale, NJ: Erlbaum.

Tolaas, J. (1991). Notes on the origin of some spatialization metaphors. *Metaphor and Symbolic Activity, 6,* 203-218.

Winkielman, P., & Cacioppo, J. T. (2001). Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation increases positive affect. *Journal of Personality and Social Psychology, 81,* 989-1000.

Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion.* Mahwah, NJ: Erlbaum.

Sensitivity to Confounding in Causal Inference: From Childhood to Adulthood

E. Christina Schofield (christis@ucla.edu)
Department of Psychology, Box 951563
Los Angeles, CA 90095-1563 USA

Patricia W. Cheng (cheng@psych.ucla.edu)
Department of Psychology, Box 951563
Los Angeles, CA 90095-1563 USA

Abstract

A necessary condition for correctly assessing causality is the absence of confounding causes. This set of experiments assesses whether people are sensitive to confounding when they infer causation. Two stories were constructed, one in which two candidate causes and an outcome perfectly covaried (confounded), and one in which one candidate cause occurred independently of a second candidate cause that perfectly covaried with an outcome (unconfounded). If people reason by forming contrast groups that hold alternative causes constant, then in the confounded case, subjects should say that it is impossible to determine causality when two candidate causes are confounded. In the unconfounded case, subjects should be able to say that the candidate is causal. If people are not sensitive to confounding, then they should attribute causality to both candidates in the confounded case, and the results for the target candidate should be the same as those in the unconfounded case. This experiment was conducted with children and adults. Children saw one of the two conditions, while adults saw both conditions. Both children and adults make a distinction between confounded and unconfounded causes when making attributions of causality. Our results show that children are able to state the indeterminacy of confounded causes at an age much earlier than previously documented.

Introduction

One view of how children learn is that they approach the world as scientists and form theories about the world (e.g. Gelman, 1996) using information about variation and covariation to establish causal connections (Gopnik, Sobel & Schulz, 2001). Further, they intervene upon the world in order to discover these relationships (Schulz, 2003). Although children may have misconceptions in their explanations, as when a child states that he thinks God made the sun out of gold and lit it with fire (Siegler, 1998), the presence of such misconceptions does not mean that children are unable to use the data present in the environment to form correct causal attributions. Given that adults have had many more experiences than children, we should not expect children's theories to be the same as adult's theories, especially for complex phenomena. What is important is whether the same process is utilized when determining causality. In particular, this paper seeks to examine whether both children and adults are sensitive to confounding when there are two candidate causes for a novel outcome.

As well as being a potential means of improving science instruction and for examining whether children assess causality in the same manner as adults, assessing children's sensitivity to confounded causes is also important for differentiating between two models of causal attribution: the unconditional ΔP model (Jenkins & Ward, 1965), and the focal sets approach (Cheng and Novick, 1992).

Under the unconditional ΔP model, people compare the frequency of e, an effect of interest, when c, a potential cause, is present, with the frequency of e when the potential cause is not present:

$$\Delta P = P(e|c) - P(e|\sim c)$$

If ΔP = 0, then the candidate is considered noncausal; if ΔP is noticeably greater than 0, then the candidate is thought to cause the outcome, and if ΔP is noticeably less than 0, the candidate is thought to prevent the outcome. In the unconditional version of this model, people ignore confounding and pool over all the information known about the candidate cause. Using this formulation, if both candidate causes perfectly covary with each other and the outcome, then both will be judged as causal. Under the focal sets approach, the same formula can be used, but is evaluated only when comparing across groups where alternative causes are held constant. If people utilize the focal sets approach, they could make a determination when no confounding was present, but would be indeterminate in the case of perfect covariation because no focal set could be formed.

One study looking at third, sixth and ninth graders, as well as non college young adults and undergraduate college students found that before the ninth grade, students were unlikely to state that there was insufficient evidence to determine causality when there is confounding (Kuhn, Amsel, O'Loughlin, 1988). But, these experiments involved causes for which the students were likely to have prior theories, and people interpret ambiguous data in ways that are consistent with their prior beliefs (Darley & Gross, 1983). Kuhn et. al. do not indicate whether students who did not notice the indeterminacy were answering in a manner consistent with their prior theory. Also, this set of studies

focuses on the coordination of theory and evidence, and one of criteria used for assessing student's answers was their ability to justify their responses. If this is an unconscious process, students might be sensitive to the differences between conditions, yet unable to justify their responses. In the present study, the task is made much simpler, by presenting the students with a novel effect and asking for their causal attribution without asking for a justification.

Data from two experiments are presented. Both experiments assess whether people differentiate between confounded and unconfounded causes. In one experiment, the subjects are undergraduates, while in the other experiment the subjects are pre-school age children. In both experiments, participants were presented with two possible causes for a novel event, and were asked to determine the cause of the novel event. In one condition the two possible causes were independently occurring, while in the other conditions, the two candidate causes always occur together. If people are sensitive to confounding they should be able to make a causal attribution in the first condition but not the second.

Methods

These experiments were designed to assess the extent to which people are sensitive to the independent occurrence of potential causes of an effect when making judgments of causality. The first experiment was conducted on adults. Even if adults are able to succeed in this task, it could easily be due to prior training. The second experiment was conducted on children. The similar materials were used for both experiments. Below we describe the methods for both experiments before reporting the results.

Experiment 1

Participants 10 undergraduates at the University of California, Los Angles enrolled in the Introduction to Psychology Course participated in the study. Students receive class credit for participating in the study and were recruited using an on-line bulletin board for this course.

Design This experiment had two conditions and utilized a within subjects design. In one condition, the two possible causes of an unusual event were perfectly correlated (confounded). In the other condition, the same two possible causes occurred independently of one another. Subjects were asked about the causality of the candidate causes in turn. The ordering of the stories, as well as the order in which the subject was asked about each candidate cause, was counterbalanced across subjects.

Materials Two passages of approximately the same length were constructed, (one story was 668 words and the other was 681 words). Both passages tell the story of bunny rabbits that went to two different parties.

In both stories, the parties occur at the same time and on the same day. The day of the party, half of the bunnies ate candy. At one of the parties the bunnies ate cake, while at the other party they did not. In the confounded condition all the bunnies who ate candy also ate cake, in the non-confounded condition half of the bunnies who ate candy also ate cake, and vice versa. All the bunnies at the cake party grew new pink wings; none of the bunnies at the "no cake" party did. Two avoid confusion between the two stories, in one story, the bunnies ate green grass candy and yellow cheesecake; in the other story the bunnies ate blue berry candy and orangey orange cake.

At the end of the story, participants were asked about the causality of each of the causal candidates in the story.

1) Does Yellow Cheese Cake/ Blue Berry Cake all by itself make bunnies grow new pink wings? Yes, No, Impossible to tell.
2) Does Green Grass Candy/ Orangey Orange Candy all by itself make bunnies grow new pink wings? Yes, No, Impossible to tell.

The text of the story was accompanied by illustrations, with an appropriately colored wedge in the bunnies' stomachs representing the cake, and a candy shaped object in the bunnies' stomach representing green grass candy.

Because we were attempting to revise the stimuli in order to make the directions clearer for the children, the stimuli underwent slight modification across the 10 subjects. The conditions remained the same, but there were slight changes in wording and pictorial presentation across groups.

The stories were shown as a power point presentation on a 15" computer screen.

Procedure Participants were randomly assigned to conditions that differed based on the ordering of the stories and assessment questions. Participants were then told that they were going to hear a story about bunny rabbits in two little bunny towns; that something interesting was going to happen to these bunny rabbits, and that they were going to try to figure out what happened.

Participants looked at the illustrations on the screen as the experimenter read the story aloud. At the end of the story, participants were asked about the causality of each of the causal candidates in the story.

The experimenter wrote down their answers on an answer sheet as they progressed through the story.

Experiment 2

Participants 16 pre-school children from the Bellagio day care center at the University of California, Los Angeles participated in the study. Nine male and seven female children between the ages of 4;5 and 5;7, with a mean age of 4;11 participated in the study. One student was excluded from the analysis for answering incorrectly on questions about the facts of the stories presented. The rest of the students answered all of the questions correctly (as explained later).

Design This experiment had two conditions and utilized a between subjects design. In one condition, the two possible causes of an unusual event were perfectly correlated (confounded). In the other condition, the two possible causes of an unusual event occurred independently of one another. The order in which children were asked about each candidate cause was counterbalanced across conditions.

Materials The stories presented to the children had the same content as the stories presented to the adults, with two differences. In both conditions, children saw green grass candy and yellow cheesecake. (This was possible because subjects only saw one story, which ruled out the possibility of carryover between stories.) The children's assessment procedure also differed from that of the adults.

Children were first asked for their spontaneous attribution. "Do you think that it is possible to figure out why the bunnies grew new pink wings?". If the child answered yes then the following questions were asked.

1) Why do you think these bunnies [pointing to those who went to the cake party] grew new pink wings?
2) Why do you think these bunnies [pointing to those who went to the no cake party] did not grow new pink wings? The ordering of these two questions was counterbalanced across conditions.

Because children sometimes do not answer in the free response, do not address both of the causal candidates, or do not address the causal candidates in their responses (i.e. "Bunnies grew wings because they wanted to"), additional probes were added, asking about each of the candidate causes separately. Children were told about statements that other children had made while reading this story. Children were asked to say whether they thought these statements were "definitely right, definitely wrong, or impossible to tell." The statements they were asked to judge were

1) GREEN GRASS candy all by itself makes bunnies grow pink wings.
2) YELLOW CHEESE CAKE all by itself makes bunnies grow pink wings.
3) YELLOW CHEESE CAKE and GREEN GRASS candy together make bunnies grow pink wings.

If the child had previously indicated that the yellow cheesecake was causal, they were not asked about the yellow cheesecake again, (and the same for the other candidates).

Procedure Children were randomly assigned to conditions. Children were video taped during the session. In order to accustom children to the camera, they were first introduced to the camera and allowed to see themselves on the LCD screen. Children were then told that they were going to hear a story about bunny rabbits in two little bunny towns; that something interesting was going to happen to these bunny rabbits, and that they were going to try to figure out what happened.

Participants looked at the illustrations on the screen as the experimenter read the story aloud. At the end of the story, children were asked 4 questions to assess whether they understood and remembered the content of the story. The experimenter pointed to a picture of the bunnies with the candy in their tummies and asked "What did these bunnies eat?", the correct answer being candy (or cake and candy in the confounded case). The experimenter then pointed to the bunnies without candy in their tummies and asked "Did these bunnies eat candy?", the correct answer being no. The experimenter then pointed to a picture of the bunnies at the cake party and asked "What did these bunnies eat at the party?" (this question was omitted in the confounded condition if children answered cake and candy to the first question above), the correct answer being cake. The experimenter then pointed to a picture of the bunnies at the no-cake party and asked "Did these bunnies eat cake?", the correct answer being no. Children who did not correctly answer all questions were excluded from the study.

Results

Experiment 1

Adults subjects were sensitive to confounding when they make causal judgments. In the confounded condition, when asked whether cake caused new pink wings, all 10 subjects said it was impossible to tell. When asked whether candy caused new pink wings, all 10 subjects said it was impossible to tell. In the unconfounded condition, when asked whether cake caused new pink wings, 8 subjects said cake did cause pink wings (the correct answer); 1 subject said it was impossible to tell; and 1 subject said cake did not cause pink wings. When asked whether candy caused pink wings, 6 subjects said candy did not cause pink wings (the correct answer); 4 subjects said it was impossible to tell.

Using McNemar's test for 2-related samples of categorical data, we see that the pattern of responses differed across conditions for both of the causal candidates. Subjects were more likely to say the cake was causal in the unconfounded condition than in the confounded condition, and more likely to say it was impossible to assess causality in the confounded condition than in the confounded condition ($p < 0.05$, exact statistic, binomial distribution used). Subjects were more likely to say that the candy was not causal in the unconfounded condition than in the confounded condition, and were more likely to say it was impossible to tell in the confounded condition than in the unconfounded condition ($p < 0.05$, exact statistic, binomial distribution used).

Using a χ^2 for each set of data, we see that subjects are picking the response that corresponds with the focal set theory reliably better than chance in most cases. Each of following χ^2 analyses uses three cells, corresponding to the three possible responses for the task (yes, no , impossible to time). In the confounded condition, subjects all said that it was impossible to tell if the cake was causal ($\chi^2 = 20$,df=2, $p<0.05$) and it was impossible to tell if the candy was causal ($\chi^2 = 20$,df=2, $p<0.05$). In the unconfounded condition, subjects were more likely to say that the cake was causal

than any other answer choice ($\chi^2 = 9.8$,df=2, $p<0.05$). None of the subjects said the candy was causal ($X^2 = 10$, df=1, $p<0.05$) and were evenly split between saying the candy was not causal or that there was not enough information to assess the relationship.

Experiment 2

Children were also sensitive to confounding when they make causal judgments, but this data has more variability associated with it than the adult version of the experiment.

Children were categorized into one of five causal attribution categories: the cake is causal, the candy is causal, both causal (jointly or independently), it is impossible to tell, and other causal attribution. If children made a spontaneous causal attribution, this was taken as the value for the measure. Otherwise, the value for this measure was taken from the child answers to questions about each individual candidate.

Using a χ^2 analysis for this data, we see that the pattern of responses differed across conditions, ($\chi^2 = 10.18$,df=4, $p<0.05$). In the confounded condition, 3 children said both were causal and 4 children said it was impossible to tell. No other responses were given by children in the confounded condition. In the unconfounded condition, 4 children said cake was causal, 1 child said candy was causal, 2 children said they both were causal, and one child gave an alternate attribution. No children said that it was impossible to establish causality.

Conclusions

Both children and adults make a distinction between confounded and unconfounded candidate causes when making attributions of causality. All adults said that it was impossible to tell whether the cake or the candy alone caused the wings in the confounded conditions. Because the conditions used in this experiment are consistent with examples used in scientific methodology classes, it is possible that the adults have had prior experience in science classes that train them to be able to do this. The same is not true of the children, however.

The child data suggests that children are able to state that is impossible to attribute causality when two candidate causes perfectly covary at a much earlier age than documented by previous studies. Children as young as 4 years old, less than a third as old as previously believed, made such attributions. They were also able to use frequency data to make a causal attribution.

The difficulty in using data to prove that a theory is correct or incorrect may not be due to student's inattention to confounded causes. Instead, it may be that that having to justify their responses, or an interaction with prior knowledge is the problem. When introducing children to the abstract idea of confounding, it may be useful to build of their intuitions of causal attribution by first having them make judgments of potential causes of novel effects. Their answers to these problems could then be used as a basis for discussion about the abstract concept of confounding.

Acknowledgments

The preparation of this paper was supported by an NSF fellowship to the first author and an NIH Grant MH64810 to the second author.

References

Carpenter, T.P., Fennema, E., Penelope, P. L, Chiang, C., & Loef, M. (1989) Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, 26, 499-531.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365-382.

Darley, J.M. & P,H.Gross (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20-33.

Gelman, S. A. (1996) Concepts and Theories. In T.K.F Au and R. Gelman (Eds.) *Perceptual Development* (pp117-150). Academic Press., San Diego, CA.

Gopnik, A., Sobel, D.M., Schulz, L.E. (2001) Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37 (5), 620-629.

Hunting, R.P., Davis, G. & Pearn, C.A. (1996) Engaging whole-number knowledge for rational-number learning using a computer-based tool. *Jounal for Research in Mathematics Education,* 27 (3), 354-379.

Jenkins, H. M, & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, 79 (1, Whole No. 594): 17.

Kuhn, D., Amsel, E, & M. O'Loughlin (1988) *The development of scientific thinking skills*. Academic Press, Inc: New York, NY.

Schulz, L.E. (2003, April) The play's the thing: Intervention and Causal inference. In L.E. Schulz (Chair) *Understanding children's causal knowledge: Exploring the origins of causal inference.* Symposium conducted at the meeting of the Society for Research in Cognitive Development, Tampa, FL.

Siegler, R.S. (1998) C*hildren's Thinking* (3rd ed.). Prentice Hall, Upper Saddle River, NJ.

Teaching Structural Knowledge in the Control of Dynamic Systems: Direction of Causality makes a Difference

Wolfgang Schoppek (wolfgang.schoppek@uni-bayreuth.de)
Department of Psychology, University of Bayreuth
D-95440 Bayreuth, Germany

Abstract

Recent publications about humans controlling dynamic systems have emphasized the role of specific rules or exemplar knowledge. Although it has been shown that small systems can be controlled with these types of knowledge, there is evidence that general knowledge about the structure of a system plays an important role, too, particularly when dealing with systems of higher complexity. However, teaching structural knowledge has often failed the expected positive effect. The present work investigates details of acquisition and use of structural knowledge. It is hypothesized that guiding subjects to focus on dependencies rather than effects supports them in applying structural knowledge, especially when the application is practiced in a strategy training. An experiment with N=95 subjects supported the hypothesis of the usefulness of the dependency perspective, but revealed an adverse effect of the strategy training. Differences between subgroups studying different majors have been found that give rise to questions about the relation between prior knowledge and instruction. The results have interesting implications for models of how structural knowledge is represented as well as for methods of teaching system control efficiently.

Humans have to deal with dynamic systems throughout their lives. Especially in industrial environments, people are confronted with new systems such as production lines frequently. Therefore it is worthwhile to study how humans learn to control dynamic systems, and how instruction can support the learning process.

In cognitive psychology, a common paradigm for studying the control of dynamic systems can be characterized by the following features: The systems simulate some fictitious device or environment that most people have no specific experience with (e.g. a tank with sea animals in a biology lab, used by Vollmeyer, Burns, & Holyoak, 1996). This is to ensure an equally low level of prior knowledge. Discrete linear additive equations are used for simulation, one equation per output variable. There is the opportunity to assign values to the input variables in each simulation step, which is referred to as "trial". A number of trials, e.g. six simulated hours, make up a "round". The objective for participants is to attain a specific goal state either at the end of a round, or as soon as possible and to maintain the state. A prominent, yet simple example is the "Sugar Factory" (Berry & Broadbent, 1984) that has been used to investigate questions about implicit vs. explicit knowledge and about rule vs. exemplar learning (e.g. Dienes & Fahey, 1995; Fum & Stocco, 2003; Lebiere, Wallach, & Taatgen, 1998)

Research with this paradigm has shown that subjects largely prefer acquiring and using exemplar knowledge rather than structural knowledge, i.e. subjects memorize specific actions taken in specific situations together with their outcomes. This strategy can be successful under certain conditions: First, when the system has a small problem space like, for example, the Sugar Factory (144 states); second, when the same goal state has to be attained repeatedly (Vollmeyer et al., 1996), which means that only a small fraction of a possibly large problem space is relevant. Simulation studies with the Sugar Factory have shown that it can be successfully controlled by using either declarative representations of specific actions (Lebiere et al., 1998), or learned production rules that also represent specific interventions (Fum & Stocco, 2003). In conditions, however, where subjects have to deal with huge problem spaces (e.g. because the system is more complex and subjects have to attain a number of different goal states), the exemplar strategy is no longer useful[1]. Instead, it is more reasonable to use general knowledge about the causal structure of the system to navigate through the problem space. I will refer to this type of knowledge as "structural knowledge".

In principle, complete structural knowledge is sufficient to control a system even without specific experience. Although correlations between structural knowledge and performance have been reported (Funke, 1993), experiments where structural knowledge was taught, usually failed to demonstrate its superiority (Putz-Osterloh, 1993; Schoppek, 2002). One reason for this is that deriving

[1] The inclination to use exemplar knowledge even when it is inappropriate may explain why subjects generally perform at very low levels when they are asked to control complex dynamic systems that are new to them.

specific actions from structural knowledge is a skill that has to be practiced in addition to learning the structure. This view is corroborated by results from studies where the application of structural knowledge has been practiced extensively (Preussler, 1998). A second reason for the difficulties of applying structural knowledge is that knowledge about causal relations is acquired under a different perspective than it is applied when controlling a system. This issue is elaborated in the following paragraphs.

Verbal protocols of successful system controllers and simulation studies (Schoppek, 2002) have helped identify efficient strategies for acquisition and application of structural knowledge. A good strategy for exploring the causal structure of a system is to vary input variables one at a time to identify the immediate effects of the input variables and the momentum of the system, which is produced by effects of output variables onto each other. For example, a subject could put some lime into the animal tank to observe the effect onto the oxygen content of the water, then set lime input back to zero and observe how the oxygen content changes on its own.

A common application strategy starts with (1) predicting the next state of the system under the assumption of no interventions, continues with (2) calculating the differences between the predicted and the desired state, (3) selecting a free input variable, (4) calculating the input value, and ends with (5) applying the intervention. In the course of this strategy, for each output variable all their dependencies are considered in turn. This consideration of dependencies is a marked difference compared with the focus on effects that is prevalent during acquisition of structural knowledge.

Thus we can distinguish two perspectives on causal relations: One looking for effects of a given cause, the other looking for possible causes of a given effect. The first perspective is prevalent during exploration of a new system, the second is more adaptive during system control. In the following, I will use the word “effects” to characterize constructs related to the first perspective, and “dependencies” to characterize the second perspective.

The distinction of perspectives on causal relations has a number of implications. The first has to do with the question what given information cues the retrieval of what other information. During exploration, when input variables are manipulated and effects are observed, associations from cause to effect are learned, resulting in a structure where representations of input manipulations act as cues for representations of changes in output variables. When the task is to control a system and the dependencies of output variables are considered, output variables should be learned as cues for input variables.

A second implication concerns the mechanism of chunking, which plays an important role in successful problem solving (Newell, 1990, Gobet & Simon, 1996). The effect perspective suggests chunking together single effects of a variable (which can be an input or an output variable), whereas the dependencies perspective suggests chunking together all dependencies of an output variable. Again, the second possibility seems to be more adaptive in system control, because having all dependencies in one chunk relieves the problem solver from extensive memory search, a process that consumes much time, poses high demands on working memory, and is thus error prone.

A second issue in the context of helping humans to use structural knowledge has to do with strategy instruction. Undoubtedly, extensive practice under supervision of experienced operators is effective, but also very costly. Thus it is important to find ways of leveraging structural knowledge efficiently. The way followed here was to base a training program on a strategy that has proven successful in a computer simulated cognitive model of controlling a system similar to the present one (Schoppek, 2002).

To summarize, the aim of the present work is to investigate ways of teaching structural knowledge about dynamic systems, either indirectly by manipulating the perspective on causal relations, or directly by practicing the application of structural knowledge. Specifically, I tested the hypothesis that guiding subjects to focus on dependencies rather than effects enhances performance. By measuring access to causal knowledge with a speeded judgment task I investigated if the different perspectives are also reflected in the representation of structural knowledge. The results may show new ways of teaching structural knowledge and extend our understanding of the use of this type of knowledge.

Experiment

The system I used in this experiment is a simulation of the influences of three fictitious medicines onto the levels of three fictitious peptides in the blood. The medicines are called MedA, MedB, and MedC; the peptides are called Muron, Fontin, and Sugon. The effects of the substances onto each other are simulated with the following discrete linear equations:

(1) $\text{Muron}_t = 0.1\ \text{Muron}_{t-1} + 2\ \text{MedA}_t$

(2) $\text{Fontin}_t = \text{Fontin}_{t-1} + 0.5\ \text{Muron}_{t-1} - 0.2\ \text{Sugon}_{t-1} + \text{MedB}_t$

(3) $\text{Sugon}_t = 0.9\ \text{Sugon}_{t-1} + \text{MedC}_t$

In a neutral state with Muron = Sugon = 0 and Fontin = x, the system is stable. Once some of the medicines are administered, the system gains momentum. Note that the

amount of Fontin in the blood can only be reduced through Sugon, which depends on MedC. Since Sugon decomposes slowly, large time delays of changes in medication have to be dealt with. Subjects interacted with the system through an interface consisting of two tables showing the states of the variables in all trials, and input boxes where they could enter values for the medicines. One round comprised six trials, introduced to the subjects as "simulated hours".

Structural knowledge was tested with a speeded causal relation judgment task. All names of input and output variables were shown on a screen in a spatial arrangement that matched that of the simulation interface. This was done to assure that variables could be identified by both, their names and their locations. Then the name of an output variable was highlighted on the right side of the screen, followed by the highlighting of another variable name on the left side with an ISI of 500 ms. The subject was asked to respond with pressing one of two keys as quickly and accurately as possible to indicate her judgment if there was a causal relation between the highlighted variables or not. All 18 possible input-output and output-output relations were shown in one test. Eight of these relations had to be answered with yes, 10 with no. The procedure was arranged such that knowledge of dependencies should result in faster judgments compared with pure knowledge of effects. This is expected because the variable that was highlighted first (the effect) is assumed to act as a prime for the variable highlighted second (the cause) only when causal relations have been memorized under the perspective of dependencies.

Subjects and Design

N=95 subjects, studying different majors at the University of Bayreuth, participated in the experiment. Subjects were paid 10 € for their participation.

The factor "type of knowledge" with the levels "knowledge of effects" (Eff) and "knowledge of dependencies" (Dep), and the factor "strategy training" with the levels "no training" and "training" were varied between subjects. A third, quasi-experimental factor "field of study" with the three levels arts/humanities, law/economy, and science was also analyzed. In principle, subjects were randomly assigned to one of the four conditions. A few exceptions from complete randomization were due to the objective to have approximately equal distributions of field of study in each condition.

Procedure

The experiment began with a general instruction about the system. All subjects went through a standardized exploration phase guided by the experimenter. The exploration was designed to demonstrate all causal relations between the variables of the system. Subjects were guided to analyze the observed effects and asked to enter them in cards provided by the experimenter. The procedure in this phase was different for the two knowledge conditions: In the Dep condition, the experimenter consistently asked for dependencies, and the cards were sorted by the "dependent" variables Muron, Fontin, and Sugon. In the Eff condition, the experimenter consistently asked for effects, and the cards were sorted by the "independent" variables MedA, MedB, and MedC. At the end of this phase, the experimenter examined the knowledge of the subject orally, again consistently asking either for dependencies or for effects. Subjects had to recall all possible relations with the respective numeric weights before moving on to the next phase (all subjects achieved that).

Subjects in the "no strategy training" condition could then explore the system for one round (six simulated hours) on their own. Subjects in the "strategy training" condition went through a number of exercises where they practiced a method of predicting future states of the system. As mentioned above, this was the first part of a strategy tested earlier in a cognitive model. Only a part of the complete strategy was selected to keep the training short. Nevertheless, all effects (condition Eff) or dependencies (condition Dep) were needed and rehearsed in these exercises.

Next, all subjects were given the control problems. All problems comprised six simulated hours and were given with the objective that the goal states had to be reached as soon as possible, and to be maintained. Table 1 shows the initial states and the goal states for the four control problems. Initially, all variables except Fontin were zero. In order for the subjects to familiarize themselves with the control task, they were given two rounds for Problem 1.

Table 1: The four control problems given to the subjects

Problem 1: Fontin = 50	→	Muron = 200 , Fontin = 1000
Problem 2: Fontin = 900	→	Muron = 100
Problem 3: Fontin = 2000	→	Fontin = 1000
Problem 4: Fontin = 50	→	Muron = 400 , Fontin = 900

Results

To measure control performance, the solution error was calculated by summing the natural logs of the absolute differences between the goal values and the actual values for each time step of a round (Müller, 1993). A perfect solution is indicated by a solution error of zero. Since the results of Problem 2 were close to ceiling, they were excluded from the analysis. I analyzed the mean solution error of the remaining problems as dependent variable in

an ANOVA with the factors "type of knowledge" (knowledge about effects, "Eff" vs. knowledge about dependencies, "Dep"), "strategy training" (with vs. without training), and the quasi-experimental factor "field of study" of the participant (arts/humanities, law/economy, science). The means aggregated across all fields of study are listed in Table 2.

The ANOVA yielded significant main effects of all three factors, "type of knowledge" ($F = 3.94$, $df = 1$, $MSE = 5.57$, $p = .05$), "strategy training" ($F = 5.97$, $df = 1$, $MSE = 8.45$, $p < .05$), and "field of study" ($F = 13.24$, $df = 2$, $MSE = 18.75$, $p < .01$). As expected, subjects who were guided to acquire knowledge of dependencies were more successful in controlling the system (mean solution error = 2.1, SD = 1.3) than subjects who were guided to acquire knowledge of effects (M = 2.6, SD = 1.5). Contrary to expectation, subjects who underwent the strategy training performed lower (M = 2.6, SD = 1.5) than those without strategy training (M = 2.0, SD = 1.3). Subjects studying arts or humanities performed worst (M = 3.2, SD = 1.4, n = 33), followed by subjects studying law or economy (M = 2.3, SD = 1.2, n = 30). Most successful in controlling the system were science students (M = 1.6, SD = 1.1, n = 32).

There is a significant interaction between "field of study" and "type of knowledge" ($F = 3.29$, $df = 2$, $MSE = 4.65$, $p < .05$). Detailed analyses revealed that a strong effect of "type of knowledge" was only present in the group of subjects who studied arts/humanities (see Figure 1). No other effects reached statistical significance (all $p > .05$).

Table 2: Solution error of system control in the various conditions of the experiment

		Eff	Dep	
Strategy training	yes	**2.9** (1.5) n = 24	**2.4** (1.5) n = 26	**2.6** (1.5) n = 50
	no	**2.4** (1.5) n = 22	**1.7** (0.9) n = 23	**2.0** (1.3) n = 45
		2.6 (1.5) n = 46	**2.1** (1.3) n = 49	**2.4** (1.4) n = 95

To test the expectation that knowledge of dependencies results in faster response times in the speeded structural knowledge test, I calculated an ANOVA with the same factors as described above and the mean response times for hits in the first test as dependent variable. (Three subjects with mean response times of greater than 3800 ms were excluded from the analysis. Raw values were ln-transformed for the ANOVA). The expected effect of

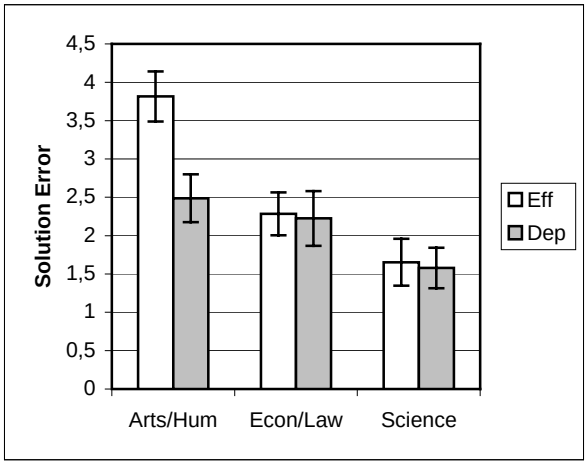


Figure 1: Means and standard errors of solution error of controlling the system (smaller values indicating better performance)

"type of knowledge" was confirmed by the analysis ($F = 7.83$, $df = 1$, $p < .01$), (1559 ms vs. 1237 ms, Dep faster). However unexpectedly, there was also a main effect of "strategy training" ($F = 11.24$, $df = 1$, $p < .01$), (1576 ms vs. 1236 ms, with training faster). No other effects were significant at the level of $\alpha = .05$. The results of the second structural knowledge test were analogous to the first test.

Similar analyses with the discrimination index (an index of how well subjects can discriminate between relations and no relations, cf. Snodgrass & Corvin, 1988) as dependent variable yielded no significant effects. Discrimination indices were relatively high in all conditions (di = 0.89).

Discussion

The experiment has confirmed the hypothesis that guiding subjects to focus on dependencies of output variables rather than on effects of input variables can enhance performance in controlling a complex dynamic system. Although there is an effect in the complete sample, the major contribution came from the subjects studying arts/humanities. Presumably, this group has the least experience with abstract representations of dynamic systems and thus learned something new when focusing on dependencies instead of effects. If the other groups did not benefit from the manipulation because they take the dependencies perspective on their own, or because of some other strategy cannot be told with the present data.

The results of the speeded causal judgment task indicate that focusing on dependencies vs. effects affects the

mental representation of causal relations. The task was arranged to enable priming from output to input variables, but not the other way round. Subjects in the Dep condition were significantly faster in judging the relations, supporting the assumption that they have established stronger associations between output to input variables than subjects in the Eff condition.

The two findings are raising the question about their relation. Are these stronger associations a cause for better performance or are they just a side effect of the experimental manipulation? If the relation was causal, there should be a substantial (negative) correlation between response time in the causal judgment task and solution error in the control problems. The respective correlation is r=.05 in the whole sample. Hence, the faster reaction times in the Dep condition are probably a side effect of the manipulation. This, in turn, supports the hypothesis that the positive effect of knowledge of dependencies on performance is based on the chunking aspect, i.e. the integration of single effect representations according to output variables. It is possible that especially science students have built such chunks on their own, even in the Eff condition. (Note that subjects in the Eff condition were not prevented from gaining knowledge of dependencies). Figure 2 shows a sketch of the hypothetical structure of a dependency chunk "Dep01" (the causal weights are omitted for clarity). The shaded substructure "Eff01" is a chunk that represents the single causal relation between MedC and Sripon. The structure, whose construction in a learning process appears straightforward, mirrors the equations defining the behavior of the system remarkably. The solid lines indicate slot-value relations. Dotted lines indicate the associations between the name of the dependent variable and names of influencing variables, which may have been learned under the Dep condition. These associations can explain the effects in the speeded judgment task, but are not necessary for the usefulness of dependency chunks in control tasks. This interpretation is in line with the assumption of Boucher & Dienes (2003) that there are two ways of learning associations, one resulting in activating relations, the other resulting in chunks that combine the associated information. Baker, Murphy and Vallée-Tourangeau (1996) suppose that these two ways may be attributed to different modules of the mind. Research on causal reasoning has discovered many other cases where concept-driven symbolic processing must be assumed in addition to pure associative learning to explain the phenomena (Waldmann, 1996).

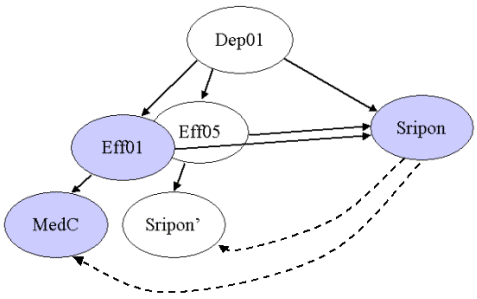


Figure 2: Hypothetical structure of a dependency chunk; solid lines indicate slot-value relations, dotted lines indicate associations.

Unexpectedly, the strategy training had an effect on answering speed in the causal judgment task (with training faster). According to the above interpretation subjects must have rehearsed relations between each output variable y and the variables affecting y during the training. In the Dep condition, this is obvious. Since in the Eff condition subjects were asked for all variables that had an effect on the output variable in question, they had to search memory for names of input variables while the name of the output variable was present in working memory. Thus, subjects have learned associations from output to input in that condition, too.

The adverse effect of the strategy training was also unexpected. The training had been inspired by results from cognitive tutoring that subskills can effectively be trained based on single production rules (Anderson, 1993), and thus, practicing only the most difficult part of a larger strategy appeared reasonable. However, the success of this kind of training depends on the compatibility of the practiced subskills with the subjects' own strategies. This condition seemed to be hurt in the present case. Subjects might have applied the practiced method of predicting the next state, and after successful completion were unclear about what to do next and how to use the result. An alternative explanation is that the practiced strategy has interfered with the subjects' own strategies, resulting in mixtures of incompatible strategy fragments. (see e.g. Vosniadou, 1997 for the difficulties of integrating new knowledge with prior knowledge).

In future efforts to train the application of structural knowledge it should be assured that subjects have at least an idea of the whole strategy. This could be achieved by introducing abstract labels for all subgoals and practicing the whole strategy at least once before possibly focusing on the most difficult part of it (Catrambone, 1998).

In general, the results of the experiment show that variations of structural knowledge do affect performance in the control of dynamic systems. This extends the view that mainly exemplar knowledge or very specific rules are used for controlling systems (Dienes & Fahey, 1995; Fum & Stocco, 2003; Lebiere et al., 1998). It is important to note that not knowledge about single causal relations as measured by the discrimination index of the causal judgment task makes the difference (there were no effects of the experimental factors on di), but rather the way of using it, obviously depending on prior knowledge, and the way of chunking it into larger units.

Acknowledgments

The research reported here was supported by the University of Bayreuth. I would like to thank the anonymous reviewers for valuable hints and Lucie Necasova, Tereza Kvetnova, and Nha-Yong Au for collecting the data.

References

Anderson J.R. (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Baker, A.G., Murphy, A., & Vallée-Tourangeau, F. (1996). Associative and normative models of causal induction: Reacting to versus understanding cause. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The Psychology of Learning and Motivation 34* (pp. 3 - 46). San Diego: Academic Press.

Berry D.C., & Broadbent D.E. (1984). On the relationship between task performance and associated verbalisable knowledge. *The Quarterly Journal of Experimental Psychology, 36A,* 209 - 231.

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science, 27,* 807-842.

Catrambone R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General, 127,* 355 - 376.

Dienes Z., & Fahey R. (1995). Role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 848 - 862.

Fum, D., & Stocco, A. (2003). Outcome evaluation and procedural knowledge in implicit learning. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society.*

Funke, J. (1993). Microworlds based on linear equation systems: A new approach to complex problem solving and experimental results. In G. Strube & K.-F. Wender (Eds.), *The cognitive psychology of knowledge*. (pp. 313-330). Amsterdam: Elsevier (North-Holland).

Gobet F., & Simon H.A. (1996). Recall of Random and Distorted Chess Positions: Implications for the Theory of Expertise. *Memory & Cognition*, 24, 493-503.

Lebiere C., Wallach D., & Taatgen N. (1998). Implicit and explicit learning in ACT-R. In F.E. Ritter, & R. M. Young (Eds.), *Proceedings of the Second European Conference on Cognitive Modelling* (ECCM-98) (pp. 183 - 189). Nottingham, UK: Nottingham University Press.

Müller, H. (1993). *Complex problem solving: Knowledge and reliability*. Bonn: Holos.

Newell A. (1990). *Unified Theories of Cognition: The 1987 William James Lectures.* Cambridge, MA: Harvard University Press.

Preußler W. (1998). Strukturwissen als Voraussetzung für die Steuerung komplexer dynamischer Systeme. [Structural knowledge as precondition for the control of complex systems] *Zeitschrift für Experimentelle Psychologie, 45,* 218 - 240.

Putz-Osterloh W. (1993). Strategies for knowledge acquisition and transfer of knowledge in dynamic tasks. In G. Strube, & K. F. Wender (Eds.), *The cognitive psychology of knowledge* (pp. 331 - 350). Amsterdam: Elsevier.

Taatgen N.A. (2001). A model of individual differences in learning air traffic control. In E. M. Altmann, A. Cleeremans, C. D. Schunn, & W. D. Gray (Eds.), Fourth international conference on cognitive modeling (pp. 211 - 216). Mahwah, NJ: Lawrence Erlbaum Associates.

Schoppek W. (2001). The influence of causal interpretation on memory for system states. In J. D. Moore, & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 904 - 909). Mahwah, NJ: Erlbaum.

Schoppek W. (2002). Examples, rules, and strategies in the control of dynamic systems. *Cognitive Science Quarterly, 2,* 63 - 92.

Snodgrass J.G., & Corvin J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34 - 50.

Vollmeyer R., Burns B.D., & Holyoak K.J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20,* 75 - 100.

Vosniadou, S. (1997) The development of the understanding of abstract ideas, in K. Harnqvist & A. Burgen (Eds.) *Growing up with Science*, Jessica Kingsley Publishers

Waldmann M.R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The Psychology of Learning and Motivation 34* (pp. 47 - 88). San Diego: Academic Press.

Deductive rationality in human reasoning: Speed, validity and the assumption of truth in conditional reasoning

Walter J. Schroyens (Walter.Schroyens@psy.kuleuven.ac.be)
Laboratory of Experimental Psychology, University of Leuven
Tiensestraat 102, Leuven, B-3000, Belgium

Abstract

We proffer the thesis that, in the process of defeating an inference on the basis of a factual truth that falsifies it, people move from a hypothetical truth-value to a factual truth-value of the conclusion. We will present evidence that shows (a) that some people spontaneously make a truth assumption and constrain their inferences to logically valid inferences, (b) that people tend to abandon the truth-assumption when they have factual evidence to the contrary, (c) that people, however, can and do in fact reason logically when they are informed about the rules of the language game (i.e., the truth-assumption) and (d) that adhering to the truth-assumption in the face of conflicting evidence to the contrary requires an investment of time and effort. The findings are discussed in relation to contemporary theories of human reasoning.

General Introduction

We all reason: We draw inferences from the multiple sources of information we are confronted with and make decisions based on them. This allows us to move around in a changing world where the capability to comprehend the contingent nature of our environment determines for a large part our successes as an individual, as well as a species. The study of human reasoning is therefore important to advance our understanding of the general mechanisms of thought.

The turn of the century has provided the stage of a paradigm shift in human reasoning research. The nineties provided the scene for polemical debates as regards basic human reasoning competence. This basic reasoning competence (i.e., the basic machinery that allows us to draw inferences) was mostly studied by means of abstract knowledge-lean inference problem. By using arbitrary relations (e.g., 'if the letter is an A, then the number is a 2") no content-specific background knowledge would be triggered to influence the reasoning process towards accepting or rejecting the conclusion. Abstraction was made of the specific content of that about which people were reasoning. It is within this research milieu that theories became specified as regards human deduction. In the study of human deduction one studies necessary inferences derived from certain premises. One asks people to draw logically valid inferences, and these are defined as inferences that must be necessarily true if the premises are true. Presently there is an increasingly prominent body of evidence that shows the pervasive influence of content and belief (Cummins et al., 1991). Our beliefs are uncertain (i.e., they are true to a certain degree: e.g., even Newton's mechanics are not universally applicable). This observation induced a shift towards the study of the subjective probabilistic properties of that about which we are reasoning as well as commonsense reasoning or reasoning under uncertainty.

The present research is situated within this timely clash between experimental paradigms and associated theoretical approaches. Theorists sometimes like to boost the polemics between dichotomized opposites (it does make for simpler, and hence more easily publishable reading). For instance, it is claimed that theories that have focused on reasoning under certainty (i.e., deductive reasoning) are incapable of being extended to reasoning under uncertainty (i.e., probabilistic reasoning). The 'core argument' (Oaksford & Chater, 1998) is that common-sense reasoning is non-monotonic, whereas logic systems are monotonic: valid inferences cannot be invalidated; they remain valid. The validity of everyday inferences however would be revisable. For instance, when being given the argument:

'If it is a bird, then it flies;
Tweety is a bird who, thus, can fly"
almost everybody will accept it. At the same time, when subsequently being told that Tweety is an ostrich, almost everybody will reject the original inference and will state that Tweety cannot fly.

The rationality debate in the cognitive science of human reasoning is partly muddled by a failure to distinguish the defeasibility of a conclusion from the non-monotonicity of an inference. For instance, Oaksford and Chater's (1998) core argument is subverted when taking count of the distinction between truth and validity. Monotonicity concerns the validity of inferences; defeasibility concerns the truth of conclusions and this "distinction between *validity* and *truth* … is basic to deductive logic [and] many people find the distinction difficult to grasp" (Glass & Holyoak, 1986, p. 338). The abovementioned definition of logical validity use the notion of truth but the truth of a valid conclusion is always hypothetical (*if* the premises are true, then the conclusion must also be true). The truth-value of a defeated inference however is not hypothetical. It is factual: it hinges on a factual truth (i.e., our belief, at a particular moment in time and space that something is true in the 'real' world).

The present study intends to show the importance of the truth-assumption and by consequence the hypothetical nature of the truth of logically valid inferences. We proffer the thesis that in the process of defeating an inference people move from a hypothetical to a factual truth-value of this conclusion. I present evidence showing (a) that at least some people make the truth-assumption and spontaneously constrain their inferences to logically valid inferences, (b) that people abandon a truth-assumption when they have factual evidence to the contrary, (c) that people, however, can and do in fact reason logically when they are informed about the rules of the language game (i.e., the truth-assumption) and (d) that adhering to the truth-assumption in

the face of conflicting evidence to the contrary requires an investment of time and effort. In the general discussion we will then return to the theoretical and conceptual issues that are touched by the evidence for people's propensity to exhibit deductive rationality in reasoning hypothetically on the basis of a truth-assumption.

Table 1
Formal representation and standard nomenclature of the four basic conditional inference problems and their default conclusions.

	Logically valid		Logically Invalid	
	Affirmation	Denial	Affirmation	Denial
Premises	Modus Ponens:MP	Modus Tollens:MT	Affirm consequent: AC	Denial Antecedent: DA
Major	[If p then q]	[if p then q]	[if p then q]	[if p then q]
Minor	[p]	[not-q]	[q]	[not-p]
Conclusion	[q]	[not-p]	[p]	[not-q]

Experiment

To investigate the truth-assumption in representing the information with which we are confronted and about which we reason, and its import apropos validity and deductive rationality in human reasoning we will make use of well-known content effects in conditional reasoning. In the following I first introduce these effects. Next, I present them within a dual-processing framework. This yields some additional predictions concerning the functional and temporal relations of two conceptually distinct types of reasoning (and the corresponding distinction between hypothetical versus factual truth).

Content Effects. Table 1 presents the most commonly studied conditional inference problems. These problems are formed by an affirmation or denial of the antecedent [p] or consequent [q] of a conditional of the form [if p then q]. The content of the conditional utterance can be almost anything, e.g.:

(1) If you turn the key, then the car will start.
(2) If you heat water to 100°C, then it will boil.
(3) If you push the brake, then the car will stop.
(4) If you jump into the swimming pool, then you'll get wet.

The content effects that are observed with such realistic conditional-inference problems show that the reasoning process is strongly affected by the factual truth of the premises and/or conclusion (Politzer & Bourmaud, 2002).

At a general level the content effects are summarized as an effect of the number of factual counter-examples. For instance, the conclusions for AC and DA are falsified by situations that reflect the possibility that the antecedent is false [not-p] while the consequent is nonetheless observed [p]. When the conditional captures a causal statement, such [not-p and q]-cases reflect *alternative causes*. For instance, when we ask people to generate alternative causes for conditionals (1) and (2), they generally come up with relatively few as compared to the number of alternative causes they can generate for conditionals (3) and (4). The conclusions of MP and MT are countered by situations that represent the contingency where [p] is satisfied whereas [q] is not satisfied. When the conditional enunciates a causal statement, such [p and not-q]-cases reflect *disabling conditions*. For instance, when we ask people to generate alternative causes for conditionals (1) and (3), they generally have little difficulty coming up with a relatively high number of factors that might prevent the effect from occurring. For conditionals (2) and (4) people can only come up with few disabling conditions. The most robust finding in reasoning with conditionals like (1), (2), (3) and (4) above, is that people are less likely to accept MP/MT when there are many (vs. few) disablers and are less likely to accept AC/DA when there are many (vs. few) alternatives.

We proffer the thesis that belief effects in conditional reasoning and the presumed problematical nature of these effects for systems of deduction are due to a failure to play the language game of deduction. When one does not ask people to assume that the premises are true, people are not asked to reason deductively. Studies that investigate content-effects in conditional reasoning often do not even mention the truth-assumption. This implies that no implications can be drawn as regards people's deductive rationality (i.e., their propensity or capability to infer logically valid inferences). To demonstrate the importance of the truth-assumption in deduction reasoning, we decided to stress the truth-assumption and its implication that any inference made under this assumption is hypothetically true. The experiment was set up so we could compare performance on problems that did not stress the truth-assumption with problems that did stress the truth-assumption.

Expectations are relatively straightforward. When people are reasoning on the basis of the truth-assumption they will exhibit more deductive rationality as compared to situations where they reason in an unconstraint context. Deductive rationality in the present study is measured by the proportion of inferences that are valid relative to the norm of classic logic. That is, when people reason in a stressed truth-assumption context, they will endorse more logically valid MP and MT inferences. The logically invalid AC/DA arguments would not be affected by an increased impetus the hypothetical nature of inferences made under the truth assumption. Indeed, the counterexamples to MP/MT would be excluded or impossible, if the conditional were true. However, the counter-examples to AC/DA (i.e., alternative causes) are consistent with a conditional utterance of the form [if p then q]. Indeed, the utterance "If you jump into

the swimming pool, then you'll get wet" does not say '*if and only if* you jump into the swimming pool, then you'll get wet". In sum, there should be an interaction between the logical-validity of the inference and the impetus that is placed on the truth-assumption.

Dual Processing. We noted that the present research is situated within the timely clash between experimental paradigms and associated theoretical approaches. Being faced with the task of reconciling the 'old' (deductive certainty) and the new (probabilistic uncertainty), there is an increasing popularity of so-called dual processing frameworks. There presently seems to be a growing consensus that a distinction can be made between two types of rationality, or systems of reasoning (see, e.g., Evans & Over, 1996; Johnson-Laird, 1983; Stanovich & West, 2000). Dual-process theories of reasoning draw on the distinction between, on the one hand, highly contextualized associative, heuristic, tacit, intuitive or implicit processes that are holistic, automatic, experiential in nature, and relatively undemanding of cognitive capacity and, on the other hand, de-contextualised, rule-based, analytic, explicit processes that are relatively slow, and demanding of cognitive capacity.

There is a commonality in almost all dual-processing theories. About the functional relation between the two reasoning systems it has been argued that there is a primacy of System 1 processes (Stanovich & West, 2000). Evans and Over (1996) similarly discussed the override function of System 2 (Explicit, Rationality-2 in their terminology). This functional relation parallels the distinction and relation between generate and test procedures (Chater & Oaksford, 1999) or, analogously, the conclusion formulation and validation stages proffered in the highly influential mental models approach to reasoning (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991, 2002). We can associate factual/probabilistic reasoning and hypothetical/deductive reasoning with respectively System-1 and System-2 thinking. The override function of System-2 as regards the output of System-1 consequently allows us to specify some additional expectations concerning the potential effect of stressing the truth-assumption.

In the dual-processing framework it is assumed that System-2 processes are secondary to the workings of System-1 processes. This implies that if we can inhibit system-2 thinking, the effects of its functionality will be reduced. That is, we would expect the effect of stressing the truth-assumption to be reduced under conditions that are not conducive to system-2 thinking. We can expect, the other way round, that when we can instigate system-2 thinking, the effect of its potential override function would be increased. This means that the effect of stressing the truth-assumption would be strongest under conditions that allow people to engage in the resource-dependent and time-consuming system-2 type of thinking.

We asked one group to reason as quickly as possible, thereby reducing the potential import of the system-2 thinking (see Schroyens, Schaeken, & Handley, 2003) and the expected effect of stressing the truth-assumption. A second group was asked to think carefully. Given that people are less likely to engage in system-2 thinking under speeded inference conditions (as compared to the standard-inference conditions), we can expect that the inhibitory effect of stressing the truth-assumption will be annulled. That is, the other way around, only people who have the time and motivation to engage in system-2 type thinking will exhibit the effect of stressing the truth-assumption.

Method

Design. Participants served as their own control as regards inference type (logically valid: MP/MT vs. logically invalid: AC/DA), the number of alternative causes (few vs. many), the number of disabling conditions (few vs. many), and the impetus that was placed on the truth-assumption (no vs. strong). A between-groups factor was formed by the impetus that was placed on speed vs. accuracy.

Materials. We collected 16 conditionals utterances for which people in a pre-test were able to generate few or many alternatives and few or many disablers (see, e.g., items 1-4 presented above). The set contained four items for each of these four types of conditionals with few/many alternatives/disablers. Each conditional served as the major premise for each of the four types of inference problems (MP/MT/AC/DA, see Table 1).

The inference problems were cast into two booklets. A first booklet contained the 32 items that did not mention the truth-assumption and a second booklet with 32 other items that stressed the truth-assumption. (The specific item content was counter-balanced across the two truth-assumption conditions). Each counterbalancing set contained two items of an MP/AC/DA/MT argument about a conditional with few/many alternatives/disablers (2x4x2x2=32). The non-stressed condition presented the problems as follows.

If you turn the key, then the car will start.
You turn the key.
It follows:
The car will start.

Participants marked their evaluation of this conclusion on a 7-point scale ranging from (1) very uncertain that the conclusion follows to (7) very certain that the conclusion follows. In the stressed truth-assumption condition the problem was presented in the following format:

If you assume that it would always be true that:
If you turn the key, then the car will start.
And you know for sure:
You turn the key.
Then it would follow:
The car will start.

Participants marked their evaluation of this conclusion on a 7-point scale ranging from (1) very uncertain that the conclusion follows if one assumes that the rule is true to (7) very certain that the conclusion follows if one assumes that the rule is true.

The instructions to the speeded inference conditions mentioned that they were to evaluate the problems fast and should not stay too long with any particular problem. After the 3rd and the 6th sheet of paper, with four problems per page, an extra page was inserted which reminded them that

they were to make their judgment 'as quickly as possible'. In the accuracy conditions this reminder said that they were to 'think carefully' and that their evaluations of the conclusions should be 'as accurate as possible'.

Procedure. Participants received both problem booklets at the beginning of the session (the standard problems first; the truth-assumption problems second). About half the students in each of two 11th and two 12th grade classes received the problems with speeded-inference instructions, whereas the other half received the accuracy instructions. The students in accuracy groups were told that the one who generated the most correct conclusions of a predetermined subset would receive 10 Euro. To the speeded groups it was said that the person who solved the problems fastest (at a minimum accuracy level) would also receive 10 Euro.

Participants. Participants were 72 11th and 12th grade student from a Belgian, Flemish high school. Thirty-four students received the speeded-inference instructions; the remaining 38 pupils ended up in the accuracy conditions.

Results

Certainty ratings (1-7) were transformed to the [0,1] probability interval and submitted to analyses of variance. Figure 1 presents the effect of alternatives on the logically invalid inferences (AC/DA), and the effect of disablers on the logically valid inferences (MP/MT) in the standard conditions that do not mention the truth assumption. These standard problems replicate the standard findings. First, the number of disablers affected the certainty ratings of the logically valid inferences: Participants are more certain that the conclusion follows when there are few counterexamples, .81 vs. .69; $F(1,70) = 53.06$, $p < .01$. Second, the invalid inferences also showed the standard counterexample effect of few vs. many alternatives: Participants rate the conclusions less certain when more counterexamples can be found for it, .81 vs. .60; $F(1,70) = 153.75$, $p < .01$. Figure 1 also shows that the counterexample effect is larger on the logically invalid inference, as compared to the logically valid inferences; $F(1,70) = 21.38$ $p < .01$.

Figure 2 shows the size of the counterexample effect (few vs. many) as a function of the timing constraint, logically validity and the assumption of truth. Figure 2 clearly shows that the counterexample effect on the logically valid inferences is reduced when people make the truth-assumption, $F(1,70) = 14.21$, $p < .001$, but only so when individuals reason without a timing constraint and focus on accuracy, $F(1,70) = 19.67$, $p < .01$. The counterexample effect does not approach significance in this condition, .866 vs. .853. The interaction between speed and truth at the level of the valid inferences was significant, $F(1,70) = 5.41$, $p < .05$. No such interaction was observed at the level of the invalid inferences ($F = .003$), and the third-level interaction indeed tended to approach statistical significance, $F(1,70) = 2.85$, $p < .10$. Specific comparisons showed that, as expected, the counterexample effect on the valid inferences re-appears when people evaluate the conclusions as fast as possible, .873 vs. .808: $F(1,70) = 9.66$, $p < .01$. That is, stressing the truth-assumption does not reduce the counterexample effect on the logically valid inferences when people are reasoning under a timing constraint ($F < 1$). At the level of the logically invalid inferences, we see an overall reduction of the counterexample effects, $F(1,70) = 13.22$, $p < .01$. This might suggest that stressing the truth-assumption tends to

Figure 1

Certainty ratings of the logically valid and invalid arguments under standard conditions that do not mention the truth assumption.

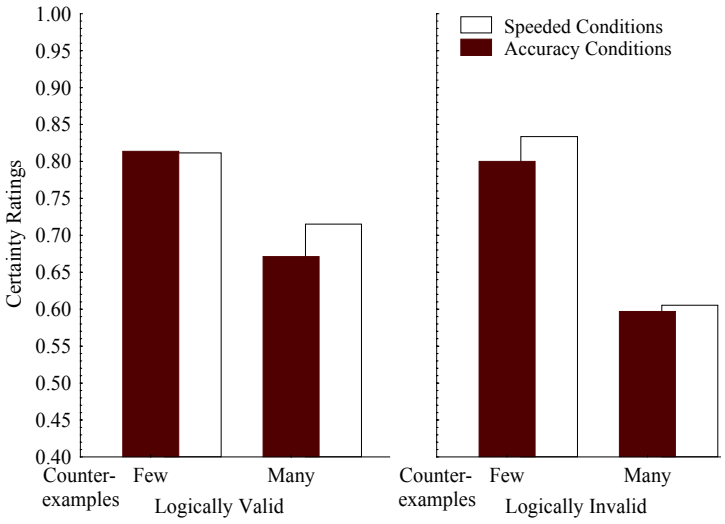
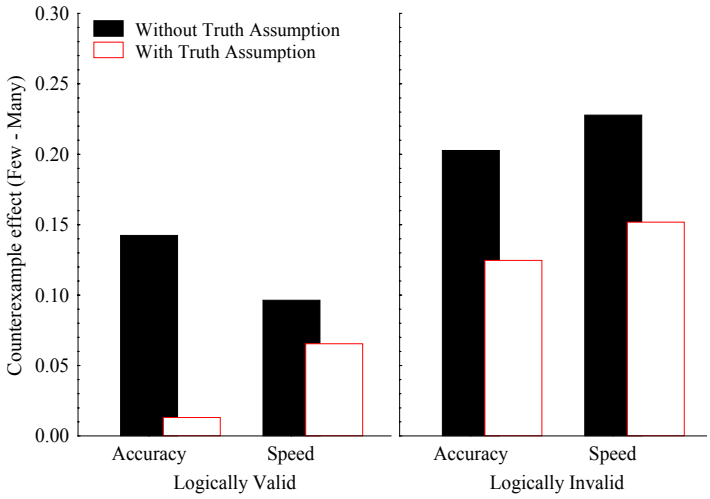


Figure 2

Counterexample effects on the logically valid and invalid arguments as a function of a timing constraint and the explicit presence of the truth assumption.



induce an overall inhibition of background knowledge. The fact that the truth-assumption effect on the valid inferences depends on the timing constraint tells us that this is not the entire story. Also, as noted before, the counterexample effects on the valid inferences are completely annulled when the truth-assumption is stressed (under accuracy conditions), whereas Figure 2 shows that they are still very much present on the invalid inferences under the same conditions. The reduced counterexample effects on the invalid inferences presented under truth conditions concurs with the idea that some people adopt a bi-conditional interpretation of 'if'. The alternative causes are then theoretically or hypothetically (i.e., under the assumption of the truth of the utterance) impossible.

Discussion

Our findings corroborate several of the claims we have made regarding deductive rationality in human reasoning. First, in order to reason deductively one has to make and adhere to the truth-assumption. We observed that the counter-example effect on the logically valid inferences is indeed smaller than that on the logically invalid inferences. The counter-examples to logically valid inferences are indeed (hypothetically) impossible – this is actually why these inferences are logically valid. Second, though we have evidence that some people spontaneously exhibit deductive rationality in adhering to the truth-assumption, other people clearly abandon the truth-assumption in the light of factual evidence to the contrary. The probabilistic counterexample effects on the logically valid inferences attest to this.

The speed/accuracy manipulation and the effects of stressing the truth-assumption provide strong support to our analyses of deductive rationality within a dual processing scheme. First, the overall increase in deductive rationality (as measured by the increase in the certainty ratings of the logically valid inferences) under conditions that stress the truth-assumption lends support the centrality of the truth-assumption in the notion of logically validity and human deductive reasoning. Second, the annulment of the counter-example effects on the logically valid inferences under conditions that make it clear that the truth of inferences about factually false utterances is a hypothetical truth, is also in agreement with the thesis that people inhibit factual knowledge that conflicts with the hypothetical truth of the utterances people reason about. Third, the dependency of the counter-example effect annulment on the time and effort people take to provide an evaluation of the conditional inferences, concurs with (and was predicted on the basis of) the thesis that probabilistic content-driven reasoning is primary to the effortful abstract, analytic hypothetical reasoning processes that can serve to override the output of the fast and frugal heuristic processes.

General Discussion

Our study shows the import and importance of the truth-assumption as regards deductive rationality in human reasoning. In the current general discussion we will touch upon some wider theoretical and conceptual issues. We will first consider the rational basis for the truth-assumption. Next, we will consider the import and importance of the truth-assumption as regards arguments that have been made in discussions of the non-monotonic and/or defeasible nature of human reasoning.

An Implicit vs. Explicit Truth-Assumption. We found support for thesis that at least some people make the truth-assumption and actually stick to it. It remains the case, however, that the majority of people will abandon the truth-assumption. The sizable counterexample effects on the logically valid inferences evidence this. One can only claim that the truth-assumption is abandoned when it is made in the first place. The question that then arises is whether those people who do not follow the truth-assumption (by taking count of factual knowledge to the contrary) actually made it in the first place.

It is our contention that when people form a representation of the utterances they are confronted with, they initially and implicitly make the assumption that the proposition expressed by it is true. This is in accordance with the Gricean maxims of conversation (which by themselves are related to Kant's four a priori categories of quantity, quality, relevance and modality): we generally assume/ensure that our or the speaker's contribution is truthful, relevant and as informative as possible, though not more detailed than required by the context (Grice, 1975). The truth-assumption is an implicit assumption (see, e.g., Schroyens, Schaeken, & d'Ydewalle, 1999). It is partly because it is an implicit assumption (at least to start with) that it is easily abandoned. The rational basis of the truth-assumption can be found in the idea of bounded rationality or cognitive economy. There is a representational cost attached to considering all possibilities, both true and false.

Most current theories presume the truth-assumption. This is not very surprising when one considers that truth is ontologically primordial to falsity: Non-truth presumes truth – as non-being presumes being. The mental-models theory (Johnson-Laird & Byrne, 2002) is the single one theory that is most explicit in invoking the truth-assumption. Indeed, it forms the basis of the truth-principle as regards the representation of the meaning of conditionals of the form [if p then q]. This principle states that people initially represent only represent true possibilities. Oaksford, Chater, & Larkin (2000) seems to have the only theory for which it is difficult to see whether it incorporates the truth-assumption. They do not seem to distinguish true from false utterances. There are only degrees of truth (i.e., probabilities). This restriction to factual truth (verisimilitude) is problematical because there is plenty of evidence that shows that people can reason hypothetically and deductively.

Truth, Validity and Non-Monotonic Reasoning. We situated the present study within the timely clash between paradigms focusing on deductive or probabilistic reasoning and presented the core argument that is made against logic theories. Theories of human deduction would not be capable to cope with the defeasible nature of human reasoning.

Our introductory analyses of the core argument against mental logic have shown that the issues are more complex: The defeasibility of a conclusion does not necessarily imply the non-monotonicity of an inference. Let us reiterate our

arguments against the claim that logic is in trouble because it is monotonic, while commonsense reasoning would not be Indeed, we have come to the somewhat controversial conclusion that it remains an open question whether commonsense reasoning is non-monotonic (even though we know it is defeasible).

We know that the counterexamples to the Modus Ponens argument (MP: if p then q, p, therefore q) are cases that naive reasoners (as opposed to logicians) consider impossible if the conditional utterance is true (Evans, Ellis, & Newstead, 1996). When they assume that [if p then q] is true, most people generally judge that it would be impossible that there are [p and not-q]-contingencies: situations wherein the consequent does not follow from the antecedent. In short, when people defeat a logically valid inference this simply indicates that peoples' intuitive notion of validity does not match that of logical validity. The pervasive 'belief effects' show that reasoners are much more concerned with the factual truth of a conclusion (Tweety the ostrich does not fly), as compared to the hypothetical truth of such conclusions (if it were true that all birds fly then Tweety the ostrich would fly).

Since logical validity encompasses the truth assumption, defeating a necessary inference marks the abandonment of this truth-assumption. By consequence it remains undetermined whether people have reasoned non-monotonically (i.e., revised a judgment of logical validity into a judgment of logical invalidity). When we assume, for arguments sake, that people actually aim to derive logically valid inferences, the defeasibility of inferred inferences shows that people shift from one notion of validity (i.e., logical validity, which includes the truth-assumption) to another notion of validity (let us call it 'psychological validity', which gives more weight to factual truth and allows a truth-assumption to be annulled). It seems one succumbs to the fallacy of equivocating two distinct concepts (logical and psychological validity), when defeasibility of an inference is taken to indicate non-monotonicity of human reasoning.

Because classic logic is monotonic while everyday reasoning is presumably non-monotonic (or at least defeasible), it has been stated that neither mental-models theories nor mental-logic theories are capable of explaining common-sense reasoning. It is hard to see why polemics have been created when defeating inferences is actually at the heart of mental models theory. Mental-models theory holds to a three-stage processing scheme. People first generate initial (incomplete) representations of what they think is possible if the premise are true (model-construction); they then integrate the representation of the multiple source of information that form a reasoning problem (model-integration). This allows them to generate a putative conclusion, which, third and most importantly, at least some people at least sometimes attempt to test by looking for a counterexample. A conclusion is rejected and/or modified in the light of conflicting information. That is, defeasible reasoning is in no way beyond the reach of mental-models theory, quite on the contrary:

"It is worth given up, not the thesis that human beings are capable of rational thought, but the idea that what underlies this ability is a mental logic. There can be reasoning without logic. More surprisingly, perhaps, there can be valid reasoning without logic" (Johnson-Laird, 1983, p. 40).

Acknowledgments

The present research was done with the support of the Flanders (Belgium) Fund for Scientific Research and the Research Council of the University of Leuven. We also like to express our grateful acknowledgments to Sunile Maes and Lieven Brebels for their help in collecting the data.

References

Chater, N., & Oaksford, M. (1999). The probabilistic heuristics model of syllogistic reasoning. *Cognitive Psychology, 38(2)*, 191-258.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition, 19,* 274-282.

Evans, J. St. B. T., Ellis, C. E., & Newstead, S. E. (1996). On the mental representation of conditional sentences. *Quarterly Journal of Experimental Psychology, 49A,* 1086-1114.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality in reasoning*. Hove, UK: Psychology Press.

Grice, H. P. (1975). *Logic and conversation.* In P. Cole, & J. L. Morgan (Eds.), Studies in syntax: Speech acts,Vol. 3, (pp. pp. 41-58).New York: Academic Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109(4)*, 646-678.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 883-899.

Politzer, G., & Bourmaud, G. (2002). Deductive reasoning from uncertain conditionals. *British Journal of Psychology, 93*, 345-381.

Schroyens, W., Schaeken, W., & d'Ydewalle, G. (1999). Error and bias in meta-propositional reasoning: A case of the mental model theory. *Thinking and Reasoning, 5(1),* 29-65.

Schroyens, W., Schaeken, W., & Handley, S. (2003). In Search of Counter Examples: Deductive Rationality in Human Reasoning. *Quarterly Journal of Experimental Psychology, 56A(7),* 1129-1145.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645-726.

Enhancing Example-Based Learning in Hypertext Environments

Julia Schuh (j.schuh@iwm-kmrc.de)
Virtual Ph.D. Program: Knowledge Acquisition and Knowledge Exchange with New Media
Konrad-Adenauer-Strasse 40, 72072 Tuebingen Germany

Peter Gerjets (p.gerjets@iwm-kmrc.de)
Multimedia and Hypermedia Research Unit, Knowledge Media Research Center
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Department of Applied Cognitive Psychology and Media Psychology, University of Tuebingen
Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Abstract

In previous research, Gerjets, Scheiter, and Tack (2000) demonstrated that learners experience serious difficulties in utilizing instructional examples according to their profitability when interacting with a hypertext-based learning environment. In this paper we focus on possible causes of these difficulties and on different instructional methods for improving learners' utilization of worked-out examples in hypertext environments. The results of two experimental studies are reported.

Learning from Worked-Out Examples: The Role of Example Processing Strategies and Example Design

Research over the last 15 years in the domain of learning and problem solving has demonstrated that instructional examples play an important role for knowledge acquisition in domains like mathematics, physics, or programming (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). In particular for initial skill acquisition, learning from worked-out examples seems to be superior to actively solving training problems (Sweller, van Merriënboer, & Paas, 1998). However, numerous findings also indicate major drawbacks of example-based learning. In particular, poor learners tend to overuse examples during problem solving without reflecting on their appropriateness (VanLehn & Jones, 1993). In addition, learners have difficulties identifying relevant information in worked-out examples and are often distracted by examples' surface features (Ross, 1989). Furthermore, Renkl (1999) assumes that students often suffer from illusions of understanding when learning from worked-out examples. I.e., they may have the false impression of having grasped the solution rationale of an example problem. Finally, learners have difficulties generalizing solutions from examples to novel problems (Catrambone & Holyoak, 1989; Reed, Dempster, & Ettinger, 1985).

A number of empirical studies have identified features of *example processing strategies* and *example design* that are efficient for successful knowledge acquisition (cf. Atkinson, Derry, Renkl, & Wortham, 2000).

Important *strategical aspects* mainly concern the *adequate selection and elaboration* of instructional examples. Reed, Ackinclose, and Voss (1990) showed that learners failed to select sufficiently complex instructional examples for learning although the profitability of these examples for *subsequently solving test problems* could be demonstrated. However, Reed, Willis, and Guarino (1994) found that learners who were allowed to select worked-out examples *while solving test problems* were able to select suitable examples. Additionally, it has been shown that *self-explanations* are an important aspect of good learners' example processing (Chi et al., 1989; Pirolli & Recker, 1994; Renkl, 1997). In particular, anticipations of solution steps and inferences with regard to the relations between solution steps, goals, and abstract principles have been proven useful for knowledge acquisition.

With respect to *design issues* it could be shown that *multiple examples* can support schema induction which helps learners to solve novel problems (Cummins, 1992). Providing multiple examples with *different surface features* might further improve this process of abstraction (Quilici & Mayer, 1996). Additionally, it has been proposed that the provision of *completion problems* - where learners have to fill in some details of worked-out examples' solution steps - is a helpful instructional device as it fosters self-explanations (Van Merrienboer, 1990). In particular, presenting completion problems along with evaluative feedback on subjects' gap-filling performance seems to improve learning outcomes. For instance, Stark (1999) showed that learners benefit from such a combination of completion problems and feedback and stresses the point that completion problems foster example elaboration whereas giving feedback on the learning success might prevent learners from *illusions of understanding*.

From these findings on learning from examples it can be argued that *strategies of example selection and processing* as well as *features of example design* have to be taken into account to improve learning outcomes.

The aspect of adopting suitable strategies gains increasing importance the more the control of the learning process is left up to the learner. In learning situations where the learner can *select* instructional material as well as determine the *sequence* and the *pace* of presentation, the importance of strategies increases (Gerjets, Scheiter, & Tack, 2000). Therefore, an identification of suitable strategies of information utilization and an examination of whether learners can adopt these strategies is highly relevant when more focus is put on self-regulated learning in the field of instruction.

Example-Based Hypertext Environments

One domain in which these issues of learner control are stressed is the field of hypertext-based learning where the user can select among different kinds of information and where he can choose according to his goals when the information is to be presented and in which order (Rouet & Levonen, 1996).

On the one hand, this allows for great flexibility and adaptivity of learning and problem solving. Generally, it is assumed that the nonlinear structure of hypertext environments improves learners' ability to use knowledge in a flexible way, so that they learn to apply one information unit to serve different purposes in a variety of situations. Non-linearity also enables learners to utilize information units according to their goals and to their prior knowledge. With regard to example-based learning, providing multiple examples with different surface features in a nonlinear hypertext environment allows the learner to compare examples within one problem category as well as to compare examples between different problem categories. These comparisons are fundamental for processes of abstraction in that they allow learners to identify structural features that define different problem categories. Therefore, non-linearity and the resulting opportunities of flexible information utilization may be especially suitable when learning from examples.

On the other hand, learners can "face new problems in selecting and accessing relevant information" (Rouet, Levonen, Dillon, & Spiro, 1996, p. 3). Problems can arise if learners do not possess the necessary prerequisites to cope with the demands that have been imposed to them by redirecting control over the learning process to them (Rouet & Levonen, 1996). Learning with a nonlinear hypertext increases the amount of control demands by making it necessary that learners permanently make decisions about the profitability of individual information units with regard to their current learning tasks. Even if all information provided is relevant to the current task, the information items may differ with respect to their profitability in terms of their processing costs and their contribution to improving the learning outcome (cf. Pirolli & Card, 1999). Therefore, learners may have to develop adequate strategies of information selection and processing in order to make use of the potential benefits of hypertext-based information presentation.

Based on these considerations a question of central importance in example-based learning with hypertext is whether learners are capable of utilizing examples according to their profitability, i.e., select, sequence, and compare them in a suitable way. Most research on learning from examples up to now has focused on learning situations where learners have been forced to process the examples provided in a predefined sequence and, in some studies, even for a fixed amount of learning time. However, it is not clear whether these findings can be easily transferred to more natural learning situations that allow subjects to select information in different sequences and to control their own pace of studying.

Results of Previous Experiments

In a series of previous experiments Gerjets, Scheiter, and Tack (2000) demonstrated that learners experience difficulties in hypertext environments with regard to their ability to utilize examples according to their profitability. These experiments were conducted using a web-based hypertext environment for training and testing in the domain of combinatorics (HYPERCOMB). During the learning phase subjects could retrieve abstract information on six problem categories from the domain of combinatorics. Depending on the experimental condition, this abstract information was either not augmented by any additional instructional information or was augmented by one or three worked-out examples that illustrated the six problem categories. Learners could retrieve the information they wanted to study and could determine the pace and sequence of information presentation. When they had the impression that they had learned sufficiently well learners could switch to a test phase where they had to solve three test problems. Automated logfile analyses were used to track subjects' strategic navigation behavior. Additionally, subjects' problem-solving performance was registered.

In order to investigate strategic adaptation to different instructional situation Gerjets et al. (2000) studied learners with either *low or high domain-specific prior knowledge* using different instructional versions of HYPERCOMB (no or one example or three examples per problem category) *with or without time pressure*.

As a result of their experiments Gerjets et al. (2000) showed that learners have difficulties in selecting the most profitable information in a specific instructional situation.

A comparison among the three instructional conditions yielded no beneficial effects of *merely providing* examples compared to providing only abstract information. However, if subjects *made use* of the examples in a suitable way (e.g. by comparing different examples) this clearly improved their learning and problem solving performance compared to subjects who made insufficient use of the instructional material. These findings on information profitability were contrasted with learners' actual information utilization behavior. Despite the fact that example processing proved to be useful, about half of the subjects demonstrated poor example processing strategies as they neither processed each example in the one-example condition more than once nor did they study more than one example per problem category in the three-example condition.

Hypotheses: Possible Explanations for Learners' Problems to utilize instructional examples according to their profitability

There might be two different explanations for learners' problems to utilize instructional examples according to their profitability, which will be described in the following paragraphs. Furthermore, two experimental conditions will be outlined that were designed to counteract these hypothesized causes of subjects' failures in using examples adequately.

Non-linearity

A first explanation is related to the fact that the experimental material is designed as a nonlinear hypertext environment. According to Niederhauser, Reynolds, Salmen, and Skolmoski (2000) additional control demands caused by non-linearity may result in extraneous cognitive load (cf. Sweller, Van Merrienboer, & Paas, 1998), which in turn impedes learning activities. Learners may suffer from

cognitive overload due to additional control and navigational demands caused by the nonlinear environment.

Additionally, learners may be in general overwhelmed by the need to *decide which information is profitable to select in which situation* (cf. Rouet et al., 1996). As a result subjects may be unable to utilize examples according to their profitability, either because of cognitive overload or because of inadequate navigational decisions. In order to counteract these problems of cognitive overload and of information selection we introduced a *linear-hypertext condition* of HYPERCOMB that contained exactly the same information as the *nonlinear-hypertext condition* and that forced learners to recognize every information available in a predefined order. Thereby, only pacing was left up to the learner. Eliminating the need to select and sequence information should reduce extraneous cognitive load and should free cognitive resources for processing instructional examples adequately. Furthermore, profitability judgements are less critical in a linear-hypertext condition.

Illusions of understanding

A second explanation for the insufficient use of the examples provided in HYPERCOMB is not related to the non-linearity of the information presentation but is related to the fact that learners may suffer from *illusions of understanding* when learning from worked-out examples (Renkl, 1999). To prevent learners from such illusions we introduced an instructional condition with incomplete examples and feedback where we presented fragmentized example solutions and asked the learners to complete these gaps by selecting one of two possible multiple-choice answers. After the completion, learners were provided with feedback concerning the correctness of their answers. This procedure may improve intensive example processing as it may help learners to realize that they are far away from an in-depth understanding of the example solutions. As a result, learners may notice that examples proved a profitable source of information and are helpful in order to overcome these comprehension failures.

In experiment 1 a nonlinear version of HYPERCOMB with three complete worked-out examples per problem category (baseline condition) was compared to a linear version in order to test the first hypothesis that subjects' inadequate use of examples results from additional navigational and control demands in nonlinear hypertext.

In experiment 2 the baseline condition was compared to a condition with three incomplete examples with feedback in order to test the second hypothesis that subjects' failures in using examples adequately results from an illusion of understanding.

We expected that both instructional manipulations should increase the time spent on processing examples and thereby improve learning outcomes. Additionally, we assumed that the instructional devices would especially foster learning outcomes of subjects with low prior knowledge who may suffer from control demands as well as from illusions of understanding to a greater extent that learners with high prior knowledge.

Experiment 1: Linear Hypertext

Method

Participants Subjects were 80 students of the Saarland University, Germany, who either participated for course credit or for payment. Average age was 23.4.

Materials and procedure Subjects used the HYPERCOMB environment for learning and problem solving. First, a short introduction to the domain of combinatorics was presented. During the subsequent learning phase subjects could retrieve abstract information on six problem categories (defined by their associated formula) from the domain of combinatorics. Additionally, three worked-out examples that varied with regard to their complexity and their cover story were provided for each problem category. In the test phase subjects were instructed to solve three probability word problems. Neither the abstract information nor the worked-out examples of the learning phase were available during the test phase.

Design and dependent measures As a first independent variable two levels of domain-specific prior knowledge were introduced. Additionally, two different instructional conditions were implemented (2 x 2 design):

In the *nonlinear-hypertext condition (baseline)* learners could choose by themselves which information to retrieve (i.e., abstract information and three different examples per problem category) and in which sequence to pursue. Learners could retrieve all information pages as often as they wanted and they could study them as long as they wanted. The learners themselves controlled the learning process so they could as well neglect all the provided information as study them very carefully. This condition served as baseline condition.

In the *linear-hypertext condition* the same instructional material as in the first condition was presented in a linear fashion. Learners had to follow a so-called *guided tour* through the hypertext environment by using "next"-buttons to get from one page to another. All problem categories were explained successively. For each problem category, first the abstract information page was displayed followed by three example pages. Every information page was presented only once, learners could not look back to information already seen and they could not skip any of the information. Therefore, selection and sequencing of the information pages were controlled by the system and only pacing was left up to the learner.

In the test phase subjects had to solve three word problems by marking the appropriate solution principle and the values of two variables for each of the test problems in a multiple-choice form. No calculations had to be made. One error was assigned for each wrong answer (i.e., subjects could obtain overall error rates between 0 and 9). Problem-solving time as well as learning time on example pages and on abstract pages was recorded by using logfiles. Following the test phase subjects had to pass a knowledge test with multiple-choice questions related to abstract concepts from the domain of combinatorics. Similar questions were posed as a pretest at the beginning of the experiment to register subjects' domain-specific prior knowledge. Subjects were

assigned to high and low prior knowledge groups by means of a median splits according to their pretest results.

Results and Discussion

First, we compared high and low prior-knowledge subjects learning either in the nonlinear or in the linear-hypertext condition within the two levels of prior knowledge with regard to their pretest errors (table 1). An overall ANOVA (instructional condition x prior knowledge) yielded no differences between the instructional conditions (F < 1).

Table 1: Time data (in sec) and error rates (in %) as a function of prior knowledge and instructional condition

Instructional condition	**Nonlinear**		**Linear**	
Prior knowledge	**High**	**Low**	**High**	**Low**
Pretest errors	28.6	63.2	30.3	64.3
Time on example pages	608	465	948	1069
Problem-solving time	602	550	617	606
Problem-solving errors	33.3	42.2	27.2	39.4
Knowledge-test errors	11.0	32.0	18.5	26.5

Time data With regard to example-processing time, an overall ANOVA (instructional condition x prior knowledge) yielded a significant main effect of instructional condition. Subjects in the linear-hypertext condition spent more time on example pages than subjects in the nonlinear-hypertext condition ($F(1,76) = 36.39$; $MS_E = 122363.7$; $p < .001$). Additionally, it could be shown that there was no difference between subjects with high and low prior knowledge concerning example processing time ($F < 1$). The interaction between instructional condition and prior knowledge was marginally significant ($F(1,76) = 2.84$; $MS_E = 122363.7$; $p < .10$). The increase of example-processing time due to the linear information presentation was slightly more pronounced for subjects with low prior knowledge ($t(38) = 5.37$; $p < .001$) than for subjects with high prior knowledge ($t(38) = 3.13$; $p < .01$).

Furthermore, we analyzed whether there was a *trade-off* between example-processing time and problem-solving time, in that learners in the nonlinear hypertext-condition might need less time for studying the examples but more time for later problem solving. However, an ANOVA for problem-solving time yielded no significant results (all Fs < 1).

Performance data In order to test the assumption that an increase in example-processing time leads to better problem-solving performance, we conducted an overall ANOVA (instructional condition x prior knowledge) that, however, only yielded a significant main effect for prior knowledge ($F(1,76) = 5.59$; $MS_E = 399$; $p < .05$). There neither was a main effect for instructional condition nor was there an interaction between instructional condition and prior knowledge (both Fs < 1).

Besides problem-solving performance we also analyzed knowledge-test performance as an indicator of learning success. An overall ANOVA (instructional condition x prior knowledge) for knowledge-test errors yielded a main effect for prior knowledge ($F(1,76) = 10.51$; $MS_E = 400.1$; $p < .01$) which was due to the high degree of item overlap between the pretest and the knowledge test at the end of the experiment. There was, however, no main effect for instructional condition as well as no interaction (both Fs < 1).

To conclude, providing subjects with linear hypertext increased example-processing time to a large extent as expected. However, this increase in time spent on examples was not accompanied by the expected gains in performance. Furthermore, it could be shown that learning in the nonlinear-hypertext condition was even more efficient because subjects needed less example-processing time for achieving the same level of performance without increases in problem-solving time. Therefore, our first explanation that subjects shallow example processing in hypertext environments observed in previous experiments (Gerjets et al., 2000) can not be traced back to additional control and navigational demands caused by nonlinear information presentation: Reducing these demands does not result in improved performance although example utilization behavior is intensified.

Thus, it seems that merely quantitative increases in example-processing time are not sufficient to ensure successful learning. Therefore, in experiment 2 we implemented incomplete examples with feedback as an instructional method that focuses on more qualitative improvements of example processing instead of only increasing example-processing time. This is in accordance with our second hypothesis that superficial example processing can be traced back to illusions of understanding when learning from worked-out examples.

Experiment 2: Incomplete Examples with Feedback

Method

Participants Subjects were 80 students of the Saarland University, Germany who either participated for course credit or payment. Average age was 23.7 years.

Materials and procedure Subjects used the same HYPERCOMB environment for learning and problem solving as the subjects in the nonlinear-hypertext condition in experiment 1. It consisted in a short introduction to combinatorics, a learning phase with abstract information and three worked-out examples per problem category, and finally a subsequent test phase with three probability word problems.

Design and dependent measures As a first independent variable two levels of domain-specific prior knowledge were introduced. Additionally, two instructional conditions were implemented (2 x 2 design):

As a *baseline condition* we used the nonlinear-hypertext condition from experiment 1 where subjects could decide by themselves which information to review (abstract information and three fully worked-out examples per problem category) and in which sequence to pursue.

In the *feedback condition* the solution steps of the worked-out examples where fragmentized and subjects where asked to fill these gaps by choosing among two multiple-choice answers. It is, however, important to note that subjects could decide by themselves whether they filled in the gaps and used the opportunity to receive feedback or not. Every

example solution was fragmentized two or three times and the gaps were related to structural features of the problem categories. After having determined the gap-filling answer subjects automatically received feedback on whether their answer was right or not. In case of choosing the wrong alternative the right answer was presented.

The subsequent test phase was identical to experiment 1. As dependent measures error rates and time date were recorded. In the feedback condition the frequency of feedback utilization was additionally registered. As in experiment 1 subjects had to pass a knowledge test with multiple-choice questions related to abstract concepts from the domain of combinatorics after the test phase. Subjects' answers to similar questions at the beginning of the experiment were used to distinguish between low and high prior-knowledge subjects.

Results and Discussion

The results of experiment 2 are shown in table 2. A first comparison revealed that there were no differences between the instructional conditions with respect to pretest errors ($F < 1$). (cf. table2).

Table 2: Time data (in sec) and error rates (in %) as a function of prior knowledge and instructional condition

Instructional condition	**Baseline**		**Feedback**	
Prior knowledge	**High**	**Low**	**High**	**Low**
Pretest errors	28.6	63.2	33.6	63.7
Time on example pages	608	465	692	785
Problem-solving errors	33.3	42.2	36.4	46.1
Knowledge test errors	11.0	32.0	21.5	27.0

Time data In order to test the hypothesis that subjects in the feedback condition process the examples more intensively than subjects in the baseline condition, we conducted an overall ANOVA (instructional condition x prior knowledge) for the time spent on example pages. However, this ANOVA only yielded a marginally significant main effect for instructional condition ($F(1,76) = 3.67$; $MS_E = 222102.5$; $p < .10$) with subjects in the feedback condition spending more time on studying example pages than subjects in the baseline condition. There was neither a main effect for prior knowledge ($F < 1$) nor an interaction ($F(1,76) = 1.25$; $MS_E = 222102.5$; $p > .40$).

Performance data With regard to problem-solving errors an overall ANOVA (instructional condition x prior knowledge) yielded no main effect for the instructional condition ($F < 1$). Thus, although there was a slight increase in example-processing time in the feedback condition this increase was not accompanied by respective improvements in problem-solving performance. For prior knowledge, the analysis yielded a main effect ($F(1,76) = 4,08$; $MS_E = 658.8$; $p < .05$). No interaction between the two factors could be demonstrated ($F < 1$).

Additionally, we conducted an ANOVA (instructional condition x prior knowledge) for knowledge-test errors, which yielded no significant main effect for the knowledge test errors ($F < 1$). The interaction between these two factors was marginally significant ($F(1,76) = 3.56$; $MS_E = 381.3$; $p < .10$).

To conclude, at first sight asking subjects to fill in gaps and providing feedback on these gap-filling activities does not seem to be an effective way of improving subjects' learning outcomes in example-based hypertext environments. However - because the use of feedback was not obligatory - it can be expected that only subjects who retrieved feedback sufficiently would often show better learning outcomes.

Therefore we calculated the correlation between the number of times subjects used feedback and the resulting learning outcomes in the feedback condition for high and low prior-knowledge subjects separately. These analyses show that subjects with low prior knowledge indeed benefited from an extended use of feedback (correlation between frequency of feedback utilization and problem-solving errors: $r = -.45$; $p < .05$; knowledge-test errors: $r = -.44$; $p < .05$, one-tailed test) whereas there were no or only weak associations between frequency of feedback utilization and learning outcomes for high prior-knowledge subjects (problem-solving errors: $r = -.12$; $p > .30$; knowledge-test errors: $r = -.31$; $p < .10$).

To sum up, subjects with low prior knowledge who made sufficient use of feedback clearly improved their problem-solving performance compared to subjects with low prior knowledge who made insufficient use of the instructional material. This is in line with our second hypothesis that the provision of incomplete examples with feedback may be useful to reduce illusions of understanding. These findings on the profitability of feedback information for learners with low prior knowledge were contrasted with their actual information utilization behavior in a next step of analysis: Despite the fact that the use of feedback proved useful for learners with low prior knowledge, they did not retrieve feedback more often than learners with high prior knowledge ($t(38) = -.42$; $p > .60$; 2-tailed test). To conclude, although the use of feedback fostered problem-solving and knowledge-test performance of low prior-knowledge subjects they did not use it more extensively. Thus, similar to the findings of example utilization it could be demonstrated again that subjects may experience serious difficulties in utilizing beneficial information provided in hypertext-environments according to its profitability.

Conclusions

With regard to the impact of two different instructional manipulations reported in this paper the following conclusion can be drawn. Although a linear information presentation increases example-processing time this does not automatically lead to improvements in learning outcomes. Linear presentation might reduce extraneous cognitive load due to control and navigational demands in nonlinear hypertext environments, however, this may also imply that learning advantages of nonlinear environments are neutralized. I.e., learners in a linear environment no longer have the opportunity to select and sequence information according to their needs. This lack of opportunity to self-control information utilization also may impair important processes of example comparison. Therefore, there may be a trade-off between the benefits and the drawbacks of non-linearity.

With regard to the provision of incomplete examples with feedback it could be demonstrated that this

instructional device was beneficial for subjects with low prior knowledge but that these subjects often made insufficient use of it.

Therefore, learners do not only have problems in utilizing worked-out examples according to their profitability for learning but also in using feedback extensively when learning in nonlinear environments. On the one hand, learners skip helpful information like feedback if they can control their learning process by themselves. On the other hand, when restricting learner control by presenting information in a linear environment learning becomes less efficient.

Therefore, the development of a learning environment where both of these findings are combined might be most successful. It can be assumed that learning in a nonlinear hypertext environment might be improved by forcing subjects to recognize information units of crucial importance for learning - like a minimal number of examples or the use of feedback on example completions. When developing learning environments, the specific learning situation must be considered to guarantee the advantages of non-linearity and at the same time to reduce the drawbacks by forcing the user to recognize profitable information. There must be a balance between the control that is given to the learner and the system control.

Acknowledgments

This work was supported by German Research Foundation (Collaborative Research Center 378: Resource-Adaptive Cognitive processes at the Saarland University, Saarbruecken). We thank Carina Kraemer for conducting the experiments and Simon Albers for programming work.

References

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2001). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181-214.

Catrambone, R., & Holyoak, K. J. (1990). Learning subgoals and methods for solving probability problems. *Memory and Cognition, 18*, 593-603.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Cummins, D. D. (1992). Role of analogical reasoning in the induction of problem categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1103-1124.

Gerjets, P., Scheiter, K., & Tack, W. H. (2000).Resource-adaptive selection of strategies in learning from worked-out examples. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings from the 22nd Annual Conference from the Cognitive Science Society* (pp. 166-171). Mahwah, NJ: Erlbaum.

Niederhauser, D. S., Reynolds, R. E., Salmen, D. J., & Skolmoski, P. (2000). The influence of cognitive load on learning from hypertext. *Journal of Educational Computing Research, 23*, 237-255.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorise statistics word problems. *Journal of Educational Psychology, 88*, 144-161.

Reed, S. K., Ackinclose, C. C., & Voss, A. A. (1990). Selecting analogous solutions: Similarity versus inclusiveness. *Memory & Cognition, 18*, 83-98.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 106-125.

Reed, S. K., Willis, D., & Guarino, J. (1994). Selecting examples for solving word problems. *Journal of Educational Psychology, 86,* 380-388.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21, 1-29.

Renkl, A. (1999). Learning mathematics from worked-out examples: Analysing and fostering self-explanations. *European Journal of Psychology of Education, 14*, 477-488.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 456-468.

Rouet, J.-F., & Levonen, J. J. (1996). Studying and learning with hypertext: Empirical studies and their implications. In J.-F. Rouet, J. J. Levonen, A. Dillon & R. J. Spiro (Eds.), *Hypertext and cognition*. Mahwah, NJ: Erlbaum.

Rouet, J.-F., Levonen, J. J., Dillon, A., & Spiro, R. J. (1996). An introduction to hypertext and cognition. In J.-F. Rouet, J. J. Levonen, A. Dillon & R. J. Spiro (Eds.), *Hypertext and cognition*. Mahwah, NJ: Erlbaum.

Stark, R. (1999). Lernen mit Lösungsbeispielen. Göttingen: Hogrefe.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*, 643-675.

Pirolli, P. & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235-275.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 527-580). Cambridge, MA: MIT Press.

VanLehn, K., & Jones, R. M. (1993). Better learners use analogical problem solving sparingly. In P. E. Utgoff (Ed.*), Machine Learning: Proceedings of the Tenth Annual Conference*. San Mateo, CA: Morgan Kaufmann.

Van Merrienboer, J. J. G. (1990). Strategies for programming instruction in high school: program completion vs. program generation. *Journal of Computing Research, 6*, 265-286.

A Probabilistic Framework for Model-Based Imitation Learning

Aaron P. Shon, David B. Grimes, Chris L. Baker, and **Rajesh P.N. Rao**
{aaron, grimes, clbaker, rao}@cs.washington.edu
CSE Department, Box 352350 University of Washington Seattle WA 98195 USA

Abstract

Humans and animals use imitation as a mechanism for acquiring knowledge. Recently, several algorithms and models have been proposed for imitation learning in robots and humans. However, few proposals o er a framework for imitation learning in a stochastic environment where the imitator must learn and act under real-time performance constraints. We present a probabilistic framework for imitation learning in stochastic environments with unreliable sensors. We develop Bayesian algorithms, based on Meltzo and Moore's AIM hypothesis for infant imitation, that implement the core of an imitation learning framework, and sketch basic proposals for the other components. Our algorithms are computationally efficient, allowing real-time learning and imitation in an active stereo vision robotic head. We present results of both software simulations and our algorithms running on the head, demonstrating the validity of our approach.

Imitation learning in animals and machines

Imitation is a common mechanism for transferring knowledge from a skilled agent (the *instructor*) to an unskilled agent (or *observer*) using direct demonstration rather than manipulating symbols. Various forms of imitation have been studied in apes [Visalberghy and Fragaszy, 1990, Byrne and Russon, 2003], in children (including infants only 42 minutes old) [Meltzoff and Moore, 1977, Meltzoff and Moore, 1997], and in an increasingly diverse selection of machines [Fong et al., 2002, Lungarella and Metta, 2003]. The attraction for machine learning is obvious: a machine with the ability to imitate has a drastically lower cost of reprogramming than one which requires programming by an expert. Imitative robots also offer testbeds for cognitive researchers to test computational theories, and provide modifiable agents for contingent interaction with humans in psychological experiments.

Few previous efforts have presented biologically plausible frameworks for imitation learning. Bayesian imitation learning has been proposed to accelerate Markov decision process (MDP) learning for reinforcement learning agents [Price, 2003]; however, this framework chiefly addresses the problem of learning a forward model of the environment [Jordan and Rumelhart, 1992] via imitation (see below), and the correspondence with cognitive findings in humans is unclear. Other frameworks have been proposed for imitation learning in machines [Breazeal, 1999, Scassellati, 1999, Billard and Mataric, 2000], but most of these are not designed around a coherent probabilistic formalism such as Bayesian inference. Probabilistic methods, and Bayesian inference in particular, are attractive because they handle noisy, incomplete data, can be tuned to handle realistically large problem sizes, and provide a unifying mathematical framework for reasoning and learning. Our approach is unique in combining a biologically inspired approach to imitation with a Bayesian framework for goal-directed learning. Unlike many imitation systems, which implement only software simulations, this paper demonstrates the value of our framework through both simulation results and a real-time robotic implementation.

Components of an imitation learning system

The observer must surmount a number of problems in attempting to replicate the behavior of the instructor. Although described elsewhere [Schaal et al., 2003, Rao and Meltzoff, 2003], we briefly reformulate them as follows:

1. **State identification:** Ability to classify high-dimensional sensor data into a lower-dimensional, relevant state robust to sensor noise. State identification should differentiate between the internal state of the observer (proprioceptive feedback, etc.) and the state of the environment, including the states of other agents, particularly the instructor.
2. **Action identification:** Ability to classify sequences of states in time.
3. **State mapping:** Transformation from the egocentric coordinate system of the instructor to the egocentric coordinate system of the observer.
4. **Model learning:** Learning forward and inverse models [Blakemore et al., 1998] to facilitate interaction with the environment.
5. **Policy learning:** Learning action choices that maximize a reward function, as observed from the actions selected by the instructor in each given state.
6. **Sequence learning and segmentation:** Ability to memorize sequences of key states needed to complete

an imitation task; ability to segment imitation tasks, and to divide tasks into subtasks with particular subgoal states.

A Bayesian framework for goal-directed imitation learning

Imitation learning systems that learn only state and action mappings (without modeling the environment or the instructor's goals) ignore the separability of the instructor's intent from the actions needed to accomplish that intent. Systems that use deterministic models rather than probabilistic ones ignore the stochastic nature of realistic environments. We propose a goal-directed Bayesian formalism that overcomes both of these problems. The notation s_t denotes the state (both internal and external to an agent) at time t, and a_t denotes the action taken by an agent at time t. s_G denotes a special "goal state" that is the desired end result of the imitative behavior. The key to viewing imitation learning as a model-based, goal-directed Bayesian task is to identify:

Forward model: Predicts a distribution over future states given current state(s), action(s), and goal(s)—$P(s_{t+1}|a_t, s_t, s_G)$. Models how different actions affect environmental state.

Inverse model: Infers a distribution over actions given current state(s), future state(s), and goal(s)—$P(a_t|s_t, s_{t+1}, s_G)$. Models which action(s) should be selected to transition from one environmental state to another.

Prior model: Infers a distribution over actions given current state(s) and goal(s)—$P(a_t|s_t, s_G)$. Models the policy (or preferences) followed by a particular instructor in transitioning through the environment to achieve a particular goal.

Thus the prior model involves learning an MDP (or a partially observable MDP), while the forward model involves learning a "simulator" of how the environment (possibly including other agents) reacts to actions performed within it. Learning inverse models is a notoriously difficult task [Jordan and Rumelhart, 1992], not least because multiple actions could have mapped from s_t to s_{t+1}. However, using Bayes' rule, we can infer the distribution returned by the inverse model using the forward and prior models:

$$P(a_t|s_t, s_{t+1}, s_G) \propto P(s_{t+1}|a_t, s_t, s_G) \Pr(a_t|s_t, s_G) \qquad (1)$$

Equation 1 can be used to either select the maximum a posteriori action to complete a state transition, or to sample over a distribution of alternatives, refining the model (and representing an exploration- exploitation tradeoff reminiscent of reinforcement learning). Sampling from the distribution over actions is also called *probability matching*. Evidence exists that the brain employs probability matching in at least some cases [Herrnstein, 1961, Krebs and Kacelnik, 1991].

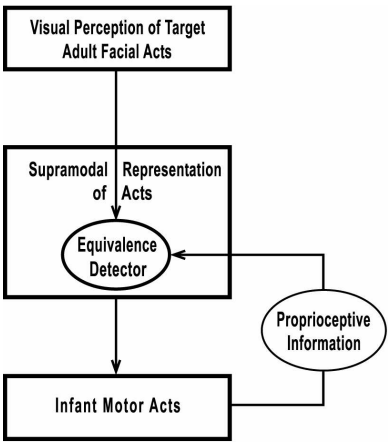


Figure 1: **AIM hypothesis model for infant imitation:** The AIM hypothesis of Meltzo and Moore [Meltzo and Moore, 1997] argues that infants match observations of adults with their own proprioceptions using a modality-independent representation of state. Our computational framework suggests an efficient, probabilistic implementation for this hypothesis.

Fig. 1 graphically represents Meltzoff and Moore's Active Intermodal Mapping (AIM) hypothesis [Meltzoff and Moore, 1997]. According to this cognitive model, imitation begins with an infant (or other agent) forming a representation of features in the outside world. Next, this representation is transformed into a "supra-modal," or modality-independent, representation of those features. An equivalence detector matches the current modality-independent representation of the instructor's state with a modality-independent representation of the infant observer's state. Proprioceptive feedback guides the infant's motor output toward matching the instructor's state. Our framework for Bayesian action selection using learned models captures this idea of imitation as a "matching-to-target" process.

Fig. 2 depicts a block diagram of our architecture. Like AIM, our system begins by running several feature detectors (skin detectors, face trackers, etc.) on sensor inputs from the environment. Detected features are monitored over time to produce state sequences. In turn, these sequences define actions. The next step is to transform state and action observations into instructor-centric values, then map from instructor-centric to observer-centric coordinates. Observer-centric values are employed to update probabilistic forward and prior models in our Bayesian inference framework. Finally, combining distributions from the forward and prior models as in Eqn. 1 yields a distribution over actions. The resulting distribution over actions is converted into a single motor action the observer should take next, with an efference copy conveyed to the feature detectors to cancel out the effects of self-motion.

State and action identification

Deriving state and action identity from sensor data involves task- and sensor-specific functions. Although it is impossible to summarize the extensive

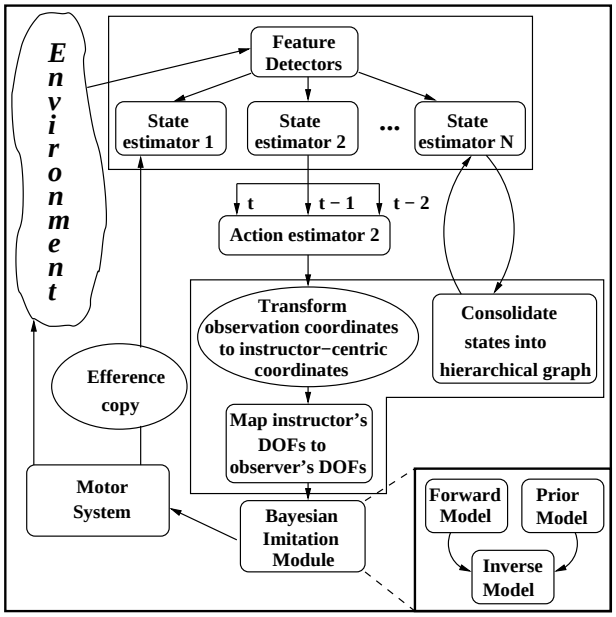


Figure 2: **Overview of model-based Bayesian imitation learning architecture:** As in AIM, the initial stages of our model correspond to the formation of a modality-independent representation of world state. Mappings from instructor-centric to observer-centric coordinates and from the instructor's motor degrees of freedom (DOFs) to the observer's motor DOFs play the role of equivalence detector in our framework, matching the instructor's motor output to the motor commands of the observer. E erence copy provides proprioceptive feedback to close the motor control loop.

body of work in action and state identification here, we note recent progress in extracting actions from laser rangefinder and radio [Fox et al., 2003] and visual [Efros et al., 2003] data. In most cases, computational expediency necessitates employment of dimensionality reduction techniques such as principal components analysis, Isomap [Tenenbaum et al., 2000], or locally linear embedding [Roweis and Saul, 2000]. Saliency detection algorithms [Itti et al., 1998] may also help reduce high-dimensional visual state data to tractable size.

Learning state mappings

A prerequisite for any robotic imitation task is to determine a mapping from the instructor's state to the observer's [Nehaniv and Dautenhahn, 2002]. We view this state mapping problem as an instance of subgraph isomorphism, where the goal is to match subgraphs from the instructor (corresponding to effectors, e.g. limbs) to their corresponding graphs in the observer. In the simulation and robotic head results shown below, the mappings are trivial; developing detailed graph-theoretic approaches to mapping from instructor states to observer states remains an ongoing topic of investigation.

Learning forward models

Numerous supervised and unsupervised approaches (see e.g. [Jordan and Rumelhart, 1992, Todorov and Ghahramani, 2003]) have been proposed to learn models of the environment, and to discover policies to maximize rewards obtained from the environment. Evidence demonstrates that infants learn forward models of how their limbs, facial muscles, and other body parts react to motor commands, a process referred to by Meltzoff and Moore [Meltzoff and Moore, 1997] as "body babbling." Such forward model learning could occur both prenatally and during infancy. We anticipate using well-established supervised algorithms to acquire forward models of environmental dynamics. Unsupervised learning of forward and inverse models to generate motor policies is a well-known problem in the reinforcement learning community (see [Kaelbling et al., 1996] for a survey). In reinforcement learning, an agent's internal reward signal alone is used to learn models of the environment, rather than relying on examples provided by a teacher as in imitation learning.

Sequence learning and segmentation

Realistic imitation learning systems must be able to learn sequences of states that define actions, and to segment these sequences into meaningful chunks for later recall or replay. Part of our ongoing work is to define how semantically meaningful chunks can be defined and recalled in real time. Recent developments in concept learning (e.g., [Tenenbaum, 1999]) suggest how similar environmental states might be grouped together, enabling development of hierarchical state and action representations in machine systems.

A Bayesian algorithm for inferring intent

Being able to determine the intention of others is a crucial requirement for any social agent, particularly an agent that learns by watching the actions of others. Recent studies have revealed the presence of "mirror neurons" in monkey cortex that fire both when an animal executes an action and when it observes others performing similar actions. These findings suggest a neurological substrate for intent inference in primates [Rizzolatti et al., 2000]. One appealing aspect of our framework is that it suggests a probabilistic algorithm for determining the intent of the instructor. That is, an observer can determine a distribution over goal states based on watching what actions the instructor executes over some period of time. This could have applications in machine learning systems that predict what goal state the user is attempting to achieve, then offer suggestions or assist in performing actions that help the user reach that state. The theory could lead to quantitative predictions for future cognitive studies to determine how humans infer intent in other intelligent agents.

Our algorithm for inferring intent uses applications of Bayes' rule to compute the probability over goal states given a current state, action, and next state obtained by the instructor, $P(s_G|s_{t+1}, a_t, s_t)$. This probability distribution over goal states represents the instructor's intent. One point of note is that $P(s_{t+1}|a_t, s_t, s_G) \equiv P(s_{t+1}|a_t, s_t)$; i.e., the forward model does not depend on the goal state s_G, since the environment is indifferent

to the desired goal. Our derivation proceeds as follows:

$$P(s_{t+1}|a_t, s_t, s_G) = \frac{P(s_{t+1}, s_t, a_t, s_G)}{P(a_t, s_t, s_G)} \quad (2)$$

$$P(s_{t+1}|a_t, s_t, s_G) = \frac{P(s_G|s_{t+1}, a_t, s_t)}{P(s_G|a_t, s_t)} \frac{P(s_{t+1}, a_t, s_t)}{P(a_t, s_t)} \quad (3)$$

Because $P(s_{t+1}|a_t, s_t, s_G) \equiv P(s_{t+1}|a_t, s_t)$, and since $\frac{P(a_t,s_t)}{P(s_{t+1},a_t,s_t)} = \frac{1}{P(s_{t+1}|a_t,s_t)}$:

$$P(s_G|a_t, s_t) = P(s_G|s_{t+1}, a_t, s_t) \quad (4)$$

$$\frac{P(s_G, a_t, s_t)}{P(a_t, s_t)} = P(s_G|s_{t+1}, a_t, s_t) \quad (5)$$

$$\frac{P(a_t|s_G, s_t) P(s_G, s_t)}{P(a_t, s_t)} = P(s_G|s_{t+1}, a_t, s_t) \quad (6)$$

$$P(a_t|s_G, s_t) P(s_G, s_t) \propto P(s_G|s_{t+1}, a_t, s_t) \quad (7)$$

$$P(s_G|s_{t+1}, a_t, s_t) \propto P(a_t|s_G, s_t) P(s_t|s_G) P(s_G) \quad (8)$$

The first of the terms in Eqn. 8 represents the prior model. The second term represents a distribution over states at time t, given a goal state s_G. This could be learned by, e.g., observing the instructor manipulate an object, with a known intent, and recording how often the object is in each state. Alternatively, the observer could itself "play with" or "experiment with" the object, bearing in mind a particular goal state, and record how often each object state is observed. The third term is a prior over goal states; it can be derived by modeling the reward model of the instructor. If the observer can either assume that the instructor has a similar reward model to itself (the "like-me" hypothesis [Meltzoff, 2002]), or model the instructor's desired states in some other way, it can infer $P(s_G)$.

Interestingly, these three terms roughly match the three developmental stages laid out by Meltzoff [Meltzoff, 2002]. According to our hypothesis, the first term in Eqn. 8 corresponds to a distribution over actions as learned during imitation and goal-directed actions. This distribution can be used if all the observer wants to do is imitate body movements (the first step in imitation that infants learn to perform according to Meltzoff's theory of development). The second term in Eqn. 8 refers to distributions over states of objects given a goal state. Because the space of actions an agent's body can execute is presumably much less than the number of state configurations objects in the environment can assume, this distribution requires collecting much more data than the first. Once this second term is learned, however, it becomes easier to manipulate objects to a particular end—an observer that has learned $P(s_t|s_G)$ has learned which states of an object or situation "look right" given a particular goal. The complexity of this second term could explain why it takes babies much longer to learn to imitate goal-directed actions on objects than it does to perform simple imitation of body movements (as claimed in Meltzoff's theory). Finally, the third term, $P(s_G)$, is the most complex term to learn. This is both because the number of possible goal states s_G is huge, and the fact that the observer must model the instructor's distribution over goals indirectly (the observer obviously cannot directly access the instructor's reward model). The observer must rely on features of its own reward model, as well as telltale signs of desired states (e.g., states that the instructor tends to act to remain in, or that cause the instructor to change the context of its actions, could be potential goal states) to infer this prior distribution. The difficulty of learning this distribution could explain why it takes so long for infants to acquire the final piece of the imitation puzzle, determining the intent of others. We did not explicitly design the terms in our intent inference algorithm to match childhood developmental stages; rather, the derivation follows from the inverse model formulation in Eqn. 1 and straightforward applications of Bayes' rule.

Simulation results

Fig. 3 demonstrates imitation results in a purely simulated environment. The task is to reproduce observed trajectories through a maze containing three different goal states (maze locations marked with ovals). This simulated environment simplifies a number of the issues mentioned above: the location and value of each goal state is known by the observer a priori; the movements of the instructor are observed free from noise; the forward model is restricted so that only moves to adjacent maze locations are possible; and the observer can detect when it is next to a wall (although it does not know a priori that it cannot move through walls).

The observer first learns a forward model by interacting with the simulated environment for 500 simulation steps. The instructor then demonstrates 4 different trajectories to the observer (1 to the white goal, 2 to the light gray goal, 1 to the dark gray goal), allowing the observer to learn a prior model. Fig. 3(a) shows the maze environment used in our simulations. Fig. 3(b) shows a sample training trajectory (black arrows) where the instructor moves from location (1,1) to the goal state at (3,3). The solid white line (over arrows) demonstrates the observer reproducing the same trajectory after learning. The observer's trajectory varies somewhat from the instructor's due to the stochastic nature of the environment. Fig. 3(c) shows another training trajectory, comprising 47 steps, where the instructor moves toward the white goal (goal 1). The observer's task for this trajectory is to estimate, at each time step of the trajectory, a distribution over which goal state the instructor is headed toward. During the inference process, the observer does not have direct knowledge of the actions selected by the instructor; it must infer these by monitoring state changes in the environment. The graph in Fig. 3(d) shows this distribution over goals, where data points represent inferred intent averaged over epochs of 8 simulation steps each (i.e., the first data point on the graph represents inferred intent averaged over simulation steps 1-8, the second data point spans simulation steps 9-17, etc., with the last epoch spanning 7 simulation steps). Note that the estimate of the goal is correct over all epochs. The algorithm is particularly confident once the ambiguous section of the trajectory, where the instructor could be moving toward the dark gray or the light gray goal, is passed. Performance of the algorithm would be enhanced by more training; only 4 sample tra-

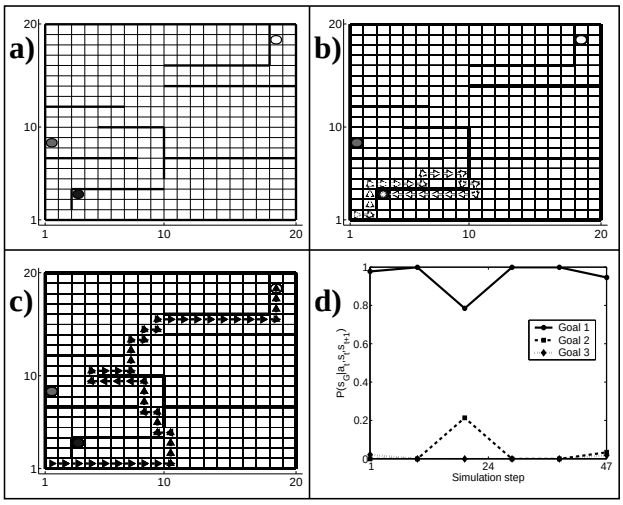


Figure 3: **Simulated environment for imitation learning:** (a) Maze environment used to train observer. Thick black lines denote walls; ovals represent goal states. Lightness of ovals is proportional to the probability of the instructor selecting each goal state (re ecting, e.g., relative reward value experienced at each state). (b) Example trajectory (black arrows) from the instructor, ending at the second goal. Reproduction of the trajectory by the observer is shown as a solid white line overlying the arrows; inference is performed as in Eqn. 1. The instructor required 23 steps to reach the goal; the observer required a slightly larger number of steps due to both the stochastic nature of the environment and imperfect learning of the forward and prior models. (c) Instructor's trajectory in the intention inference task. (d) Graph showing a distribution over instructor's goal states, as inferred by the observer at di erent time points in the simulation. Note how the actual goal state, goal 1, maintains a high probability relative to the other goal states throughout the simulation. Goal 2 brie y takes on a higher probability due to limited number of training trajectories.

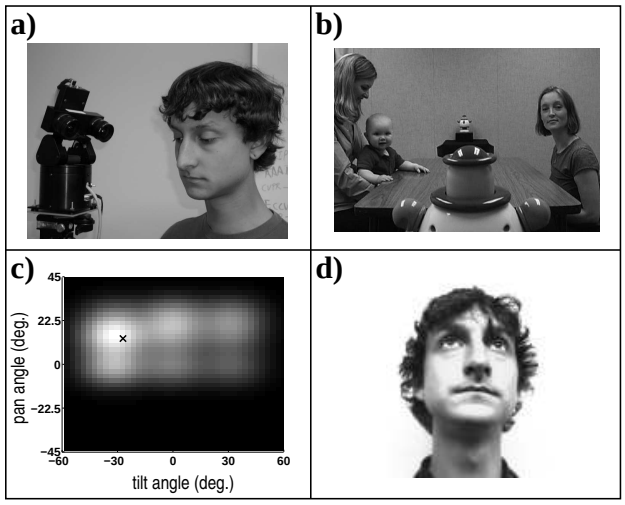


Figure 4: **Gaze tracking in a robotic head:** (a) Biclops active stereo vision head from Metrica, Inc. (b) Infants as young as 9 months can detect gaze based on head direction; older infants ($\geq$ 12 months) use opened eyes as a cue to detect whether they should perform gaze tracking (from [Brooks and Meltzo , 2002]). (c) Likelihood surface for the face shown in (d), depicting the likelihood over pan and tilt angles of the subject's head. The region of highest likelihood (the brightest region) matches the actual pan and tilt angles (black X) of the subject's face shown in (d).

jectories were presented to the algorithm, meaning that its estimates of the distributions on the right hand side of Eqn. 8 were extremely biased.

Real-time application in a robotic head

We have also implemented our probabilistic approach in a Biclops active stereo vision head (Fig. 4(a)). The head follows the gaze of a human instructor, and tracks the orientation of the instructor's head to determine where to look next. Gaze following [Brooks and Meltzoff, 2002, Scassellati, 1999] (Fig. 4(b)) represents a key step in the development of shared attention, in turn bootstrapping more complicated imitation tasks. Our system begins by identifying an image region likely to contain a face (based on detecting skin tones and bounding box aspect ratio). We employ a Bayesian pose detection algorithm [Wu et al., 2000] that matches an elliptical model of the head to the human instructor's face. Our algorithm then transforms the estimated gaze into the Biclops' egocentric coordinate frame, causing the Biclops to look toward the same point in space as the human instructor. We trained the pose detector on a total of 13 faces, with each training subject looking at 36 different targets; each target was associated with a different pan and tilt angle relative to pan 0, tilt 0 (with the subject looking straight ahead).

Fig. 4(c) depicts a likelihood surface over pan and tilt angles of the instructor's head in the pose shown in Fig. 4(d). Our system generates pan and tilt motor commands by selecting the maximum a posteriori estimate of the instructor's pan and tilt, and performing a simple linear transform from instructor-centric to egocentric coordinates. Out of 27 out-of-sample testing images using leave-one-out cross-validation, our system is able to track the angle of the instructor's head to a mean error of $\pm$4.6 degrees. [1] Our previous efforts [Shon et al., 2003] demonstrated the ability of our system to track the gaze of an instructor; ongoing robotics work involves learning policy models specific to each instructor, and inferring instructor intent based on object saliency.

Conclusion

This paper describes a Bayesian framework for imitation learning, based on the AIM model of imitation learning by Meltzoff and Moore. The framework emphasizes imitation as a "match-to-target" task, and promotes separation between the dynamics of the environment and the policy a particular teacher chooses to employ in reaching a goal. We have sketched the basic components for any imitation learning system operating in realistically large-scale environments with stochastic dynamics and noisy sensor observations. Our model naturally leads to a Bayesian algorithm for inferring the intent of other

[1] We de ne error as:
$$\mathcal{E} = \sqrt{\left(\theta_{pan} \quad \hat{\theta}_{pan}\right)^2 + \left(\theta_{tilt} \quad \hat{\theta}_{tilt}\right)^2}$$
where θ is the true angle, and $\hat{\theta}$ is our system's estimate of the angle.

agents. We presented preliminary results of applying our framework to a simulated maze task and to gaze following in an active stereo vision robotic head. We are currently investigating the ability of the framework to scale up to more complex robotic imitation tasks in real-world environments. We are also exploring the connections between our probabilistic framework and findings from developmental psychology.

Acknowledgements

We thank Andy Meltzo for extensive discussions and feedback, and for supplying Figs. 1 and 4(b). APS was funded by a NSF Career grant to RPNR, and DBG was funded by NSF grant 133592 to RPNR. We also thank the anonymous reviewers for their helpful comments.

References

[Billard and Mataric, 2000] Billard, A. and Mataric, M. J. (2000). A biologically inspired robotic model for learning by imitation. In Sierra, C., Gini, M., and Rosenschein, J. S., editors, *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 373–380, Barcelona, Catalonia, Spain. ACM Press.

[Blakemore et al., 1998] Blakemore, S. J., Goodbody, S. J., and Wolpert, D. M. (1998). Predicting the consequences of our own actions: the role of sensorimotor context estimation. *J. Neurosci.*, 18(18):7511–7518.

[Breazeal, 1999] Breazeal, C. (1999). Imitation as social exchange between humans and robots. In *Proc. AISB99*, pages 96–104.

[Brooks and Meltzo , 2002] Brooks, R. and Meltzo , A. (2002). The importance of eyes: How infants interpet adult looking behavior. *Developmental Psychology*, 38:958–966.

[Byrne and Russon, 2003] Byrne, R. W. and Russon, A. E. (2003). Learning by imitation: a hierarchical approach. *Behavioral and Brain Sciences*.

[Efros et al., 2003] Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *ICCV '03*, pages 726–733.

[Fong et al., 2002] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2002). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3–4):142–166.

[Fox et al., 2003] Fox, D., Hightower, J., Liao, L., Schulz, D., and Borriello, G. (2003). Bayesian ltering for location estimation. *IEEE Pervasive Computing*.

[Herrnstein, 1961] Herrnstein, R. J. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement. *J. Exp. Anal. Behaviour*, 4:267–272.

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259.

[Jordan and Rumelhart, 1992] Jordan, M. I. and Rumelhart, D. E. (1992). Forward models: supervised learning with a distal teacher. *Cognitive Science*, 16:307–354.

[Kaelbling et al., 1996] Kaelbling, L. P., Littman, L. M., and Moore, A. W. (1996). Reinforcement learning: A survey. *J. Artificial Intelligence Res.*, 4:237–285.

[Krebs and Kacelnik, 1991] Krebs, J. R. and Kacelnik, A. (1991). Decision making. In Krebs, J. R. and Davies, N. B., editors, *Behavioural Ecology (3rd edition)*, pages 105–137. Blackwell Scienti c Publishers.

[Lungarella and Metta, 2003] Lungarella, M. and Metta, G. (2003). Beyond gazing, pointing, and reaching: a survey of developmental robotics. In *EPIROB '03*, pages 81–89.

[Meltzo , 2002] Meltzo , A. N. (2002). Elements of a developmental theory of imitation. In Meltzo , A. N. and Prinz, W., editors, *The imitative mind: Development, evolution, and brain bases*, pages 19–41. Cambridge: Cambridge University Press.

[Meltzo and Moore, 1977] Meltzo , A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78.

[Meltzo and Moore, 1997] Meltzo , A. N. and Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting*, 6:179–192.

[Nehaniv and Dautenhahn, 2002] Nehaniv, C. and Dautenhahn, K. (2002). The correspondence problem. In *Imitation in Animals and Artifacts*. MIT Press.

[Price, 2003] Price, B. (2003). *Accelerating Reinforcement Learning with Imitation*. PhD thesis, University of British Columbia.

[Rao and Meltzo , 2003] Rao, R. P. N. and Meltzo , A. N. (2003). Imitation learning in infants and robots: Towards probabilistic computational models. In *Proc. AISB*.

[Rizzolatti et al., 2000] Rizzolatti, G., Fogassi, L., and Gallese, V. (2000). Mirror neurons: intentionality detectors? *Int. J. Psychol.*, 35:205.

[Roweis and Saul, 2000] Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

[Scassellati, 1999] Scassellati, B. (1999). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Lecture Notes in Computer Science*, 1562:176–195.

[Schaal et al., 2003] Schaal, S., Ijspeert, A., and Billard, A. (2003). Computational approaches to motor learning by imitation. *Phil. Trans. Royal Soc. London: Series B*, 358:537–547.

[Shon et al., 2003] Shon, A. P., Grimes, D. B., Baker, C. L., and Rao, R. P. N. (2003). Bayesian imitation learning in a robotic head. In *NIPS (demonstration track)*.

[Tenenbaum, 1999] Tenenbaum, J. (1999). Bayesian modeling of human concept learning. In Kearns, M. S., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, Cambridge, MA.

[Tenenbaum et al., 2000] Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

[Todorov and Ghahramani, 2003] Todorov, E. and Ghahramani, Z. (2003). Unsupervised learning of sensory-motor primitives. In *Proc. 25th IEEE EMB*.

[Visalberghy and Fragaszy, 1990] Visalberghy, E. and Fragaszy, D. (1990). Do monkeys ape? In *Language and intelligence in monkeys and apes: comparative developmental perspectives*, pages 247–273.

[Wu et al., 2000] Wu, Y., Toyama, K., and Huang, T. (2000). Wide-range, person- and illumination-insensitive head orientation estimation. In *AFGR00*, pages 183–188.

A Connectionist Model of the Development of Transitivity

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology, McGill University, 1205 Penfield Avenue
Montreal, QC H3A 1B1 Canada

Abbie Vogel (abbie.vogel@mail.mcgill.ca)
Department of Psychology, McGill University, 1205 Penfield Avenue
Montreal, QC H3A 1B1 Canada

Abstract

A modular connectionist model covers all six established phenomena in transitivity development in children and predicts a new effect. In contrast, a symbolic-rule hypothesis based on logic captures none of these effects and is directly contradicted by one of them. In the model a constraint-satisfaction network generates a response based on input from a feed-forward comparison module and the particular question asked. Cycles to saturate the response module implement response times.

Psychology of Transitivity

Piaget and his colleagues (Inhelder & Piaget, 1964; Piaget, 1969) designed the transitivity problem to assess the development of children's logical-inference abilities. This problem often employs sticks (or times) of different length, as shown in Figure 1. Given, for example, that a child learns that stick 2 is longer than stick 1, and that stick 3 is longer than stick 2, can the child infer that stick 3 must be longer than stick 1? This is not a perceptual problem in that the child only identifies a stick by its unique color, never seeing the actual stick lengths. Piaget's evidence suggested that correct untrained inferences, such as comparing sticks 1 and 3, did not emerge until around seven years of age, thus providing an index of the child's entry into the stage of concrete operations.

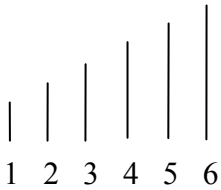


Figure 1: A six-stick version of the transitivity task.

Piaget's long-dominant view of the transitivity task as being solved by logic was ultimately contradicted in an experiment measuring the time it took for people of different ages to make various inferences (Trabasso, Riley, & Wilson, 1975). Using a six-item version of the task like that in Figure 1, Trabasso et al. trained 6-year-olds, 9-year-olds, and university students on all adjacent pairs of sticks and then asked about all possible pairs of sticks, varying the question between *Which stick is longer*? and *Which stick is shorter*? Five different effects were reported.

1. A serial-position effect: learning the adjacent pairs near the ends of the array before the pairs near the middle.
2. A distance effect: faster inferences about pairs that are farther apart in length than for pairs close together in length.
3. An anchor effect: faster inferences about pairs involving an end anchor (sticks 1 or 6) than for pairs not involving an end anchor.
4. A congruity effect: faster inferences when the term used in the question (e.g., longer) is compatible with an end anchor (e.g., the longest stick) in the pair being compared than when the question term (e.g., longer) is incompatible with an end anchor in the pair being compared (e.g., the shortest stick).
5. An age effect: older participants learned the adjacent pairs faster and made inference comparisons faster and more accurately than did younger participants.
6. Other experiments with different comparison tasks found that the distance effect diminished with increasing age (Duncan & McFarland, 1980; Sekuler & Mierkiewicz, 1977).

The first four of these effects have been replicated in a wide range of tasks involving symbolic comparisons along a dimension, e.g., numerical comparisons (Banks, 1977; Duncan & McFarland, 1980; Leth-Steenson & Marley, 2000; Sekuler & Mierkiewicz, 1977).

The distance effect was particularly damaging to Piaget's logical-inference interpretation because it is precisely opposite to what Piaget would presumably predict. Assuming that each inference takes some constant time, Piaget would have to predict that the more inferences required to make a comparison, the longer the comparison would take. For example, comparing sticks 2 and 3 requires no inference at all because participants are trained on such adjacent pairs. In contrast, comparing sticks 2 and 4 requires a single inference from two premises ($S2 < S3$ and $S3 < S4$, therefore $S2 < S4$). And comparing sticks 2 and 5 requires two inferences (the previous inference plus this one: $S2 < S4$ and $S4 < S5$, therefore $S2 < S5$). The larger the split (or difference) between sticks, the more inferences would be required. The splits are conventionally termed 1, 2, and 3 in these three comparisons, respectively.

Because of the distance effect in their response-time data, Trabasso et al. concluded that people don't use logical inference per se to solve this task. They argued instead that participants construct a spatial image of the sticks while being trained on adjacent pairs and then consult this image when asked to make another comparison. The farther the sticks are apart within this spatial image, the easier it is to make a correct comparison. Despite a recent resurgence of interest in studying and modeling transitivity, there has been no computational model that covers all six of these effects. It is, in fact, computationally unclear how the brain might construct and consult spatial images in this way.

The purpose of the present work is to build such a model with cascade-correlation (CC), a neural-network learning algorithm that has been used to simulate many other phenomena in cognitive development (Shultz, 2003). Another reason to use CC is that it searches in topology space, building the network, as well as in weight space.

Like other feed-forward neural algorithms, CC produces responses in more or less constant time, and thus is not naturally suitable for covering response-time effects. To add this capability, we used a modular system of two networks, which we call constraint-satisfaction cascade-correlation (CSCC). A CC network learned to judge the relative lengths of adjacent sticks and a constraint-satisfaction (CS) network used that information plus information contained in the question to generate a response. The number of update cycles that the response module required to settle into a steady state was taken as an index of response time.

Method

The CSCC modular network system is shown schematically in Figure 2.

Comparison Module

The CC comparison module is on the left side of Figure 2. Inputs to the comparison module describe the colors of the two sticks being compared and were coded in a binary *n*th-unit fashion (1 for the color of a stick and 0s elsewhere for colors that are not involved). The 12 inputs (the same 6 colors for each of two sticks) were fully connected to a single output unit having a sigmoid activation function, which coded a length comparison with targets of -0.5 if the left (L) stick is longer and 0.5 if the right (R) stick is longer. Comparison networks were trained on all ten adjacent pairs of sticks until all output values were within score threshold of their targets on all of these ten training pairs. Order of the two sticks being compared is counterbalanced across comparisons.

We implemented age differences by using different values of score threshold: 0.5 for adults, 0.55 for 9-year-olds, and 0.6 for 6-year-olds. This is consistent with finding that older people learn more from the same experiences than young children do (Case, Kurland, & Goldberg, 1982), a principle that has been successfully used to simulate age differences in learning in several other CC simulations (Shultz, 2003). Different score thresholds would also work for capturing developmental effects here, but these particular values produced overall proportions correct that were very close to those reported by Trabasso et al. (1975) for their different age groups. We ran 12 networks at each of the three score-threshold levels, matching the *n*s at each age level in the Trabasso et al. experiment. Full details of the CC algorithm are discussed elsewhere (Shultz, 2003).

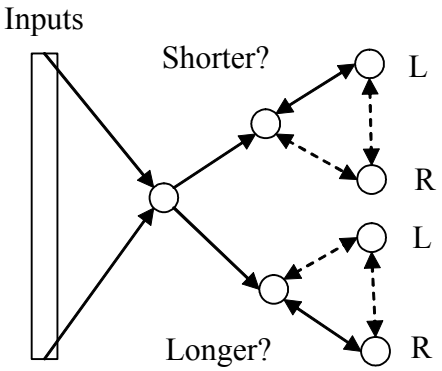


Figure 2: CSCC modular networks for transitivity.

Response Module

After training, output from the comparison module served as input to a three-unit response network. Activation on the other two units in this CS network represented the left or right sticks as being the correct response to the question being asked. As shown on the right side of Figure 2, the precise form of this response network varied according to the question being asked. Recall that the target output of the comparison network was 0.5 when the right stick was longer, and -0.5 when the left stick was longer. Consequently, if the question was *Which stick is longer*?, then there were positive weights (0.5, signified by a solid line) between the comparison unit and the right (R) unit and negative weights (-0.5, signified by a dashed line) between the comparison unit and the left (L) unit. If the question was *Which stick is shorter*?, then the signs of these weights were reversed; there were positive weights between the comparison unit and the left unit and negative weights between the comparison unit and the right unit. The basic principle underlying these weight settings is to enhance the activation value of the side unit corresponding to the stick that is longer when the question term is *longer*, and to enhance the activation value of the side unit corresponding to the stick that is shorter when the question term is *shorter*. More generally the idea is to activate the correct response and inhibit an incorrect response.

Connections between the left and right units were always negative to reflect the idea that these two units are competing with each other. Unlike the comparison unit, these two side units had no external inputs; all of their input came from inside the response network.

As is typical with CS networks, weights in this response module were bidirectional, with one weight going in each direction between any two units. As in other CS simulations (Kunda & Thagard, 1996; Shultz & Lepper, 1996), we assume here that these networks are constructed on the fly

by participants in response to their particular experimental setting and the question being posed to them. There is no assumption that participants are conscious of this construction. It is rather that the design of a response module is strongly constrained by the participant's understanding of the experimental situation and question.

All three units in a response module started out with an initial activation value of 0. At every cycle, three units were randomly selected, with replacement, to have their activations updated. In each such update, net input to the updated unit *i* was computed as:

$$net_i = in\left(\sum_j w_{ij} a_j\right) + ex(input_i) \quad (1)$$

where a_j is the activation of each sending unit *j*, w_{ij} is the relevant connection weight, $input_i$ is any external input to the receiving unit, and *in* and *ex* are parameters scaling influences internal or external, respectively, to the network. These last two parameters were both set to 0.1 in our simulations, but a wide range of values work equally well.

If this net input was positive, it was added to the receiving unit's current activation $a_i(t)$ after being scaled by the distance of that current activation from the activation ceiling of 1.0:

$$a_i(t+1) = a_i(t) + net_i(ceiling - a_i(t)) \quad (2)$$

Alternatively, if the net input was negative, it was added to the receiving unit's current activation $a_i(t)$ after being scaled by the distance of that current activation from the activation floor of -1.0:

$$a_i(t+1) = a_i(t) + net_i(a_i(t) - floor) \quad (3)$$

An overall measure of the degree to which a CS network has settled into a stable state is its *goodness*, computed as the sum of triple products of unit activation values and the relevant connection weight plus the sum of the products of external inputs and activation values:

$$goodness = \sum_{ij} w_{ij} a_i a_j + \sum_i input_i a_i \quad (4)$$

Equations 1-4 are fairly standard in the CS-network literature (Shultz, 2001). In this kind of scheme, goodness rises as units have their activations updated and eventually levels off as activations stop changing. Examples are provided in Figure 3 in terms of goodness changes over update cycles for networks with three different levels of comparison inputs. We identified the cycle at which goodness starts to reach asymptote as no goodness change greater than .02 (asymptote-threshold parameter) for 8 consecutive cycles (asymptote-patience parameter). These parameter values were selected because they correspond to our visual impressions of when goodness values approach asymptote. A range of different threshold and patience values works equally well, although sufficiently extreme parameter values can blur differences between sticks and between conditions. Figure 3 shows that networks settle quicker with higher, and more decisive, comparison activations.

To cover the congruity effect, we multiplied comparison inputs by 0.8 whenever there was an anchor stick that was incompatible with the term used in the question. This is a computational shortcut consistent with the idea that the congruity effect is based on semantic interference between incompatible terms, some of which have to be translated to a compatible form to answer the question (Banks, 1977). When asymptote was reached, the comparison unit, left or right, with the higher activation was taken as the response module's answer to the question that was posed.

Before activation-update cycles began, all the connection weights and external inputs had their initial values randomized a bit by adding or subtracting up to 10% of their initial values in a uniform distribution. This is to reflect the fact that not all participants interpret the experimental procedures and questions in exactly the same way. A wide variety of randomization values work equally well to implement such individual differences.

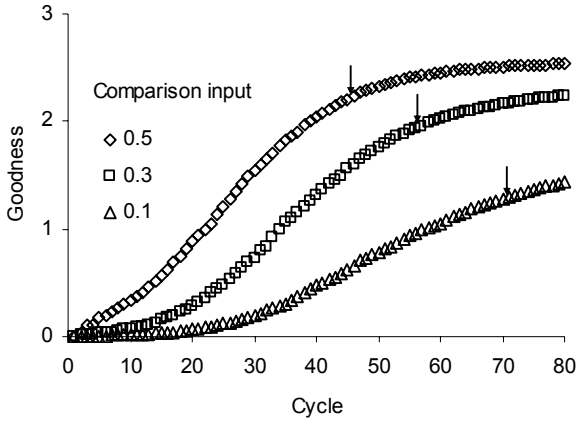


Figure 3: Increasing goodness over activation-update cycles in a CS response network at three levels of comparison input. Arrows indicate the cycle at which a goodness asymptote was reached.

Results

Learning

Although CC networks are capable of recruiting new hidden units if they are needed, none of our networks did so. This indicates that the problem of learning *n*th-unit, binary-coded, adjacent stimulus pairs is a rather simple linearly-separable problem. The mean number of epochs taken to learn the training patterns was 7.2 for score threshold of 0.6, 7.7 for score threshold of 0.55, and 10.2 for score threshold of 0.5.

The serial-position effect for training is evident in Figure 4. A score-threshold x training-pair mixed ANOVA of comparison-network error yielded a quadratic trend for training pair, $F(1, 33) = 279$, $p < .001$. With no interaction with score-threshold, this shows the serial position effect at each age: better learning of training pairs at the ends of the array than in the middle. There was also a main effect of score-threshold, $F(2, 33) = 19.59$, $p < .001$, capturing the superiority of older, deeper learners.

Inference

The theoretically-important distance effect is shown in Figure 5. In a score-threshold x split ANOVA of cycles to settle, the largest effect is a linear trend for split, $F(1, 33) = 582$, $p < .001$, confirming a strong distance effect at every score threshold, representing the three different ages. Network responses were faster the larger the split between the sticks being compared. It is also evident that the distance effect diminished a bit with decreasing score threshold, representing increasing age.

In Figure 6, a score-threshold x end-anchor ANOVA of cycles to settle reveals a main effect of anchor, $F(1, 33) = 166$, $p < .001$, simulating the finding that performance is quicker when an end anchor is present. A score-threshold x end-anchor interaction, $F(2, 33) = 6.73$, $p < .01$, predicts that the anchor effect may also diminish with increasing age.

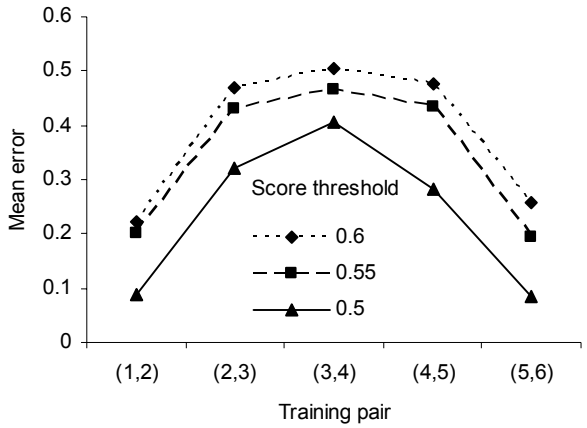


Figure 4: The serial-position effect: mean error for different training pairs and score thresholds.

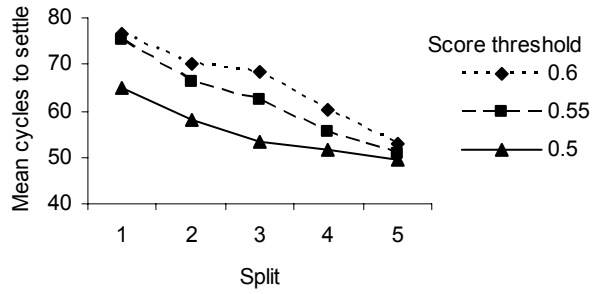


Figure 5: The distance effect: mean cycles to settle for different splits and score thresholds.

The congruity effect is plotted in Figure 7, in the form of an end-anchor x question interaction, $F(1, 33) = 288$, $p < .001$. This shows faster responding when there is compatibility between question and end anchor.

Knowledge-representation Analysis

Mean weights across 12 networks at each score-threshold level are plotted in Figure 8 in order to understand the knowledge representations acquired by the comparison networks. Recall that target output activation is negative when the left stick is longer, and positive when the right stick is longer.

Correct performance by a network can be understood by considering a few example weights. A large positive weight from the R6 input ensures positive output compared to any shorter comparison stick L1 to L5. A somewhat less positive weight from the R5 input produces positive output except when compared to the longer stick L6, in which case the stronger negative weight from the L6 input produces a negative output, signaling that the left stick is longer.

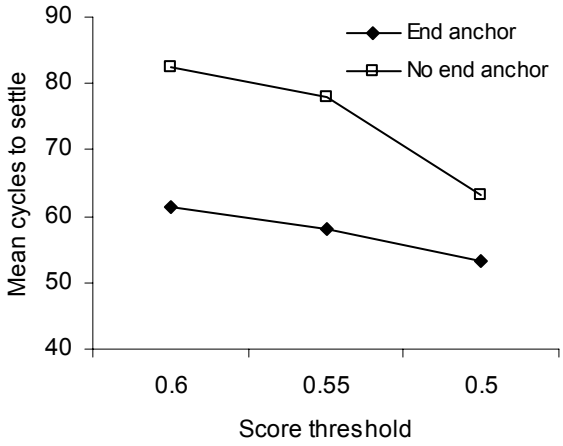


Figure 6: The anchor effect: mean cycles to settle for different score thresholds and the presence of an end anchor.

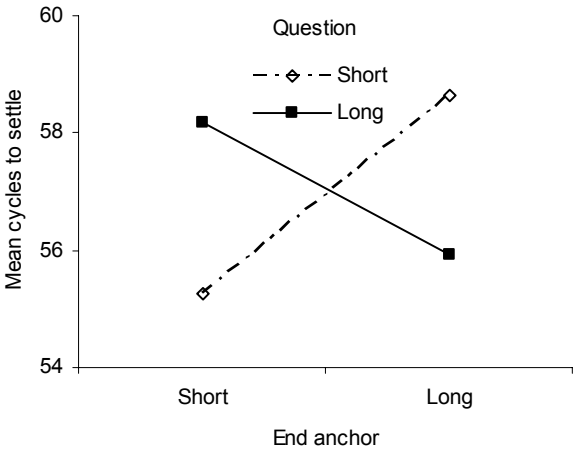


Figure 7. The congruity effect: mean cycles to settle for different size anchors and question phrasing.

The overall pattern of weights is V-shaped, seen most clearly at the lowest score-threshold of 0.5, representing adults. For the right sticks, weights are larger with increasing stick size; for the left sticks, weights are smaller with increasing stick size because the target output is negative when a left stick is longer. Fairly precise left-right

symmetry in weight values on each branch of the V is important to enable accurate judgment of pairs that are close together (e.g., L2 vs. R3).

Serial-position Effect The fact that connection weights have a steeper slope near the ends of the array than in the middle explains the serial-position effect. More distinctive weights produce larger absolute comparison outputs, which are closer to their target values, yielding less network error.

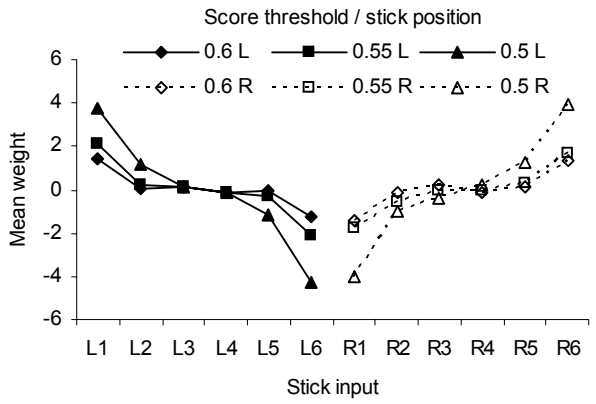


Figure 8: Mean weights in comparison networks from various input units at different score thresholds.

Distance Effect The manner in which this knowledge representation produces the distance effect is evident from the weight plot. Sticks close in size are more likely to have similar weights, producing small absolute comparison values, thus requiring more cycles to reach asymptote in the response module. In contrast, sticks that are farther apart in size are more likely to have larger differences in weight values, producing larger absolute comparison outputs, and thus requiring fewer response-module cycles.

Anchor Effect The manner in which this knowledge representation produces the anchor effect is also evident. Weights for the end anchor sticks (1, 6) have more extreme values than do weights for the other sticks, ensuring larger absolute comparison values and thus quicker responses when end-anchor sticks are involved in a comparison.

Developmental Effects The origin of developmental effects is also apparent from these knowledge representations. The lowest score-threshold of 0.5 (representing adults) produces the steepest V shape with the most easily distinguishable weights. The higher score-thresholds of 0.55 and 0.6 (representing 9- and 6-year-olds, respectively) produce progressively shallower V shapes with weights that are closer in size. The less distinctive the weights, the smaller the absolute output of the comparison module and the more cycles required to reach asymptote in the response module.

Discussion

Our model is a hybrid modular system, with a feed-forward CC network making a length comparison and a CS network using this comparison information along with question information to generate a response. This CSCC model simulated all of the established psychological effects in the development of transitivity in humans. Captured phenomena include the serial-position, distance, anchor, congruity, age-related improvement, and diminishing distance effects.

All of these effects followed naturally from the modular-networks model without any parameter tweaking or special manipulation of training patterns. In general, these effects were produced by the comparison network's natural tendency to learn to order the stimuli by length on its connection weights. The serial-position and anchor effects were due to the fact that these weights were more distinct near the ends of the array than in the middle. The distance effect arose from the fact that the relevant connection weights (those with non-zero inputs) were more distinct with sticks of more distinct size. The congruity effect arose from incompatibility between an anchor and the term used in the questioning, which was made to cause a small degradation of the comparison signal. The age-related improvement and diminishing distance effects were simulated by the familiar phenomenon of older individuals learning the problem more deeply than younger ones do.

Similar interpretations of the serial-position, distance, and anchor effects have been offered by other connectionist modelers (Leth-Steenson & Marley, 2000). But ours is the first model to capture all six effects and to offer novel connectionist interpretations of the congruity and developmental effects. Together these models show how transitivity phenomena can be explained in a neural fashion. We plan to review all of the recent simulations of human and animal data on transitivity in a fuller publication. Different models typically focus on somewhat different phenomena.

Does our model confirm Trabasso et al.'s (1975) hypothesis that people consult a visual image of an ordered spatial array of sticks to answer inference questions? The knowledge representation learned by comparison-module networks certainly does order the array of sticks by length. This learning is based merely on information about the relative lengths of adjacent pairs, without any information on how long the sticks actually are. Whether this knowledge representation constitutes a *visual* image, either in artificial networks or in real brains, is debatable. One way to investigate this issue in real brains might be to see if visual cortex becomes particularly active in brain images of people learning and solving transitivity tasks (Behrmann, Kosslyn, & Jeannerod, 1996). In any case, the simulation presented here demonstrates a fully specified functional account of transitivity development, whether assumed to be located in visual cortex, hippocampus, or other brain regions.

In contrast to these neural-network simulations, Piaget's original logical-inference view cannot account for any of these transitivity phenomena. Indeed its predictions for the effect of distance on response time are precisely the opposite of what actually occurs. Because Piaget's

hypothesis can be naturally framed in terms of a recursively-applied symbolic transitivity rule, this issue can be viewed as another instance of the symbolic rules vs. subsymbolic connections debate that has dominated cognitive science for the past 18 years. As far as psychological development is concerned, results have consistently favored the connectionist approach because it typically covers a wider range of phenomena in a more principled fashion than does the symbolic rule-based approach (Shultz, 2003).

In a more detailed publication, we will present data and analyses of correct inferences by our networks. In general these data also mirror the performance of participants in Trabasso et al.'s (1975) experiment. There is no problem with speed-accuracy tradeoffs in our simulations because response times and errors are positively related.

Although it is beyond the scope of this paper to fully evaluate the various alternative psychological theories of transitivity, these theories do not seem capable of accounting for all the phenomena treated here.

Results indicated that age effects on transitivity tasks can reflect rather small quantitative differences in depth of learning, rather than major qualitative differences in type of processing. Similar results have been found in simulations of a number of other developmental phenomena, including seriation (Mareschal & Shultz, 1999), discrimination shift learning (Sirois & Shultz, 1998), and concept learning (Shultz & Cohen, 2004). Our explanation represents a radical departure from previous interpretations of these phenomena, which have tended to suggest that older children are doing something qualitatively different than are younger children. Because of its capacity for network growth, the CC algorithm is particularly well suited to discovering whether qualitative changes are necessary for capturing developmental change. Some developmental phenomena require such qualitative growth, while other developmental phenomena do not (Shultz, 2003). In addition, our model predicted a diminishing anchor effect with increasing age that could be tested with children.

To capture the heretofore elusive congruity effect we implemented a (shortcut) neural version of the idea that semantic incompatibility between an anchor and question term can slow performance. We also plan to implement an alternative hypothesis based on the notion that particular combinations of question and items serve to bias the participant's response at the start of a random walk towards one or another decision boundary (Link, 1990). Our current response module, with its random selection of units to update, might be adapted to implement the basic features of Link's hypothesis. Comparing results across the two techniques could indicate which hypothesis provides a better explanation.

Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author. We thank J. P. Thivierge, Frédéric Dandurand, and Yuriko Oshima-Takane for helpful comments on an earlier draft.

References

Banks, W. P. (1977). Encoding and processing of symbolic information in comparative judgments. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 11). New York: Academic Press.

Behrmann, M., Kosslyn, S. M., & Jeannerod, M. (Eds.) (1996). *The neuropsychology of mental imagery*. New York: Pergamon.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*, 386-404.

Duncan, E. M., & McFarland, C. E. (1980). Isolating the effects of symbolic distance and semantic congruity in comparative judgments: An additive-factors analysis. *Memory and Cognition, 8*, 612-622.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child: Classification and seriation*. London: Routledge and Kegan Paul.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel constraint-satisfaction theory. *Psychological Review, 103*, 284-308.

Leth-Steenson, C., & Marley, A. A. J. (2000). A model of response time effects in symbolic comparison. *Psychological Review, 107*, 62-100.

Link, S. W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology, 34*, 2-41.

Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science, 11*, 149-186.

Piaget, J. (1969). *The child's conception of time*. London: Routledge and Kegan Paul.

Sekuler, H., & Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development, 48*, 630-633.

Shultz, T. R. (2001). Constraint satisfaction models. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 4). Oxford: Pergamon.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge: MIT Press.

Shultz, T.R., & Cohen, L. B. (2004). Modeling age differences in infant category learning. *Infancy*, *5*, 153-171.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219-240.

Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. J*ournal of Experimental Child Psychology, 71*, 235-274.

Trabasso, T., Riley, C. A., & Wilson, E. G. (1975). The representation of linear order and spatial strategies in reasoning: A developmental study. In R. Falmagne (Ed.), *Psychological studies of logic and its development*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Dissociations Between Regularities and Irregularities in Language Processing: Computational Demonstrations Without Separable Processing Components

Daragh E. Sibley (dsibley@gmu.edu)
Department of Psychology, George Mason University
Fairfax, VA 22030-4444 USA

Christopher T. Kello (ckello@gmu.edu)
Department of Psychology, George Mason University
Fairfax, VA 22030-4444 USA

Abstract

Two models are presented that compute a quasi-regular mapping. One was based on localist representations of items in the quasi-regular domain, the other was based on distributed representations. In each model, a control parameter termed *input gain* was modulated over the one and only level of representation that mapped inputs to outputs. Input gain caused both models to shift between regularity-based and item-based modes of processing. Performance on irregular items was selectively impaired in the regularity-based modes, whereas performance on novel items was selectively impaired in the item-based modes. Thus, each model exhibited a double dissociation without separable processing components. These results are discussed in the context of analogous dissociations found in language domains such as word reading and inflectional morphology.

Introduction

The quasi-regular nature of language has played a central role in theories of language processing in the mind and brain. On the one hand, language processes must be able to handle novel inputs, e.g., skilled readers can give reasonable pronunciations and conjugations to verbs that they have never encountered before. These abilities demonstrate how language usage can be generative on the basis of regularities. On the other hand, irregular items often exist for which the regularities do not apply. Thus, language processes must be able to override the regularities, when appropriate, with knowledge that is applicable to only a few items, or even to just one. How are language processes structured to handle both regularities, and the exceptions to those regularities?

One answer to this question is that any given quasi-regular domain is processed by two complementary routes. A *regularity-based* route is specialized to capture the regularities that span across linguistic items in the domain, and an *item-based* route is specialized to capture knowledge that is specific to items in the domain. For instance, in the words-and-rules theory (Pinker, 1999), rules are used to process regular inflectional morphologys (e.g., WALK-WALKED), and a lexicon is used to process irregular inflections (e.g., GO-WENT). In the dual-route cascaded (DRC) theory of word reading (Coltheart, Curtis, Atkins, & Haller, 93; Coltheart et al., 2001), a set of grapheme-to-phoneme correspondence rules is used to capture regularities between the spellings and sounds of words, and a system of lexical knowledge serves to override the rules when necessary (e.g., PINT does not rhyme with MINT).

Alternatively, *single-route* theories have been proposed in which the mechanisms and representations for handling regularities and irregularities are inseparable. For instance, Rumelhart and McClelland (1986) proposed a theory in which a single route of processing was used to generate the past tense of both regular and irregular verbs (also see, e.g., Joanisse & Seidenberg, 1999). Kello and Plaut (2003) proposed a theory of word reading in which the mapping from spelling to sound is mediated by a single level of learned representations (also see Plaut & Gonnerman, 2000).

A wide variety of evidence has been brought to bear on dual-route and single-route theories of language processing (for reviews, see Coltheart et al., 2001; McClelland & Patterson, 2002; Pinker, 1999; Pinker & Ullman, 2002; Plaut, McClelland, Seidenberg, & Patterson, 1996). Much of this evidence speaks to one or another particularity of a given theory. Every piece of evidence contributes to the overall debate, but here we focus on one kind of evidence that is relevant to all theories in question: dissociations between regularity-based and item-based processing.

Double dissociations have been observed in language processing, and some have been interpreted as evidence for separable regularity-based and item-based components of the language system. In the area of inflectional morphology, Ullman and his colleagues (Ullman et al., 1997) reported evidence for a dissociation between the past tense formation of regular and irregular verbs in English. They found that Alzheimer's patients, as well as aphasics with posterior lesions, were poor at generating the past tense of verbs with irregular inflections, but relatively normal with regular inflections. They found the opposite pattern for Parkinson's patients and aphasics with anterior lesions. Marslen-Wilson and Tyler (1997; 1998) found a similar dissociation in a priming paradigm with language-impaired patients.

In the area of word reading, deficits found in surface and phonological dyslexia have been interpreted analogously to those found in posterior versus anterior aphasics. For instance, Berhmann and Bub (1992) reported on a surface dyslexic patient MP for whom the ability to read exception words (particularly of low frequency) was greatly impaired,

whereas the ability to read both regular words and nonwords was mostly intact. By contrast, Funnell (1983) reported on a phonological dyslexic patient WB for whom the ability to read nonwords (even simple CVC nonwords) was greatly impaired, whereas the ability to read both easy and difficult words was mostly intact.

The impairments of these and other patients have a straightforward explanation in terms of separable item-based and regularity-based processing components. The deficits in Alzheimer's patients, posterior aphasics, and surface dyslexics all reflect damage to an item-based component of processing (e.g., a lexicon) that is responsible for irregular items (not necessarily the same component across types of deficits). The deficits in Parkinson's patients, anterior aphasics, and phonological dyslexics all reflect damage to a regularity-based component of processing (e.g., rules) that is responsible for novel items.

These double dissociations appear to challenge single-route theories because item-based and regularity-based processes are not separable in single-route theories. Proponents of single-route theories have responded to this evidence in a number of ways. In some cases, methodologies or interpretations of data have been called into question (e.g., McClelland & Patterson, 2002). In other cases, the data have been explained in terms of dissociations between semantic and phonological components of processing, rather than item-based and regularity-based components (e.g., Joanisse & Seidenberg, 1999). The research to date has left open the question of whether dissociations between the processing of novel and irregular items can be explained without reference to an architectural dichotomy in the language system.

Current Work

The primary aim of the current study was to demonstrate how a dissociation between item-based and regularity-based processing can occur in a single-route architecture without any manipulation of separable processing components, i.e., without reference to separable semantic and phonological contributions to processing. The basic idea is that a single component of processing can shift between two qualitatively different "modes" of processing as a function of one control parameter. Specifically, we present two different kinds of connectionist models that possess a control parameter termed *input gain*. We show that, in both types of models, input gain can cause a shift in processing between an item-based mode and a regularity-based mode. Furthermore, we show how this shift can give rise to a double dissociation in performance on irregular versus novel inputs.

The models were built to process an abstract, quasi-regular mapping. Properties of the mapping were analogous to basic properties of quasi-regularity in language domains. However, items did not correspond to any particular words in a particular language domain. The mapping was created primarily to facilitate analysis of the models, rather than to simulate a particular language phenomenon such as the past tense formation in English. Therefore, the models are intended and reported only as proofs-of-concept.

The first model used a single level of localist nodes to map input patterns onto output patterns. Each node represented one item in the training corpus, and the activation of each node was a function of the similarity between the item it represented, and the current input to the model. Thus, this model could be considered as *analogy-based* because both known and novel inputs were explicitly processed in terms of the similarity of their input patterns to that of all items in the corpus (see Albright & Hayes, 2003; Nakisa, Plunkett, & Hahn, 2000).

The second model used a *distributed* level of representation to map input patterns onto output patterns. Hidden representations were learned via backpropagation (Rumelhart, Hinton, & Williams, 1985), and each hidden unit contributed to the processing of many, if not all, items in the training corpus. Representations learned through backpropagation tend to map similar inputs onto similar outputs (Rumelhart et al., 1995). Thus, as in the analogy model, the distributed model processed both known and novel inputs in terms of their similarity to items in the corpus. But unlike the analogy model, hidden representations were shaped by similarities among both input and output patterns in the corpus, as well as the relationships between inputs and outputs.

In both models, input gain is a multiplicative scaling parameter on the net inputs to units, be they localist nodes or hidden units. The current simulation results show that the modulation of input gain at testing caused similar effects in both models. At low levels of input gain, both models failed to map irregular items to their appropriate outputs, but succeeded in mapping regular items and novel inputs. At high levels of input gain, both models succeeded at mapping both regular and irregular items, but performed poorly with novel inputs.

The reason why input gain caused this double dissociation was different for each model. In the analogy model, input gain modulated the intensity of competition for activation among localist nodes. Low levels of competition caused outputs to be based on the summed contributions from many partially activated nodes. Regularities across nodes were extracted in these summations to the point of overriding any exceptions to the regularities. By contrast, high levels of competition caused a winner-take-all mode of processing in which a known input correctly activated its corresponding node, whereas a novel input incorrectly activated a node corresponding to a similar, known item.

In the distributed model, input gain modulated the sharpness of a sigmoidal activation function. Low levels of input gain caused hidden units to operate mostly in their linear range, thereby emphasizing the componential (i.e., regular) relationships that were learned between inputs and outputs. High levels of input gain caused hidden units to operate mostly in their asymptotic range, thereby emphasizing the conjunctive relationships that were learned between inputs and outputs (for a discussion of

componential and conjunctive coding, see O'Reilly, 2001). Componential relationships supported only the processing of regular and novel items, whereas conjunctive relationships supported only the processing of known items.

Simulation Methods

Input and Output Representations were constructed from a 12 dimensional binary space. Out of $2^{12} = 4096$ possible input patterns, one fourth (1024) were chosen at random to constitute the corpus of items. Each chosen input pattern was associated with one output pattern. Output patterns were created in two steps. First, each input pattern was copied to its corresponding output pattern (i.e., the identity mapping. Note, however, that the results apply to all linearly separable mappings). Second, the bit value of each dimension, for each output pattern, was flipped with a 5% probability. Thus, the identity mapping was a regularity, and flipped values were exceptions to that regularity. This procedure resulted in 563 fully regular items (no flipped bits), and 461 irregular items with one to four flipped bits per item. The 3072 remaining patterns served as novel items during testing.

For the analogy model, there were 12 input units corresponding to the 12 input dimensions, and dimension values were coded as activations of ±1 on the inputs. For the distributed model, there were 24 input units, half of which coded the 12 dimension values as activations of 0 or 1. The other half were activated as flipped values of the first half, i.e., $1–x$, where x was each of the first 12 activations. The $x|1–x$ coding scheme was used because the distributed model was trained via backpropagation (this scheme was not necessary in the analogy model because it was not trained; see next two sections). In backpropagation, no learning will occur on a unit's sending weights when the activation value of that unit is zero. Therefore, the $x|1–x$ coding scheme ensured that weight derivatives were generated for every input dimension, on every training episode.

For both models, there were 12 output units corresponding to the 12 output dimensions, and dimension values were coded as targets of 0 or 1 on the outputs.

Analogy Model Architecture. In the analogy model, input units were fully connected to 1024 "logogen" units. Each logogen represented one item in the corpus, and the weights on incoming connections from input units were set according to each logogen's input pattern, i.e., +1 weights for positive input dimensions, and -1 weights for negative dimensions. Each logogen projected outgoing connections to all 12 output units, and the weights on outgoing connections were set according to each logogen's output pattern (as for incoming connections).

To process a given item, input units were first set to the item's input pattern. Logogen activations were then calculated with the normalized exponential function (see Nosofsky, 1990),

$$a_j = e^{\gamma \varepsilon I_j} \Big/ \sum_i e^{\gamma \varepsilon I_i},$$

where I was the net input to a unit, calculated as the dot product between the input vector and the incoming weight vector, γ was input gain, ε was noise sampled evenly in the range ±0.1, and i spanned all logogens. Each output unit was then calculated as the sigmoid of the dot product between the logogen vector and its incoming weight vector. Noise was included to break perfect ties between very small (e.g., two or three) numbers of activated logogens. Such ties occurred more often at high levels of input gain.

Distributed Model Architecture. In the distributed model, the input units were fully connected to 200 hidden units, and the hidden units were fully connected to the output units The number of hidden units was determined through pilot testing to be about 50 units more than the minimum needed to learn the mapping. However, results were very similar over a range of hidden unit numbers. Hidden units were calculated with the hyperbolic tangent function,

$$a_j = \tanh\left(\gamma \varepsilon I_j\right),$$

which is analogous to the logistic, except it has asymptotes at ±1 instead of 0 and 1. Input gain (γ) was fixed at 1 during training, and varied during testing (see next section). Noise (ε) was fixed at 0.1 (as in the analogy model) during both training and testing. Output units were calculated as in the analogy model.

Connection weights were initialized to random values in the range ±0.1, and weights were learned by gradient descent,

$$\Delta w_{ij} = \eta\left(\partial E / \partial w_{ij}\right),$$

where w_{ij} was the connection weight from unit j to i, η was the learning rate (fixed at 0.001), and E was cross-entropy error (Rumelhart et al., 1995). Weight changes were made each time after weight derivatives had been accumulated over all 1024 items in the corpus. Weight derivatives were calculated for each item as follows: input units were set to the item's input pattern, activation was propagated forward through the network, an error signal was calculated from the difference between actual and target outputs, and the error signal was backpropagated to generate the weight derivatives. Weight updates were repeated until every output unit was with 0.1 of its target for every item in the training corpus. This criterion was reached after 3000 passes through the corpus.

Testing Procedure. For both models, performance was assessed on each test item by setting the input units to the item's input pattern, and then determining whether the activation of each output unit was within 0.5 of its target (which was either 0 or 1). Model outputs were correct only when the activations of all 12 output units were within range. Targets for items in the corpus were set according to each item's output pattern. Targets for the 3072 novel items were set according to each item's input pattern, i.e., the identity mapping.

To dissociate item-based and regularity-based processing, input gain was varied as a single control parameter over the logogen units in the analogy model and over the hidden units in the distributed model. The reported levels of input gain were between 0.5 and 3 for the analogy model, and 0.333 and 3 for the distributed model. These ranges were chosen to show asymptotic performance at the lower and upper ends, i.e., the patterns of behavior did not change substantially beyond these ranges.

Simulation Results

Mean accuracies for the analogy model are graphed in Figure 1 as a function of input gain and item type (regular, irregular, or novel). The same are graphed for the distributed model in Figure 2.

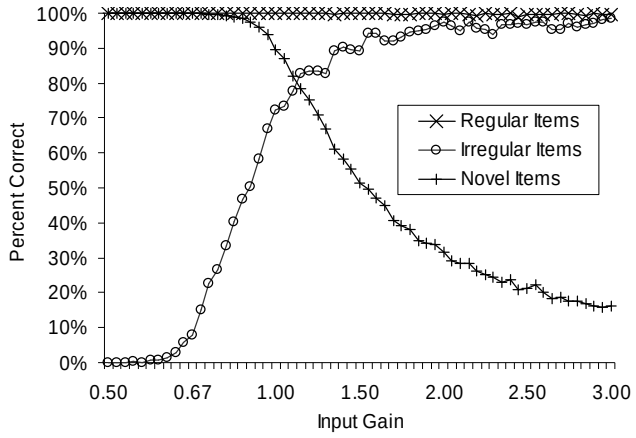


Figure 1: Mean accuracies for the analogy model

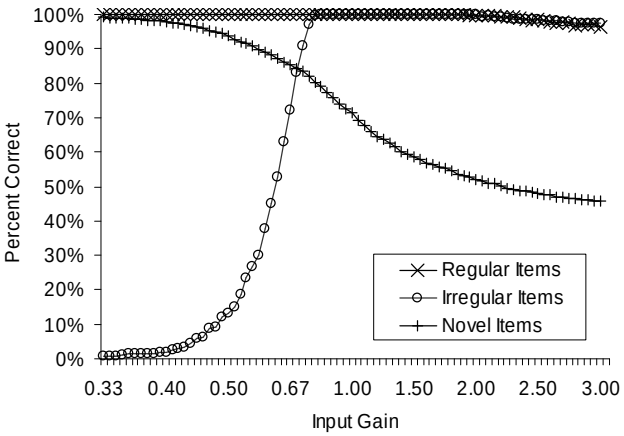


Figure 2: Mean accuracies for the distributed model

Figures 1 and 2 show that both models exhibited a clear dissociation in performance on irregular items compared with novel items. At low levels of input gain, generalization of the identity mapping to novel inputs was essentially perfect, as was performance on regular items. By contrast, performance on irregular items dropped to 0%, at which point all inputs resulted in the identity mapping. For irregular items, application of the identity mapping can be considered as a *regularization* error because, for the quasi-regular domain constructed here, the identity mapping is the regular mapping.

At high levels of input gain, performance on all items in the corpus was near perfect in both models. By contrast, mean accuracies for the novel items dropped to as low as 16% for the analogy model, and 46% for the distributed model. Of all the analogy model's erroneous responses to novel items at the highest level of input gain, 97% were output patterns that corresponded to output patterns in the training corpus. These responses can be considered as *lexicalization* errors because they are responses for other items in the model's "lexicon". The same analysis of errors made by the distributed model showed only 27% lexicalization errors (where the chance rate was 25%).

These results show that the manipulation of input gain as a single control parameter, over a single level of representation, caused a clear double dissociation in both models. To better understand the similarities and differences in processing between these models, three visualizations of the input-output mappings for each model are shown in Figure 3.

In each visualization, all 4096 points in the 12 dimensional input space are arranged on a grid such that all adjacent vertices differ by only one bit. To illustrate, near the lower left-hand corner of each plot is the vertex where all 12 input dimensions are negative. The next vertex up and the next vertex to the right each have one positive input dimension, and so on. Each grid "wraps around" such that vertices on the left edge are adjacent to the corresponding vertices on the right edge, and likewise for the top and bottom edges. Thus, the 2D space of each grid represents a portion of the similarity structure in the 12D input space. In addition, 10 evenly spaced points are interpolated in each space between each pair of vertices. Given that each side has 64 vertices ($64^2 = 4096$), there are $640^2 = 409{,}600$ points of the input space represented in each plot.

At each point, a gray scale value is plotted that represents the summed activation of four output units for the corresponding input pattern. The same four output units (chosen arbitrarily) are shown at all points in all plots. The gray scale values are calculated such that, the darker the point, the closer the outputs were to 0.5. Conversely, whiter points indicate where the outputs were at their asymptotes (0 or 1). Thus, the dark borders in each plot represent the decision boundaries in each model, that is, where one or more of the four outputs crossed the middle point between asymptotes as a function of change in the input space.

Plots are shown for each model, at three different levels of input gain: the low end (0.5 in the analogy model and 0.333 in the distributed model; top row), the high end (3 in both models; bottom row), and the point at which accuracies for irregular items and novel items are equal (1.1 in the analogy model and 0.8 in the distributed model; middle row). Overall differences in plot densities for the analogy model, compared with plot densities for the distributed model, were due to differences in the polarity of the output

units: outputs in the distributed model tended to be closer to 0 or 1, i.e., values that corresponded to white points on the plots.

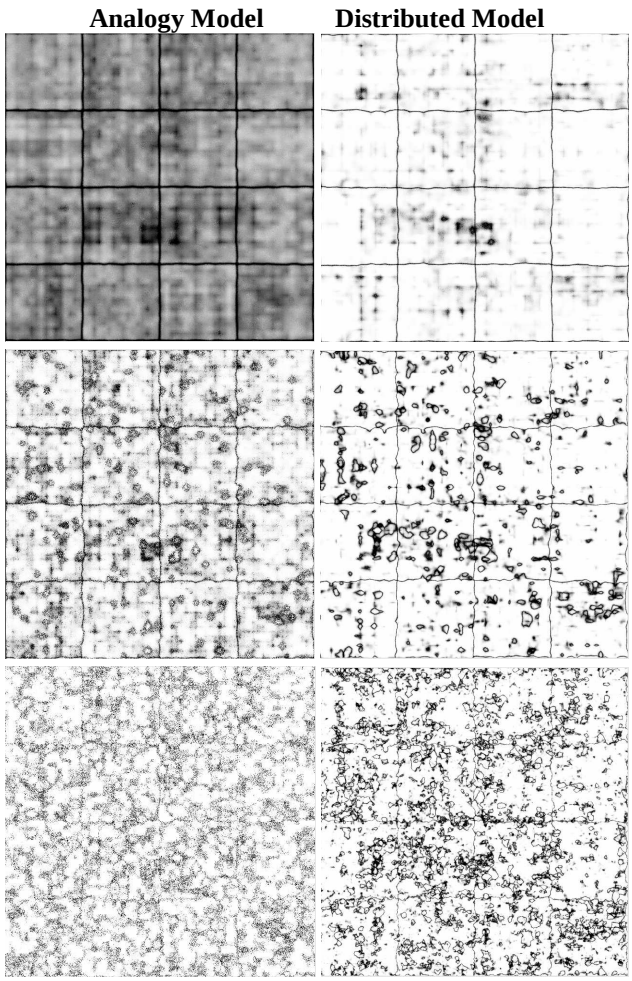


Figure 3. Visualizations for each model at low (top), medium (middle), and high (bottom) levels of input gain

The grid patterns seen in the top two plots of Figure 3 show that both models processed the identity mapping at the low end of input gain. In fact, if all 12 outputs had been represented, each plot would show a 64 by 64 grid pattern, where the grid lines fall exactly between the vertices. Thus, the grid reflects the finding that, at low input gain, the identity mapping was generalized to all inputs, including those for novel and irregular items. The grid is a depiction of regularity-based processing in each model because the identity mapping was the regularity in our quasi-regular domain.

The middle two plots show that the grid pattern became distorted for both models at moderate levels of input gain, and "pockets" of decision boundaries began to appear. Given that mean accuracies were about 80% for irregular items at these levels of input gain, one can infer that the distortions and pockets reflect the "warping" of the identity mapping that was necessary to process the irregular items. Moreover, given that mean accuracies were about 80% for novel items as well, one can infer that these distortions and pockets were mostly isolated to the irregular items. These plots show that a balance was struck at moderate levels of input gain between item-based and regularity-based processing.

The bottom two plots show that, for each model, the grid pattern was mostly replaced by pockets of decision boundaries at the high end of input gain. These pockets have a fairly simple interpretation for the analogy model. Recall that, at the high end of input gain, 97% of the errors for novel items were lexicalizations. What this means is that the pockets show where known inputs were mapped correctly, and where novel items were mapped incorrectly to similar known items. These "item pockets" are a depiction of item-based processing in the analogy model.

In the distributed model, the pockets cannot be readily interpreted as item pockets because a substantial number of novel items were mapped correctly at the high end of input gain (46%), and the proportion of lexicalization errors for novel items was not much above chance (27%). It appears that the distortions needed for accurate mappings of irregular items had "spread out" at high levels of input gain. Because the mapping of regular items is mostly correct at the high end of input gain, one can infer that the decision boundaries spread out over untrained (novel) regions of the space more than they did over trained (known) regions. It is this selective spread of decision boundaries that indicates item-based processing at the high end of input gain.

Conclusions

The current simulations provide a new demonstration of how double dissociations can occur without separable processing components (see also Devlin & Gonnerman, 1998; Juola, 2000). Performance on novel versus irregular stimuli was dissociated by shifting between regularity-based and item-based modes of processing. Unlike previous demonstrations, these modes existed at the ends of a continuum created by one control parameter.

It is important to acknowledge that the current work only opens the door to an alternative to the rules/lexicon and phonology/semantics explanations of double dissociations. It is unclear whether input gain would provide a satisfying account of specific empirical results. For instance, input gain would not appear to handle dissociations in which all regular items, both novel and known, are impaired (Marslen-Wilson & Tyler, 1997, 1998; Ullman et al., 1997). Also, the current simulations did not include subregularities or variations in the frequency of items. These factors have been simulated successfully (Kello, Sibley, & Plaut, submitted), but only as demonstrations. Subregularities allowed for model errors that were more like patient errors, but further work is necessary to test the simulated errors.

The current simulations also raise a number of larger questions, such as: Are there any testable differences between the analogy and distributed models presented here? Do these simulation results have implications for current

theories of word reading and inflectional morphology? Are the reported models consistent with the localization of regularity-based and item-based processing in the brain, to the extent that evidence exists for such localization? What might be the neural bases of input gain? These and other questions await further research.

Acknowledgments

This work was funded in part by NIH Grant MH55628, and NSF Grant 0239595. The computational simulations were run using the Lens network simulator (version 2.6), written by Doug Rohde (http://tedlab.mit.edu/~dr/Lens). We thank David Plaut for his input on precursors to this work.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition, 90*, 119-161.

Behrmann, M., & Bub, D. (1992). Surface dyslexia and dysgraphia: Dual routes, single lexicon. *Cognitive Neuropsychology, 9*, 209-251.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*, 589-608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204-256.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience, 10*, 77-94.

Funnell, E. (1983). Phonological processes in reading: New evidence from acquired dyslexia. *British Journal of Psychology, 74*, 159-180.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading, and dyslexia: Insights from connectionist models. *Psychological Review, 163*, 491-528.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: a connectionist model. *Proceedings of the National Academy of Sciences of the United States of America, 96*, 7592-7597.

Juola, P. (2000). Double dissociations and neurophysio-logical expectations. *Brain & Cognition, 43*, 257-262.

Kello, C. T. (2003). The emergence of a double dissociation in the modulation of a single control parameter in a nonlinear dynamical system. *Cortex, 39*, 132-134.

Kello, C. T. & Plaut, D. C. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language, 48*, 207-232.

Kello, C.T., Sibley, D.E., & Plaut, D.C. (submitted). Dissociations in performance on novel versus irregular items: Single-route demonstrations with input gain in localist and distributed models. *Manuscript under review.*

Marslen-Wilson, W.D. & Tyler, L.K. (1997). Dissociating types of mental computation. *Nature, 387*, 592-594.

Marslen-Wilson, W.D. & Tyler, L.K. (1998). Rules, representations and the English past tense. *Trends in Cognitive Science, 2*, 428-436.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6*, 465-472.

Nakisa, R., Plunkett, K. & Hahn, U. (2000). Single- and dual-route models of inflectional morphology. In P. Broeder & J. Murre (Eds.), Models of language acquisition: Inductive and deductive approaches. New York: Oxford University Press.

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation, 13*, 1199-1241.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*, 291-321.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Plaut, D. C. & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes, 15*, 445-485.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. New York: Basic Books.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*, 456-463.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In C. Yves & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp 1-34). Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

Rumelhart, D. E., McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart, McClelland, J. L., and The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216-271). Cambridge, MA: MIT Press.

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., & et al. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience, 9*, 266-276.

Van Orden, G.C., Pennington, B.F., & Stone, G.O. (2001). What do double dissociations prove? *Cognitive Science, 25*, 111-172.

Fodor's 'Guilty Passions': Representation as Hume's Ideas.

Peter Slezak (p.slezak@unsw.edu.au)
Program in Cognitive Science, School of History & Philosophy of Science
University of New South Wales, NSW 2052 AUSTRALIA

Abstract

Jerry Fodor (1985) has joked that philosophers have always been prone to eccentric worries such as an anxiety about the existence of tables and chairs, but with the issue of mental representation they have found a problem that is real and crucial for progress in the cognitive sciences. However, given Fodor's 'methodological solipsism' of computational symbols and their 'formality condition', Jackendoff (1992) has facetiously asked "Why, if our understanding has no direct access to the real world, aren't we always bumping into things?" It is no accident that Jackendoff's parody recalls Samuel Johnson's famous retort to Berkeley's "ingenious sophistry" by kicking a stone. There is an acute irony in the fact that cognitive science has simply rediscovered the philosophers' traditional worry about tables and chairs. Accordingly, it is not surprising that Fodor's latest book *Hume Variations* endorses the classical Empiricist 'idea' idea of Locke, Berkeley and Hume. The paper explores Fodor's concept of ideas as mental objects in relation to its historical antecedents.

Precursors

For some time, Fodor (1978, 1998) has been making hints *en passant* comparing his favored theory of mind with that of early modern Empiricist philosophers. For example, in his *Concepts*, he said "To a first approximation ... the idea that there are mental representations is the idea that there are Ideas *minus* the idea that Ideas are images." "Hume taught that mental states are relations to mental representations, and so too does RTM" (Fodor 1998, p. 8,9). In this light, it is hardly surprising that modern problems might be simply the reinvention of old problems in a new guise. Now, with his *Hume Variations* (2003), Fodor has come out of the closet, admitting to having harbored something like a "guilty passion" for Hume's *Treatise of Human Nature*. Fodor's enthusiasm for Hume is based on his view that Hume's account of the mind "seems, in a number of respects, to anticipate the one that informs current work in cognitive science" (p. 2). Indeed, Fodor suggests "Hume's *Treatise* is the foundational document of cognitive science: it made explicit, for the first time, the project of constructing an empirical psychology on the basis of a representational theory of the mind; in effect, on the basis of the Theory of Ideas" (p. 134). More specifically, Fodor says, "it remains fully plausible that cognitive processes are constituted by causal interactions among mental representations, that is, among semantically evaluable mental particulars" (p. 135). Translated, this means an 'atomistic' account of concepts, of which he says, "To be sure, on this view, we're not after all so far from billiard balls" (p. 137). Fodor adds, "Either that, or we really are entirely in the dark". Indeed, there are ample grounds to wonder about both the degree of current illumination and also Fodor's history.

Independently of his earlier *obiter dicta*, Fodor's formulations have always been evocative of traditional accounts of 'ideas' as the 'direct objects' of perception and understanding. This is the compelling conception according to which we don't perceive the objects of the world directly, but only indirectly as mediated by our mental representations or ideas of them. Fodor's analysis of 'propositional attitudes' as "relations between organisms and internal representations" has always suggested a tri-partite, 'object' conception of concepts common to traditional and contemporary representative theories (Bechtel 1998, Slezak 2002):

$$\text{world} \leftrightarrow \text{representation} \leftrightarrow \text{mind}$$

Fodor protests that his view is a minority opinion but the problematic tripartite structure has always been dominant (see von Eckardt 1993, Slezak 2002), despite the periodic complaints of 'pragmatists', and 'direct perception' advocates such as Hume's critic Thomas Reid and his more recent counterparts such as Hilary Putnam (2000).

Fodor suggests that Hume's account not only anticipates current work in cognitive science but "thinking seriously about our theory of mind in relation to Hume's might help with the project" (p.2). Thus, Fodor's small book may be seen as an extended historical footnote or appendix to earlier work, particularly his *Concepts* (1998). By 'outing' himself as adherent of the classical 'idea' idea, Fodor illustrates an important approach to theorizing in cognitive science.

With some important differences, Fodor's enterprise shares its approach and purpose with Chomsky's (1966) neglected *Cartesian Linguistics* which sought to understand the body of theoretical insight of the premodern period, to appraise their contemporary relevance and to find ways to exploit them for advancing contemporary inquiry. For his part, Chomsky offered no explicit analysis of the relation of Cartesian linguistics to current work on the grounds that the modern reader should have little difficulty in drawing these connections for himself (1966, p. 2). Chomsky was undoubtedly too optimistic in this regard, perhaps contributing to the neglect of this important contribution to both classical scholarship and contemporary cognitive science. By contrast, Fodor's book is weighted in the opposite direction with primary focus upon current theories of the mind.

As Fodor points out, Hume has suffered from a procrustean hindsight according to which most of what he

took to be important about his own philosophy has been dismissed as not philosophy at all, but empirical psychology (p. 5). As Fodor notes wryly, "Mastering the science of human nature doesn't sound a lot like analyzing concepts" (p. 5). The changed philosophical climate today has more or less effaced the distinction between philosophical inquiry and science, thereby permitting us to see Hume in a clearer light.

Of course, Fodor chooses Hume for this exercise because Hume "holds a fairly rudimentary and straightforward version of the sort of cognitive psychology that interests me" (p. 2). That is, Fodor has a partisan rather than purely exegetical purpose - namely, to use Hume as a vehicle for advertising the virtues of his own theory of mental representation. As far as it goes, this is an important and interesting exercise - not least, because the parallels and divergences help us to get a clearer picture of Fodor's own significant position on issues central to cognitive science today.

Fodor's book reveals something of the mutually illuminating connections between the disjoint literatures of cognitive science and the history of early modern philosophy. However, Fodor is not vindicated simply because he was anticipated by Hume. The 'Whig' approach to history cuts both ways, and Hume's neglected critic Reid derives a renewed interest precisely because of Fodor's project and its partisanship. Reid gets short shrift from Fodor, relegated with other 'pragmatists', direct-realists and Wittgensteinians to dismissive footnotes. However, a *Reid Variations* would tell a more compelling alternative story than Fodor allows.

As Fodor notes, Hume's representative Theory of Ideas (TOI) was itself derived from Descartes. Consequently, throughout the book Fodor refers to the doctrine he defends as 'Cartesian' though in this case the adjective is intended to modify 'representationalism' and not the more usual 'dualism'. However, Hume's conception of this representationalism was, in fact, closer to Malebranche's version than Descartes's own. Though a follower of Descartes, Malebranche held a distinctive and highly problematic conception of ideas as objects in the mind of God. Indeed, Descartes shared the 'pragmatism' and 'direct realism' of Malebranche's critic Arnauld and later Reid - the very doctrine that Fodor combats as "a main concern throughout this book" (p. 12). Fodor characterizes this pragmatist doctrine as "the defining catastrophe of analytic philosophy of language and philosophy of mind in the last half of the twentieth century" (p. 73,4). Accordingly, insisting on such issues of provenance and tracing the genealogy of ideas is no mere antiquarian pedantry or exegetical nicety. If we take Fodor's enterprise seriously, on his own account and example, the historical parallels can be very instructive about our current theoretical problems. In particular, the celebrated Malebranche-Arnauld debate and its subsequent re-enactments were anticipations of Fodor's polemic with his critics today. By focusing on Hume alone, Fodor obscures this broader picture, but the pattern of recurrence is a striking fact whose significance deserves to be understood.

Curious & Melancholy Fact?

Despite Fodor's unfailing optimism, there is a kind of recurrence which deserves attention because it is a manifestation of deeper, and therefore specially illuminating, causes, - a chronic malaise that is symptomatic of deep pathology. Thus, Yolton (1984, p. 6) has noted that the burning question among philosophers in the seventeenth and eighteenth centuries is that of "objects present to the mind" - precisely Fodor's question of "concept possession" central to the recent book and his earlier *Concepts* (1998). Fodor suggests that in the intervening period since Hume the theory of ideas "seems to have made some modest progress" while acknowledging that it is, "to be sure, more modest than some have advertised" (p. 157). Although relegated to footnotes in Fodor's book, modern counterparts of Hume's critics are also prominent among leading theorists today. This looks alot less like progress, however modest, than Fodor suggests. Thus, Putnam (2000) has recently defended Reid's "natural realism" and he also cites John Austin who, significantly, invokes yet earlier writers saying: "It is a curious and in some ways rather melancholy fact that the relative positions of Price and Ayer at this point turn out to be exactly the same as the relative positions of Locke and Berkeley, or Hume and Kant." (Austin 1962, p. 61)

In a riposte to Putnam (2000), Fodor (2000) asserts bluntly, "In fact, there is no direct realist theory of perception (or of anything else that's mental)". However, Fodor ignores Putnam's concern about representations as "interface" between mind and world, though this has been the classical source of discomfort about the 'veil of ideas' central to the Humean conception. Thus, Putnam and Reid are grouped with Gibson and McDowell as among those who "reject RTM entirely" in the sense that they hold "perception isn't mediated by mental representations." Fodor adds that this flies in the face of the evidence from the success of modern psychology. Referring specifically to Reid, Fodor suggests "but for the notion of mental representation, much of what the mind does would be miraculous. The miracle theory of mind is the natural alternative to the representational theory of mind" (Fodor 2000). Fodor's extravagant humour makes it hard to tell whether he is just exaggerating for effect or plain wrong in this characterization of the Reid/Putnam view. Of course, there have been accounts purporting to deny representations altogether (Brooks 1991, Freeman and Skarda 1990, Clark and Toribio 1994, Greeno 1989, van Gelder, 1998). However, even these views are not plausibly seen as a return to something like behaviorism since, strictly speaking, they do not reject internal representations at all (see Markman and Dietrich 2000). For his part, Putnam makes the point explicitly, seeking "to distinguish carefully between the activity of "representation" (as something in which we engage) and the idea of a "representation" as an *interface* between ourselves and what we think about, and to

understand that giving up the idea of representations as interfaces requiring a "semantics" is not the same thing as giving up on the whole idea of representation" (Putnam 2000, p. 59). Thus, Fodor's appears to miss the more subtle views (not to mention explicit texts) which do not deny representations as such but only a certain notoriously problematic conception of them as mental objects with truth values. Futhermore, as Greco (1995) notes, "it is clear that Reid does not deny the existence of ideas if ideas are thought of as operations or acts of thought. Rather, Reid objects only to ideas as mental entities distinct from any operation or act" (1995, p. 283).

Jerry-mandering?

Fodor reproaches Putnam and cites Marr's (1982) research as exemplary proof of the representational pudding. However, Marr's differential equations, zero crossings and other formalisms are not obviously the widely individuated, semantically evaluable mental particulars that Fodor takes to be characteristic of representations. Marr is undoubtedly a counter-example to any theory that would deny representations altogether, but fails to address the concern with 'object' theories of Hume and Fodor. Thus, Fodor's animadversions against miraculous theories are beside the point since the "direct" theories of interest are also "indirect" in Fodor's uncontroversial sense, namely, in positing some internal, causal processes which are responsible for, and in this sense 'mediate', perception, belief and action. This way of putting the point will be agreed on all sides. Neither Putnam nor Reid would demur from this way of 'jerry-mandering' the issue since everyone is an indirect representationalist in this sense.

Granny & the Golden Mountain

Fodor's hard-core *Malebranchisme* is further confirmed and illuminated by his un-selfconscious use of the most venerable argument for ideas as mediating objects of perception: Fodor asks how he could think about his Granny if he is in New York and she is in Ohio. Or, "How can I be in an unmediated relation to Ebbets Field (alas long since demolished); or to my erstwhile dentist, who passed away a year ago in August?" (Fodor 2000). This is, of course, just the notorious Argument from Illusion, and the rhetorical force of Fodor's question relies upon the remoteness or non-existence of things we are supposed to be in a problematic "direct" relation to. Malebranche, too, remarks "it often happens that we perceive things that do not exist, and that even have never existed - thus our mind often has real ideas of things that have never existed. When, for example, a man imagines a golden mountain, it is absolutely necessary that the idea of this mountain really be present to his mind" (1712, p. 217). The classical conclusion, of course, is that we must be in a direct relation with something else - namely, an image, sense datum or 'idea'. However, the "directness" of veridical perception (or memory) is not so easily defeated in this manner, since it need not rely on some occult relation to its objects as Fodor suggests. Of course, it is not obvious that Fodor's causal theory is any better able to deal with distant or non-existent objects of thought, as Putnam has pointed out. Conceptual or inferential role theories offer an alternative conception in the spirit of Arnauld and Reid. Moreover, as just noted, both direct and indirect theories of perception are equally committed to causal intermediaries which will explain Fodor's relation to Ebbets Field, his granny and his late dentist.

Fodor's deployment of what is in effect the Argument from Illusion suggests that he may be open to the kind of charge Putnam makes against Dummett, namely, "that his picture ... is closer to the 'cognitive science' version of the Cartesian cum materialist picture than he himself may realize" (Putnam 2000, p. 58). By this Putnam means the 'Cartesian Theatre' conception minus dualism that Dennett, too, has characterized as 'Cartesian Materialism'. Despite its centrality in the tradition of ideas, Fodor nowhere attempts to escape or even address this potential difficulty, perhaps on the grounds that modern computational, symbolic accounts of representation are automatically immune from the objection. On the contrary, however, statements by Newell (1986, p.33) and others articulating the foundational symbol-system paradigm characteristically assimilate external and internal symbols in such a way as to encourage just such suspicions (see also Bechtel 1998, Lloyd 2003). It is striking that the earliest complaints in the 17th century were precisely about taking things outside the mind as a model for the things inside (Slezak 2002).

Why God Bothered

There is particular irony in the fact that the problem for resembling ideas may be, at a deeper level, the problem shared by Fodor's RTM as well. It is not only resemblance that creates difficulties for ideas or representations. Another manifestation of the same problem may be an 'externalist' conception of representations as semantically evaluable - the claim that mental processes tend to preserve semantic properties like truth. Fodor (1994, p. 9) has said that this is "the most important fact we know about minds; no doubt it's why God bothered to give us any" (1994, p. 9). However, Fodor has seen a dilemma arising from the fact that mental content doesn't appear to supervene on mental processes and, therefore, perhaps "semantics isn't part of psychology" (Fodor 1994, p. 38). This dilemma seems to arise from the fact that semantic evaluability of representations, or old-fashioned 'veridicality' of ideas, like resemblance, depends on being able to make a comparison between ideas and what they purportedly refer to. In Berkeley's idealist response to this problem we can see the precursor and analog to Fodor's (1980) methodological solipsism. In view of these parallels, it is striking, though perhaps not surprising, that Fodor (1994) sees a deep puzzle about how misrepresentation could arise if any causal or correlational theory were true. I have suggested (Slezak 2002, 2004) that the modern problem of misrepresentation is a unnoticed variant of the classical 'Argument from Illusion' and so it should not be surprising that we saw Fodor give an explicit endorsement to just this form of

argument. Fodor argues that if ideas are caused directly by external objects, we can't have misrepresentations (i.e. illusions), whereas the classical argument concludes from the fact that we have illusions, our ideas can't be directly caused by external objects.

In this regard, Fodor's latest discussions raise questions that were central to his seminal book *The Language of Thought* (1975). Dennett (1977) noted:

> Hume wisely shunned the notion of an inner self that would intelligently manipulate the ideas and impressions, but this left him with the necessity of getting the ideas to 'think for themselves'. ... Fodor's analogous problem is to get the internal representations to 'understand themselves' ... If there is any future for internal systems of representation it will not be for languages of thought that 'represent our beliefs to us', except in the most strained sense. (Dennett 1977, p. 274, 5)

Fodor acknowledges that Putnam is aware that nowadays representational theories are formulated so that there is no "user" or exempt agent as undischarged homunculus, but suggests that he fails to acknowledge that perception is direct under these accounts. Fodor asserts that, like telephone conversations with his wife, perception is mediated in all sorts of ways, but "still, it is my wife that I talk to". However, it is here that we see where the debate seems to have become derailed. Direct realists would entirely agree with Fodor's way of making this last point. It is emphatically not the "mediation" of causal processes in all sorts of ways that constitutes the potential problem for representational theories. The problem arises only from *some* of the ways that the mediating causal processes may be conceived. For example, it is not their mediation as such that makes pictures problematic as internal representations subserving imagery. Fodor's account of what makes perception (or talking on the telephone) "direct" is exactly the kind that Putnam, Austin and Reid, inter alia, would endorse.

Gallstones or Headaches?

Fodor suggests that until recently it was generally supposed that explaining having a concept is dependent on explaining what a concept is. That is, the explanation of concept possession should be parasitic on the explanation of concept individuation (Fodor 1998, p. 2). Fodor laments the reversal of this assumption about priority and the direction of explanatory dependence. Fodor's particular target in *Hume Variations* is the one identified in *Concepts*, namely, pragmatists and dispositionalists who hold that having a concept is a matter of some kind of capacity, a matter of what you are able to do as a kind of epistemic 'know how' (Fodor 1998, p. 3). Of particular concern for Fodor, and the reason for Hume's appeal, is their shared opposition to such theories:

> ... an account that renders having concepts as having capacities is intended to preclude and account that renders concepts as species of mental particulars: capacities aren't kinds of *things*; a fortiori, they aren't kinds of *mental* things. (Fodor 1998, p. 3)

Thus, Fodor insisted that "understanding what a thing is, is invariably prior to understanding how we know what it is" (1998, p. 5). He says "epistemic capacities don't *constitute* concepts, but merely *presuppose* them." (p. 20).

It is worth remarking that Yolton (2000) notes that the "pervasive notion" throughout the seventeenth and eighteenth centuries has been that of "presence to the mind" - precisely Fodor's question of concept possession.

Thus, Fodor takes it to be a truism about the possession conditions for concepts that "If concept tokens are mental particulars, then having a concept is being in a relation to a mental particular" (1998, p. 3 fn 1). However, such talk of "possession conditions" is a framework that biases our theory towards an object-account of concepts since "possession" is itself, like "mental object" a metaphor with misleading connotations when we are concerned with states of the mind-brain. The old-fashioned term 'presence to the mind' is preferable in this regard. Colloquially, we may speak of "having a headache", or "having little patience", but talk of "possession conditions" in such cases is not obviously as appropriate as for tables and chairs. For example, having a headache is more like having indigestion or being sunburnt than owning something. By contrast, having gallstones is undeniably object possession but likely to be a poor model for psychological states. Less figuratively, the question is whether we should adopt a Malebranchean-Lockean-Humean object theory or an Arnauldian-Cartesian-Reidian process, act theory. At the very least, object implications of colloquial idioms in folk-psychology propositional attitude talk should not prejudge the issue.

These foregoing remarks have a distinctly Rylean flavour, and it is perhaps not surprising that Ryle is among the culprits in Fodor's plot. Fodor suggests that "Mid-century philosophy of mind consisted largely of confusing these issues by endorsing pragmatism as a remedy for dualism" (p. 24) and Fodor regards Ryle's (1949) *Concept of Mind* is the *locus classicus* for this confusion.

However, this analysis is to misread Ryle in a revealing manner. Ryle was concerned, in the first instance, to give a remedy for certain spurious views about our mental life. That is, the conceptual confusions of interest may encourage dualism but are independent of it. For Ryle, dualism is a *consequence* of holding certain mistaken views about our mental life and not identical with these views. Thus, Ryle's criticism of the "intellectualist legend" and the distinction between 'knowing how' and 'knowing that' have nothing to do with dualism. Further, Ryle's criticism of the doctrine of the 'mind's eye' anticipates Pylyshyn's polemic against pictorial theories concluding that "imaging occurs, but images are not seen" (1949, p. 247). Exactly as Pylyshyn would argue a generation later, Ryle said that someone imagining a scene "is not being a spectator of resemblance ... but he is resembling a spectator" (1949, p. 248). Thus, for Ryle pragmatism was primarily a remedy for certain

doctrines that may lead to dualism, but may equally lead to bad theories within a purely physicalist framework.

In Fodor we see the preference for a conception of representation which reverses the trend discernable in the seventeenth century. Yolton points out:

> ... in the writings of the main figures (Descartes, Arnauld, Malebranche, Locke, Berkeley, Hume and Kant), we can follow a gradual emergence of a clear translation or transformation of the old ontological language of presence to the mind into an epistemic presence. (Yolton 2000)

Propositional attitudes are explicated in a now-standard fashion by means of the inner box metaphor. In case anyone had doubts, Fodor (1998, p. 8) makes it clear that talk of belief boxes is a little joke which can be translated into harmless functional terms. However, even when stripped of its whimsical features, the metaphor and its implications are by no means so innocent as generally assumed. Specifically, the locution encourages Fodor's objectified ontological analysis and the priority he gives to questions of what concepts *are* as opposed to how they might be "possessed" or known. Arnauld's book *On True and False Ideas* was precisely a response to this conception in Malebranche's *The Search for Truth*. Given such a conception, it is not surprising that Fodor prefers an atomistic rather than holistic account of meaning, though he notes that the two issues are strictly distinct, since he says current fashion "tends to favor mental objects that are defined by (perhaps all) of their interrelations (p. 12). Nevertheless, propositions conceived as mental atoms are more readily seen as objects having their meanings individually and a holism of a more radical variety would dispense with mental objects altogether in favour of acts, processes or dispositions. It is perhaps not implausible to see a relevant parallel with Locke's project in his *Essay* which Jolley (1999, p. 39) suggests is "self-consciously modeled on the corpuscularian theory of matter". It may be helpful to see that Fodor is true to his avowed Empiricist progenitors even to the extent of such analogous commitments. Remarkably and more directly relevant, Reid, too, captures Fodor's corpuscularism in his attack on Hume's 'ideas' "which, like Epicurus's atoms, dance about in emptiness" (Reid 1985/1997, p. 22).

Fodor considers Stroud's (1977) criticism of Hume's mental atomism and quotes the following passage:

> The Theory of Ideas restricts [Hume] because it represents thinking of having an idea as fundamentally a matter of contemplating or viewing an 'object' - a mental atom that can come and go in the mind ... (Stroud 1977, p. 225,6; quoted in Fodor 2003, p. 11).

What is remarkable about this passage is what Fodor *fails* to comment upon (though the same remarks are quoted again at p. 21), namely, Stroud's concern with having an idea as *contemplating or viewing* an object. Fodor's complete neglect of this point is especially surprising because, as already noted, it has been central to the long tradition of criticism of the 'idea' idea. Of course, Fodor (2000) is right that the causal mediation of representational theories doesn't mean that we perceive the representations themselves, but this is only to identify the problem and not to show that particular accounts actually avoid it. In his new book, Fodor's off-hand treatment of critics of representationalism suggests an insensitivity to this concern explicitly raised by Stroud which, like Putnam's concern, arises from the inherent features of a classical tri-partite conception of ideas. Rorty (1980) too was centrally concerned with what he describes as "the original sin of epistemology" (1980, p. 60), namely, the kind of representationalism originating with Descartes. Rorty describes this as "the Cartesian image of the Eye of the Mind - the very image which has often been accused of leading to the 'veil of ideas' and to solipsism" (1980, p. 94). Fodor's new book does not address such concerns although, of course, he appreciates the way that traditional scepticism arose for the reasons just noted, and he says with mild sarcasm that it led "either to the view that 'strictly speaking' nobody ever saw a piano, or to the view that 'strictly speaking' pianos are mental" (Fodor 2000). With this oblique allusion to Berkeleyan idealism, Fodor suggests that, by contrast, the representational theory "doesn't need to say anything like that now" since it has abandoned its pretensions to being epistemology, content with being only a psychology of perception. In this form representational theories hold that "Causal processes involving mental representations mediate these perceptual relations, but you don't (typically) perceive the representations themselves either directly or otherwise" (Fodor 2000).

First, it is undeniable that an explanatory scientific psychology "doesn't need to say anything like that now", but requiring mental particulars to be semantically evaluable seems to invoke precisely the sorts of problems of veridicality arising for an epistemology concerned with knowledge as true belief. However, be that as it may, the problems of concern do not arise only if the enterprise is conceived as epistemology, as the imagery debate has amply demonstrated. The problem may arise as an unnoticed consequence of certain ways of conceiving the representations, typically when they are modeled too closely on externally perceivable objects. Thus, in a frequently cited overview of the traditional theory of ideas, McRae (1965) pointed to the central notion of an idea as the immediate object of perception or thought that can be traced back to Descartes. McRae suggests that ideas make their appearance in Descartes as immediately present objects for looking at or "contemplating".

> What remains basic for the earlier Descartes, for the later Descartes, for Malebranche and for Locke, is that ideas are the immediate objects of perception, that all knowing reduces to seeing, and that seeing (however intellectual it may be) is the sole operation of which the understanding is capable. It is of secondary importance for their conceptions of what knowing is whether these immediate objects or ideas are in the brain, in the mind, or in God. (McRae 1965, p. 179)

We may add that it is of secondary importance whether these immediate objects are pictures, propositions or other mental particulars *à la* Fodor. He writes:

> ... the questions with which theories of meaning are primarily concerned are metaphysical rather than epistemic. This is as it should be; understanding what a thing is, is invariably prior to understanding how we know what it is. (1998, p. 5)

However, Fodor's impeccable principle giving priority to the question of "what a thing is" rather than "how we know what it is" does not obviously apply as well to *knowledge* as to tables and chairs. Arguably, it is exactly in the case of forms of knowledge that the epistemic questions must take precedence. Take the case of grammar which Chomsky insists has no other reality than the knowledge of a speaker-hearer: Here the question of "what a thing is" collapses into the question of "how we know what it is". Grammar is *constituted* by being a form of knowledge and, therefore, *how we know it* counts as an answer to the question of *what it is*. Similarly, in the case of meanings and concepts, unlike the case of tables and chairs, the question of "what a thing is" is plausibly construed as the question of what we know.

These critics represent the various pragmatic or dispositional views that Fodor excoriates. It is no accident that Gibson's similarly motivated 'ecological' approach dismissed by Fodor, like the closely related 'situated cognition', are theories of direct realism which have been proposed as alternatives to the representationalism of modern computational theories. This is merely one form in which the Malebranche-Arnauld (or Hume-Reid) debate is being rehearsed today.

Le plus séduisant cartésien?

We may better understand Fodor and Hume through their antecedents and their critics. Thus, I have emphasized Malebranche here in part to correct Fodor's misleading allusions to 'Cartesianism' as the provenance of Humean views of representation. Fodor's reference to Hume's Cartesianism needs to be qualified to reflect these nuances and is correct only if we understand it as reference to Malebranche's version of *la pensée cartésienne* and not that of Arnauld or Descartes himself who shared precisely the pragmatism that Fodor is battling. If the interest and relevance of Malebranche's theory today is surprising, this is because its theological trappings and overtones of mysticism have, in Nicolas Jolley's words, "so effectively concealed the seventeenth-century debate from the view of contemporary philosophers" (Jolley 1990, p. 201). Nevertheless, it is not without reason that Malebranche was characterised by a 17th Century author, as we might say of Fodor too, *En un mot, c'est le plus séduisant cartésien que je connaisse* - in a word, the most seductive Cartesian that I know (quoted in Moreau 1999, p.9).

References

Arnauld, A. (1683/1990). *On True and False Ideas*. Trans. Stephen Gaukroger. Manchester: Manchester University Press.

Austin, J. (1962). (1962). *Sense and Sensibilia*. Oxford: Oxford University Press.

Bechtel, W. (1998). Representations and Cognitive Explanations. *Cognitive Science*, 22, 3, 295-318.

Brooks, R. A. (1991). Intelligence Without Representation. *Artificial Intelligence*, 47, 139-159.

Chomsky, N. (1966). *Cartesian Linguistics*. New York: Harpers.

Clark, A., Toribio, J. (1994). Doing Without Representing? *Synthese*, 101, 3, 401-431.

Dennett (1978). A Cure for the Common Code, in *Brainstorms*, Montgomery, Vermont: Bradford Books, 90-108.

Fodor, J.A. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J.A. (1978). Propositional Attitudes, *Representation.* Cambridge, Mass.: Bradford/MIT Press, 177-203.

Fodor, J.A. (1985). Presentation. B.H. Partee, S. Peters and R. Thomason (Eds). *Report of Workshop on Information and Representation*, Washington, D.C.: NSF System Development Foundation, 106-117.

Fodor, J.A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Fodor, J.A. (2003). *Hume Variations*, Oxford: Oxford Univ Press.

Freeman, W.J. & Skarda, C.A. (1990). Representations: Who Needs Them? In J. L. McGaugh, N. Weinberger & G. Lynch (Eds). *Brain Organization and Memory Cells, Systems and Circuits*. Oxford: Oxford University Press.

Greco, J. (1995). Reid's Critique of Berkeley & Hume, *Philosophy & Phenomenological Research*, 55, 2, 279-296.

Jackendoff, R. (1992). *Languages of the Mind*, Cambridge, Mass.: Bradford/MIT.

Jolley, N. (1990). *The Light of the Soul*, Oxford: Clarendon.

Lloyd, D. (2003). Representation. *Macmillan Encyclopedia of Cognitive Science*. London: Macmillan.

Malebranche, N. (1712/1997). *The Search After Truth*. Trans. T.M. Lennon & P.J. Olscamp. Cambridge: Cambridge Univ Press.

Marr, D. (1982). Vision. Cambridge, Mass.: MIT Press.

Moreau, D. (1999). *Deux Cartésiens: La Polémique Entre Antoine Arnauld et Nicolas Malebranche*. Paris: Vrin.

McRae, R. (1965). Idea as a Philosophical Term in the 17th Century, *Journal of History of Ideas*, 26, 2, 175-190.

Newell, A. (1986). The Symbol Level and the Knowledge Level. In Z. Pylyshyn and W. Demopoulos (Eds.) *Meaning and Cognitive Structure*. Norwood, NJ: Ablex.

Putnam, H. (2000). *The Threefold Cord*. New York: Columbia University Press.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton: Princeton University Press.

Ryle, G. (1949). The Concept of Mind, Harmondsworth: Penguin.

Slezak, P. (2002). The Tripartite Model of Representation, *Philosophical Psychology*, Vol. 13, No. 3, 2002, 239-270.

Stroud, B. (1977). Hume. London: Routledge.

Van Gelder, T. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences*, 21, 615-665.

Von Eckardt, B. (1993). What is Cognitive Science, Cambridge, Mass: MIT Press.

Yolton, J. (1984). *Perceptual Acquaintance from Descartes to Reid*. Minneapolis: University of Minnesota Press.

Automatic processing of elements interferes with processing of relations

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
Ohio State University, 208 Ohio Stadium East
1961 Tuttle Park Place, Columbus, OH 43210, USA

Jackie von Spiegel (von-spiegel.2@osu.edu)
Center for Cognitive Science & Department of Psychology
Ohio State University, 208 Ohio Stadium East
1961 Tuttle Park Place, Columbus, OH 43210, USA

Abstract

This research examines mechanisms underlying the primacy in processing of elements over relations. It is hypothesized that elements are detected atomatically even when the task is to ignore them, and this automatic detection may interfere with the processing of relations. In Experiment 1, 4 year-olds and adults were asked to ignore elemental features and to match a test item to a target by detecting the numerical equivalence beween the test and the target. Results indicate that only children, but not adults, cannot ignore elements, thus suggesting that elements could be processed automatically. In Experiment 2, the same task was presented again, except that elements were perceptually-rich. This time, both children and adults exhibited difficulty ignoring elements. These findings point to two important regularities. First, attention is automatically attracted to elements, interfering with processing of relations, and this interference may make relational processing more difficult. And second, perceptual richness of elements amplifies this effect.

Introduction

Humans live in a structured environment: we encounter entities that are interconnected spatially, temporally, or conceptually into larger arrangements. Those components of structure that are entities or separable properties of these entities can be considered elements, whereas the manner in which elements are arranged can be considered relations.

However, it is not self-evident as to what constitutes an element or a relation. For example, a letter may constitute a relational entity in a letter recognition task, but it constitutes an element in a lexical decision task. Similarly, a word may constitute a relational entity in lexical decision, but (as demonstrated by Ratcliff and McKoon, 1989) it constitutes an element in a sentence comprehension task.). Because there is evidence that stimulus familiarity is established early in the course of processing and familiar stimuli are processed by dedicated circuits (Hölscher, Rolls, & Xiang, 2003; Xiang & Brown, 1998), it seems that familiar objects are good candidates for being considered elements.

Processing of structure requires processing of both elements and relations because both elements and relations carry important information: changing a relation (e.g., *the ball is under the table* instead of *the ball is on the table*) as well as changing an element (e.g., *the book is under the table* instead of *the ball is under the table*) can radically change the nature of the information. Processing of structure and the ability to recognize the processed structure at a later time is critically important for both cognition and learning.

There is multiple evidence that pointing to a primacy of processing of elements over relations in terms of processing time, as well as phylogenetic, ontogenetic, and microgenetic time. First, researchers have found that, across a broad array of tasks, elements are processed prior to (or faster than) relations (Goldstone & Medin, 1994; Ratcliff & McKoon, 1989). Second, there is evidence that processing of some relations (e.g., numeric equivalence) is available to great primates, but even for great primates this processing requires much more substantial training than processing of elements (Thompson, Oden, & Boysen, 1997). Third, there are developmental differences in processing of elements and relations, with younger children being less likely to process relations than older children and with greater age differences in the processing of relations than elements (Gentner & Toupin, 1986; Kotovsky & Gentner, 1996). Finally, there is also a large body of evidence indicating that in knowledge rich domains, novices are more likely to process elements (i.e., individual pieces of a chess position, or entities in a problem description) more ably then relations (i.e., the arrangements of pieces in the position, or equations that underlie the solution to the problem), although experts often process relations as well as elements (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; Larkin, 1983; Reed, Ackinclose, & Voss, 1990; Reingold, Charness, Schultetus, & Stampe, 2001).

Taken together, these findings suggest that elements and relations are psychologically distinct. We further contend that there might be an attentional mechanism underlying the differential processing of elements and relations: elements may be detected automatically, and this automatic detection may interfere with processing of relations.

The idea of such a mechanism has been supported by several sets of findings. First, it has been found that the likelihood of processing of relations often varies with the salience of elements. For example, when Structures 1 and 2 (e.g., two sequence of triangles monotonically increasing in

size) differ in elemental and relational correspondences (i.e., the leftmost triangle in Structure 1 has the same size as the rightmost triangle in Structure 2 in terms of an elemental match, whereas it corresponds to the leftmost triangle in Structure 2 in terms of its position), participants focused on relational matches when objects were perceptually impoverished. At the same time, they were more likely to focus on the elemental matches when objects were perceptually elaborated (see Gentner & Medina, 1998 for a review). Second, introduction of a simple warm-up task, which attracts attention to either to relations or to elements, markedly increases processing of relations, but not elements in a target task (Sloutsky & Yarlas, under review), thus indicating that processing of elements is at ceiling. Both sets of findings suggest that elements may be processed in an automatic and obligatory manner.

If this is the case, then elements should be detected even when the task is to ignore them, and these automatically detected elements may interfere with processing of relations. Furthermore, because young children may have difficulty deliberately directing their attention to some properties of stimuli, while ignoring others, it seems likely that children would exhibit these effects under a wider range of conditions than adults.

To test these hypotheses, we created a task, in which participants were asked to focus on a simple relation of numeric equivalence. We selected this relation because previous research demonstrated that even primates could match items having equivalent number of elements, regardless of what these elements were (Thompson, et al., 1997). We deemed it reasonable, therefore, that the relation of numeric equivalence should be available to 4-to-5 year-olds. The task (a variant of Garner's interference task) was presented as a "matching game," in which participants were presented with a Target having a particular number of identical elements (e.g., two identical shapes), and a Test item. If the Test item had the same number of elements, participants should identify it as a match, otherwise they should identify an item as a mismatch. The items were presented under three conditions. First, there was a "fixed" condition, in which the Target and Test items had identical elements, and matching or mismatching relation was the only source of variance. Second, there was a "correlated" condition, in which elements and relations varied together: a relational match accompanied an elemental match, and a relational mismatch accompanied an elemental mismatch. Finally, there was an "orthogonal" condition, in which relations and elements varied independently. Examples of items across the three conditions are presented in Figure 1.

If elements are not attended to in the course of relational processing, there would be no difference in speed and accuracy of matching across the three conditions. If, however, elements are processed automatically, there should be a difference in speed or accuracy between the orthogonal condition and the correlated conditions. Such a decrease would be a strong evidence for automatic processing of elements and fro interference of automatically detected elements in processing of relations. As mentioned above, we expect that even young children process elements automatically, and, therefore, we expect that even young participants would exhibit these effects.

Experiment 1

Method

The goal of this Experiment was to test the hypothesis that even early in development, elements are processed automatically, and this automatic processing of elements may interfere with processing of relations.

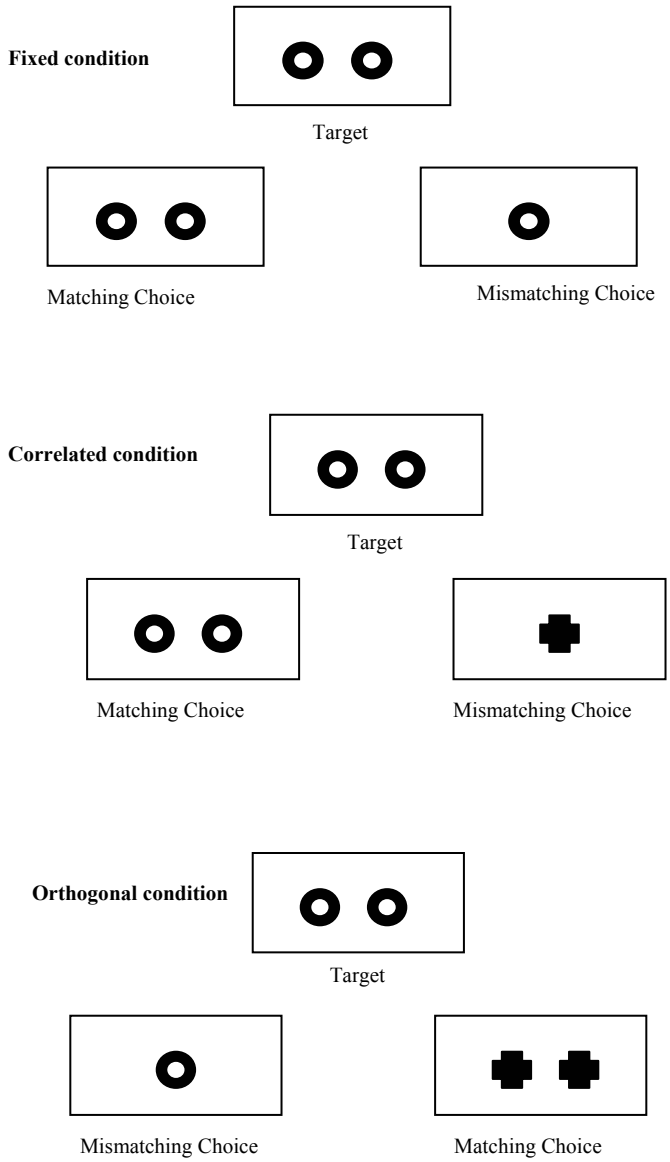


Figure 1: Example of stimuli across the three conditions.

Participants

Participants were 44 young children (Mean Age = 4.41 years, *SD* = 0.346 years; 24 girls and 20 boys) recruited from childcare centers located in middle class suburbs of the Columbus, Ohio area, with approximately equal numbers of participants in the fixed, correlated and orthogonal conditions. There was another group of 37 college undergraduates (13 women and 24 men) participating in the experiment for course credit. There was also approximately equal numbers of participants in the fixed, correlated and orthogonal conditions.

Materials

Materials were stimuli sets, each consisting of three panels. Two of these panels were Target and Choice items and these depicted simple geometric shapes (e.g., circle, triangle, cross). The third panel depicted a Trash can. Stimuli set were presented on screen with the Target and Trash can above each Choice item, with the latter one placed equidistantly to the former two. Participants were told there that if the Choice item has exactly the same number of shapes as the Target, there is a match, and they should point to the Target, whereas if the number is different, there is a mismatch, and they should point to the Trash can. There were a total of 24 trials with 12 matching and 12 mismatching trials. As mentioned above, there were three between-subjects conditions: fixed, correlated, and orthogonal.

There were exactly the same elements employed across trials in these conditions. However, within the trials, there were identical elements in all three panels in the fixed condition, elements covaried with the relation of equivalence in the correlated condition, and elements and the relation of equivalence varied independently in the orthogonal condition.

Design and Procedure

The design included two between subject factors, Condition (fixed, correlated, and orthogonal) and Age (young children and adults). Participants were randomly assigned either fixed, correlated, or orthogonal condition. The dependent variables were accuracy and latency of responses. Young children were given brief training, in which real three-dimensional objects were used to explain the rules of the "matching game." The training was identical across the three conditions.

The child participants were tested individually by a female researcher in a quiet room in their schools, whereas adult participants were tested in a lab room on campus. First, the child participants were trained on a real-object version of the computer task (this training was not used with adult participants). The researcher showed the participants two clear plastic shoeboxes. The instructions said: *This is a toy box* (pointed to box with two stars on the front) *and this is a trash can* (pointed to plain box). *There are two stars on this toy box, so this is the "two-toy" toy box. If I give you two toys, you put them in here. If I don't give you two toys, you put them in the trash can*. Then the researcher set one, two, or three toys in front of the participants and asked, "Should these go in the toy box or the trash can?" The toys were small, colorful, plastic toys (i.e. sunglasses, cars, tops). The participant was given feedback for these training trials. The participant had four trials with the two-toy toy box, after which the researcher replaced it with a one- or three-toy toy box (designated by stars on the front) and restated the instructions. Each participant had four trials with each toy box, totaling 12 training trials. If the participants were successful on the last three trials, they proceeded to the computer task. If a participant was not successful, the experiment was terminated because the participant did not demonstrate understanding of the task.

The computer task was the same as the training, except that the child participants responded by pointing to the Target or Trash can or naming them. Children's responses were entered by the experimenter. Adult participants entered their choices by pressing appropriate buttons on the keyboard. The experiment was administered on computer and was controlled by SuperLab Pro 2.0 software.

The screen was divided by a horizontal line, with the Target and Trash can above the line and the "toys" to be moved below. The toys were actually two-dimensional shapes (square, triangle, cross, circle, heart, and diamond). The researcher said to the child participants: *Now we are going to do the same thing, but on the computer. Here is the toy box (i.e., the Target) and here is the Trash can and here are the toys* (pointed to each as they were mentioned). *If the number of toys on the toy box is the same as the number of toys down here, then you tell me to put them in the box. If the number of toys on the toy box is different than the number of toys down here, then you tell me to put them in the trash.* The researcher then pressed "1" for box and "0" for trash, according to the participants' responses. The adults had similar instructions on the computer screen and two examples (one match and one mismatch). There were four warm-up trials on the computer and 24 test trials. Warm-up trials were exactly as the test trials, except that the former were accompanied by feedback.

Results and Discussion

Because procedures for children and adults differed slightly, we present their data separately. Recall that children's responses were entered by the researcher, which added time to their latencies. To adjust for this added time, we conducted a separate experiment, in which we used measured the speed of pressing buttons by the researcher. We then averaged this time across trials, and subtracted it from each child participant response.

Children. Overall child participants exhibited high accuracy of responding with 91% correct in the fixed condition, 97% correct responses in the correlated condition, and 88% correct in the orthogonal condition. There was an approaching significance difference in accuracy between the orthogonal and the correlated condition, with greater

accuracy in the correlated condition, t (27) = 1.81, p = .08. Latencies across the three conditions are presented in Figure 2.

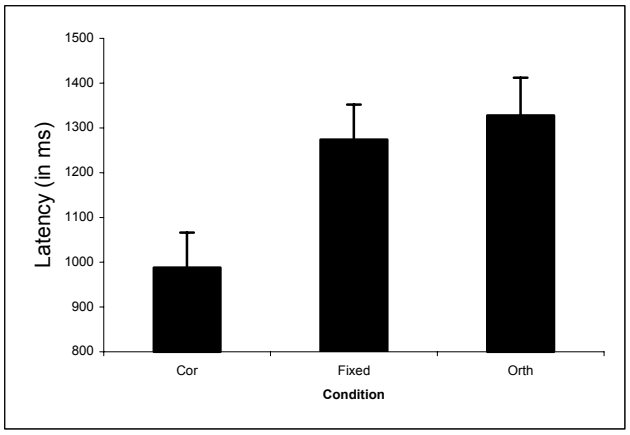


Figure 2. Children's latencies by condition. Error bars represent Standard Errors of the Mean.

These latencies were subjected to a one-way ANOVA. The analysis pointed to significant differences across the three conditions, F (2, 41) = 5, 21, p = .01. Post-hoc Tukey tests indicated that responses in the fixed and the orthogonal condition were slower than responses in the correlated condition, ps < .05. These results indicate that there was a significant speed up in the correlated condition, pointing to an automatic processing of elements.

Adults. Adults' data differed from those of young children in that there was little evidence of elements interfering with processing of relations. Adults exhibited comparable accuracy across the conditions, with 97% correct in the fixed condition, 97% correct in the correlated condition, and 95% correct in the orthogonal condition, ns, p > .3. Similarly, they exhibited comparable latencies across the conditions, 1017 ms in the fixed condition, 987 ms in the correlated condition, and 988 ms in the orthogonal condition, ns, p > .8.

Results of this experiment indicate that children, but not adults exhibit automatically processing of elements even when instructed to focus on relations.

Experiment 2

The goal of this experiment was to test the second hypothesis that perceptual richness of elements may amplfy the effects of automatic detection of elements found in Experiment 1.

Participants

Participants were 40 young children (Mean Age = 4.49 years, SD = 0.33 years; 24 girls and 16 boys). They were recruited in the same manner as in Experiment 1 and there were approximately equal numbers of participants in the fixed, correlated and orthogonal conditions. There were also 70 college undergraduates (17 women and 53 men) participating in the experiment for course credit. There was also approximately equal numbers of participants in the fixed, correlated and orthogonal conditions.

Materials

The task was set up identically to the task in Experiment 1 expect for the nature of the stimuli. Instead of the simple geometric shapes, the stimuli were perceptually rich images of common animals (e.g., bird, dog, turtle). An example of stimuli is presented in Figure 3. Again, participants were told there that if the Choice item has exactly the same number of shapes as the Target, there is a match, and they should point to the Target, whereas if the number is different, there is a mismatch, and they should point to the Trash can. There were a total of 24 trials with 12 matching and 12 mismatching trials. Again, there were three between-subjects conditions: fixed, correlated, and orthogonal.

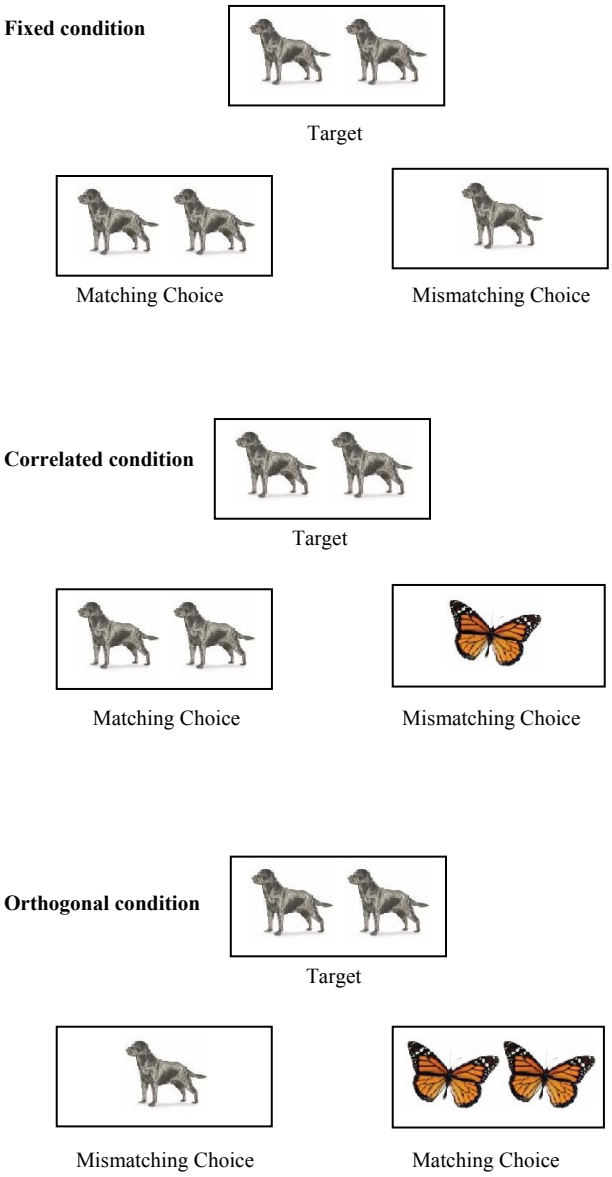


Figure 3: Example of perceptually rich stimuli across the three conditions.

Design and Procedure

The design of this experiment is identical to the design of Experiment 1. The design included two between subject factors, Condition (fixed, correlated, and orthogonal) and Age (young children and adults). Participants were randomly assigned to each level of the Condition. The dependent variables were accuracy and latency of responses. Again, young children were given brief training, identical to the training in Experiment 1 and using the same three-dimensional objects.

Results and Discussion

Data was entered and analyzed in the same manner as in Experiment 1.

Children. Similar to Experiment 1, the child participants exhibited high accuracy of responding with 94% correct in the fixed condition, 97% correct responses in the correlated condition, and 94% correct in the orthogonal condition. There were no significant differences between the accuracies of the three conditions. Latencies across the three conditions are presented in Figure 4.

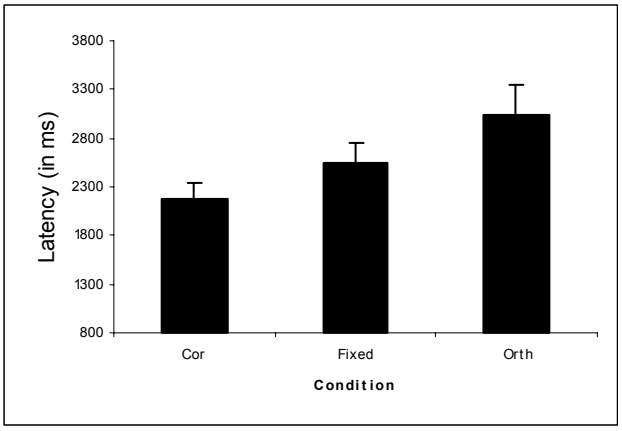


Figure 4. Perceptually rich stimuli. Children's latencies by condition. Error bars represent Standard Errors of the Mean.

These latencies were subjected to a one-way ANOVA. The analysis pointed to significant differences across the three conditions, F (2, 36) = 3.60, p = .038. Post-hoc Tukey tests indicated that responses in the orthogonal condition were slower than responses in the correlational condition, p = .03. These results indicate that there was a significant slow down in the orthogonal condition, pointing to an interference on the part of perceptually rich elements.

Adults. Adults exhibited interference effects showing somewhat lower accuracy in the orthogonal condition with 97% correct in the fixed condition, 97% correct in the correlated condition, and 92% correct in the orthogonal condition. The one-way ANOVA pointed to significant differences across the three conditions, F (2, 67) = 3.41, p = .04. Post-hoc Tukey tests indicated a significant difference in accuracy between the orthogonal and the correlated condition, with greater accuracy in the correlated condition, p = .05, and an approaching significance difference in accuracy between the orthogonal and the fixed condition, with greater accuracy in the fixed condition, p = .08. They exhibited comparable latencies to the children across the conditions, 911 ms in the fixed condition, 943 ms in the correlated condition, and 1017 ms in the orthogonal condition, ns, p > .8.

Results of this experiment indicate that increasing the perceptual richness of the elements produces more interference of automatically detected elements with processing of relations not only in children, but also in adults.

General Discussion

Two important findings stem from the reported experiments. First, when elements are perceptually impoverished, and the task is to focus on relations, young children automatically attend to elements. And second, perceptual richness of elements amplifies this effect in children, and it reveals the effect in adults.

Findings that elements are attended to automatically, even when the task is to ignore them, may explain the earlier found primacy in processing of elements. As mentioned above, elements are processed prior to (or faster than) relations (Goldstone & Medin, 1994; Ratcliff & McKoon, 1989), younger children are less likely to process relations than older children (Gentner & Toupin, 1986; Kotovsky & Gentner, 1996), and experts in a domain are more likely to process domain-important relations than novices (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981). These findings suggest that the difficulty of processing relations may often stem from participants inability to ignore irrelevant elements.

This attentional mechanism is capable of explaining several existing findings. In particular, there is evidence (see Gentner & Medina, 1998; Markman & Gentner, 1993) that when elements are perceptually-rich, participants are more likely to focus on matching elements than when elements are perceptually-impoverished. Because perceptually-rich stimuli are more likely to engage attention than perceptually-impoverished stimuli, it seems that differences reported by Gentner & Medina (1998) may stem from greater attention automatically attracted to perceptually-rich elements.

Finally, there is evidence that although young children have difficulty processing relations under regular conditions, they are significantly more likely to process relations when relations are labeled (Gentner &

Loewenstein, 2002; Kotovsky & Gentner, 1996). Again, it seem that labels attract attention to relations thus making it easier to ignore elements.

It seems that these examples demonstrate that, unless attention is attracted to relations and away from elements, participants are more likely to automatically attend to elements, and attention to elements may interfere with their processing of relations. Recall that current research used a highly familiar relation of numerical equivalence, and interference effects manifested themselves in a decresed latency or accuracy. However, it is possible that when relations are less familiar, interference may result in a failure to detect a relation.

In short, reported results indicate that elements are detected automatically. The results also indicate that perceptual richness of elements amplifies the effect of automatic detections of elements. It is possible that automatic detection of elements may interfere with processing of relations, especially when the task is to ignore elements.

Acknowledgments

This research has been supported by a grant from the National Science Foundation (REC # 0208103) to Vladimir M. Sloutsky.

References

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Chi, M. T. H., Feltovich, P. G., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Gentner, D., & Loewenstein, J. (2002). Relational language and relational thought. In E. Amsel & J. Byrnes (Eds.), *Language, literacy, and cognitive development: The development and consequences of symbolic communication* (p. 87-120).

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*, 263-297.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277-300.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20*, 29-50.

Hölscher, C., Rolls, E. T., & Xiang, J.-Z. (2003). Perirhinal cortex neuronal activity related to long-term familiarity memory in the macaque. *European Journal of Neuroscience*, *18*, 2037-2046.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797-2822.

Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & A. Stevens (Eds.), Mental models, (pp. 75-98). Hillsdale, NJ: Erlbaum.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.

Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology*, *21*, 139-155.

Reed, S. K., Ackinclose, C. C., & Voss, A. A. (1990). Selecting analogous problems: Similarity versus inclusiveness. *Memory & Cognition*, *18*, 83-98.

Reingold, E.M., Charness, N., Schultetus, R. S., & Stampe, D. M. (2001). Perceptual automaticity in expert chess players: Parallel encoding of chess relations. *Psychonomic Bulletin & Review*, *8*, 504-510.

Thompson, R. K. R., Oden, D. L., & Boysen, S. T. (1997). Language-naive chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, *23*, 31-43.

Xiang, J.-Z., & Brown, M. W. (1998). Differential neuronal encoding of novelty, familiarity, and recency in regions of the anterior temporal lobe. *Neuropharmacology*, *37*, 657-676.

High-Level Cognitive Processes in Causal Judgments: An Integrated Model

Andrea Stocco (stocco@units.it)
Danilo Fum (fum@units.it)
Stefano Drioli (drioli@psico.units.it)
Dipartimento di Psicologia, Università di Trieste
via S. Anastasio 12, I-34134, Trieste, Italy

Abstract

The problem of whether human causal judgments could be better explained by associationistic or probabilistic accounts is dealt with in the paper that reviews the basic tenets of the power PC theory (Cheng, 1997), the most famous of the probabilistic explanations, and discusses some results obtained by Fum & Stocco (2003) that are at odds with power PC predictions. An integrated model is described that is capable of explaining those findings, and a new experiment is presented in which the predictions of the model and of the power PC theory are contrasted in the case in which the causal power of a compound cue is equal to one of its components. The results clearly corroborate the model that provides, moreover, an explanation for some data that lie outside the scope of the power PC theory.

Introduction

Recent research on adult causal cognition has been focusing on two main kinds of theoretical explanations that capture many of the central findings in the field.

Associative accounts (Shanks, 1995) consider the causal reasoning performed by humans as similar to the classical conditioning happening in animals and claim that, because both processes involve the detection of the same predictive relations, they may use a common mechanism. The most famous model in this class is that of Rescorla & Wagner (1972)—henceforth R&W—that has been successfully applied to account for a series of phenomena—like blocking (Kamin, 1969), overshadowing (Price & Yates, 1993), conditioned inhibition (Chapman & Robbins, 1990), and contingency effects (Dickinson, Shanks, & Evenden, 1984), to name only a few—that were originally discovered in animals, and that have been demonstrated to play a critical role in human causal learning, too.

Probabilistic theories, on the other hand, rely essentially on the analysis of the contingencies that organisms are supposed to acquire by interacting with their environment, and try to estimate the extent to which a cue (or potential cause) can determine a given outcome. The most famous among these accounts is constituted by Cheng's power PC theory (Cheng, 1997), an extension of the probabilistic contrast model developed by Cheng & Novick (1990).

Fum & Stocco (2003) argued that associative and probabilistic models possibly cover distinct steps in human causal induction, with associative accounts describing the processes by which people (and animals) notice and extract statistical connections between events, and probabilistic models capturing the reasoning skills brought to bear in causal cognition. Investigating the role of compound cues in causal judgments, however, they obtained experimental findings that could not be explained, in their entirety, by either group of theories.

In the paper we review the power PC theory and illustrate some results obtained in Fum & Stocco (2003) that seem to falsify it. We present a new model that, while being compatible with previous data, is able to explain those puzzling results. We describe an experiment in which our model and the power PC make contrasting predictions, and we present findings that corroborate our hypothesis. We discuss some further data that, while implied by our model, are out of the scope of the power PC theory. We conclude the paper by summarizing the features of our account of human causal cognition and by outlining some possible developments.

A Probabilistic Account

Perhaps the simplest of the probabilistic models of causation is given by the ΔP rule (Jenkins & Ward, 1965) that formalizes the idea that people mentally compare the frequency of an outcome O in presence and in absence of a given cue C: $\Delta P_c = P(O|C) - P(O|\neg C)$. If the difference is around 0, the outcome is just as likely when the cue is present as when it is absent; if it approaches 1, C is perceived as producing O; if it approaches -1, the cue is seen as preventing the outcome.

Relying on this idea, Cheng & Novick (1990) developed their probabilistic contrast model assuming that, in presence of a set of possible causes for an effect, the ΔP for each cause is computed on the so-called *focal set*, defined as "a contextually determined set of events that the reasoner uses as input to the covariation process" (Cheng, 1997, p. 371). When a putative cause is taken into account, all other causal factors are kept constant within the focal set, and ΔP is computed on a background of constant alternative causes.

The transition from the probabilistic contrast model to the power PC theory was motivated by a series of problems that could not be adequately explained by the former nor by alternative associative accounts like the R&W. The power PC theory essentially computes how much a ΔP judgment should be discounted for providing an estimate of the causal power of a cue. It also detects special conditions in which the causal power cannot be deduced from ΔP.

One of the tenets of the theory is that, whenever the possible alternatives to a candidate cause C are kept under control and ΔP is non negative, $\mathbf{C}$ (i.e., the causal power of C to generate the outcome O) is given by:

$$\mathbf{C} = \frac{\Delta P_c}{1 - P(O|\neg C)}$$

According to the power PC theory, identical values of ΔP associated with different values of $P(O|\neg C)$, the *base rate*, will lead to different causal judgments. When the alternative causes are controlled, the theory predicts that, with ΔP kept constant, the causal power increases with an increase in the value of the base rate. As a special case, if the base rate is equal to 1, the causal power remains undefined, because the denominator becomes 0. If the base rate is equal to 0, the power PC reduces to the probabilistic contrast model, and the causal power depends exclusively on ΔP. Finally, if ΔP is 0, the causal power of C is 0, too.

Fum & Stocco (2003) focused on some interesting consequences of Cheng's theory concerning the role of compound cues, and set up an experiment to test them. To reduce the complexity of the theoretical framework, and to establish a clear experimental paradigm, four assumptions were made. First, the causal power of a generic cue A was defined as the probability that, all other things being equal, the cue would produce the outcome O: $\mathbf{A} = P(O|A)$. Second, a given outcome had a null probability of being obtained in absence of the cue: $P(O|\neg A) = 0$. Third, all cues were considered as independent. Fourth, all the cues were pure causes: none of them was an enabling condition (Cheng & Novick, 1991) nor needed any enabling conditions to produce its effect.

Given these assumptions, it is possible to deduce[1] some important consequences from the power PC theory. We focus here on two of them:

Irrelevance of Compound Previous experience with a cue presented in a compound form should be irrelevant to the judgment of its causal power, given that there are trials in which the cue appears alone. It is a tenet of both the power PC theory and of the probabilistic contrast model that only items in the focal set—where everything, but the candidate cause whose causal power is being evaluated, is kept constant—are taken into account to compute ΔP. Let us consider, for instance, the classical backward blocking paradigm (Chapman, 1991; Dickinson & Burke, 1996; Shanks, 1985), where compound trials of the form $(A, B \rightarrow O)$ are followed by a set of trials of the form $(A \rightarrow B)$. In this context, an adequate focal set to evaluate the causal power of the blocking cue A is constituted by trials $(A \rightarrow O)$ only, because by including $(A, B \rightarrow O)$ in the set, the cue B would also vary. The power PC theory therefore predicts that a previous presentation of a compound cue $(A, B \rightarrow O)$ should not influence the following judgment for cue A.

Equalization to Compound Sometimes it could be necessary to estimate the causal power of a cue over an inadequate focal set. Taking the example of backward blocking again into account, it should be noted that the trials $(A, B \rightarrow O)$ constitute an inadequate focal set for evaluating the causal powers of A and of B because both cues are covariant within the same set. However, this is exactly what participants in the control group of that paradigm are requested to do, and what a theory is supposed to provide an explanation for. When participants are forced to make a judgment, they should adopt the trials $(A, B \rightarrow O)$ as a focal set, and this would lead them to assign both cues the same causal power of the compound. Given the fact that the effect is never obtained without the cause, and that each possible cause appears in the set of trials $(A, B \rightarrow O)$, it is possible to demonstrate that $\mathbf{A} = P(O|A, B)$ i.e. the causal power of cue A should be equal to the probability of obtaining the outcome given the compound. The same should be true for B.

The experiment carried out in Fum & Stocco (2003) obtained findings that falsified these predictions. More precisely, contrary to the irrelevance of compound hypothesis, judgments concerning a cue A, experienced only in a compound form (i.e. together with another cue B), were significantly *higher* than judgments for the same cue experienced alone. In a similar vein, and contrary to the equalization to compound prediction, the judgments for a cue A, experienced only in a compound form, were significantly *lower* than judgments given to the compound cue embedding A.

While no theoretical explanation for these results was provided in the paper, the findings clearly suggested the existence of important factors determining causal judgements that lie beyond the scope of the power PC theory.

An Integrated Model

Trying to find an explanation for the results reported in Fum & Stocco (2003), we assume that people are able to acquire some knowledge about the contingencies that exist between cues and outcomes. A significant role is played in this phase by associative processes that contribute to the construction of an internal representation for the magnitude of the (single and compound) cues that were directly experienced. There is evidence that it is possible to spontaneously learn such knowledge by interacting with the environment (e.g., Hasher & Zacks, 1984), and we assume that people rely on this information in providing the judgments for those situations they actually encountered.

When a judgment about the causal power of a cue experienced only in compound form is required, the information about the stored magnitudes is used to infer the causal power of the individual novel stimuli, too. This process resembles reverse engineering, because people are supposed to figure out a conceivable distribution for the magnitudes of the single cues that could originate the magnitude of the compound representation.

Let us consider the top panel of Figure 1, that depicts the situation typically encountered in a blocking paradigm. The model assumes that, by interacting with the environment and by noticing the contingencies between cues and outcomes, people are able to construct an internal representation for the causal power of the cues they experience directly—for instance, A, (A, B), and others, like C. The stored magnitude could be a more or less faithful representation of the actual causal power of a given cue but, in any case, it constitutes the basis for causal judgments. To estimate the causal power of an experienced cue, people rely on its magnitude representation, and translate it into the required numerical scale.

When requested to provide an estimate for a cue that was experienced only in compound form (B, in our example), they try to figure out a sensible value for it—in our case, a magnitude for B compatible with the magnitude of both A and the compound (A, B). This process involves a comparison only between those cues that are relevant for deriving the causal

[1] We refer to the original paper for the mathematical derivations.

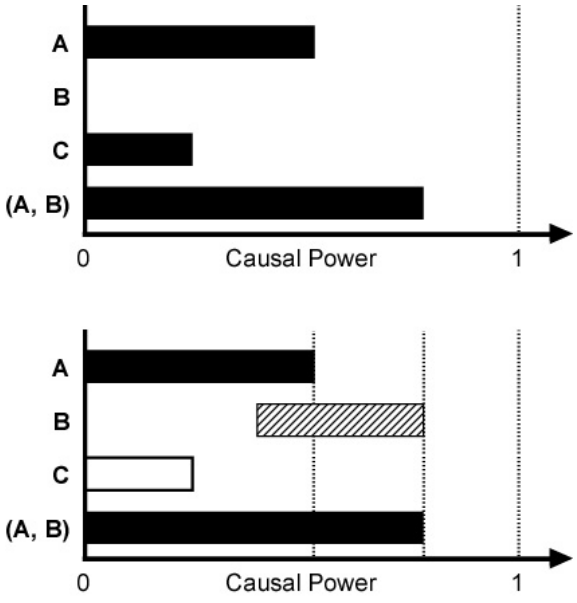


Figure 1: Phases of causal judgement. *Top (associative)*: Causal powers for cues represented as inner magnitudes. *Bottom (probabilistic)*: Inferring the value of cue *B*, that has not been experienced.

power of B (in our example A and (A,B)) and excludes the others (in our case, C). The set of cues taken into consideration conforms to the notion of focal set.

First, we shall observe that the causal power of B, which we denote through its boldface name $\mathbf{B}$, cannot be smaller than the difference between $(\mathbf{A},\mathbf{B})$, the causal power of the compound, and $\mathbf{A}$: if it were so, a certain part of the overall compound effect would remain unexplained: a "rod" shorter than $(\mathbf{A},\mathbf{B})-\mathbf{A}$ could not cover the whole length of the rod representing $(\mathbf{A},\mathbf{B})$. Therefore, $\mathbf{B}_{min}=(\mathbf{A},\mathbf{B})-\mathbf{A}$. Generally, some part of the causal power of B will be shadowed by A: if we simply subtract $\mathbf{A}$ from the compound $(\mathbf{A},\mathbf{B})$, we would in fact grossly underestimate the causal power of B. On the other hand, $\mathbf{B}$ cannot be greater than $(\mathbf{A},\mathbf{B})$, so that, $\mathbf{B}_{max}=(\mathbf{A},\mathbf{B})$.

The model therefore assumes that is "rational" to provide as a judgment for the causal power of B a value lying between $\mathbf{B}_{min}$ and $\mathbf{B}_{max}$. All the values between this range are plausible and coherent with the magnitude of the associatively experienced contingencies. The particular judgments provided by participants vary stochastically between this range. The mean expected value for $\mathbf{B}$ is therefore obtained by weighting each possible $\mathbf{B}$ by its probability $P(\mathbf{B})$:

$$\overline{\mathbf{B}}=\int_{\mathbf{B}_{min}}^{\mathbf{B}_{max}}\mathbf{B}\,P(\mathbf{B})\,\mathrm{d}\mathbf{B}$$

For any symmetrically distributed probability function $P(\mathbf{B})$, the previous equation reduces to the average between $\mathbf{B}_{min}$ and $\mathbf{B}_{max}$:

$$\overline{\mathbf{B}}=(\mathbf{A},\mathbf{B})-\frac{1}{2}\mathbf{A}$$

Explaining Previous Results

Not surprisingly, the model can accommodate the results obtained by Fum & Stocco (2003). Two main findings were reported in that paper. First, some associative effects resulted in a systematic distortion of the causal judgments provided by the participants. The model assumes that these effects are confined to the first phase of the process leading to causal judgments, where inner magnitudes of contingencies are supposed to be acquired.

The second result is more interesting, and it seems critical for the power PC theory. In order to account for backward blocking—one of the most robust and popular contingency learning phenomena—a theory should be able to explain how people make a causal judgment about a cue that has been experienced only in compound form. As previously noted, power PC either should exclude taking into account the inadequate focal set $(A,B\rightarrow O)$, denying thus itself the possibility to account for backward blocking, or should predict, by using only that available set, that the judgments about the causal power of A and B will be equal that of the compound (A,B).

Our model makes a different prediction. According to it, participants are supposed to construct a mental representation of the causal power of A and B such that, by joining (and possibly overlapping) them, they will cover that of the compound (A,B). Because the magnitude of the causal power of one of the cues, let us say $\mathbf{A}$, should be obviously comprised between 0 and $(\mathbf{A},\mathbf{B})$, the estimate for the mean causal power $\overline{\mathbf{B}}$ could be computed by averaging on the predicted values of $\mathbf{B}$, computed on all the values for $\mathbf{A}$ comprised between these extremes:

$$\overline{\mathbf{B}}=\int_0^{(\mathbf{A},\mathbf{B})}\left((\mathbf{A},\mathbf{B})-\frac{1}{2}\mathbf{A}\right)\mathrm{d}\mathbf{A}\Big/\,(\mathbf{A},\mathbf{B})$$

By solving this equation we obtain:

$$\overline{\mathbf{B}}=\frac{3}{4}(\mathbf{A},\mathbf{B})$$

The same result will hold, of course, for $\overline{\mathbf{A}}$.

In the experiment of Fum & Stocco (2003) the value for $(\mathbf{A},\mathbf{B})$ was set to 0.80. Under this condition, the model predicts that $\mathbf{A}=\mathbf{B}=0.60$. The judgments provided by participants were $\mathbf{A}=0.62$ and $\mathbf{B}=0.61$, respectively, with the difference being not statistically significant. It is useful to remind that, according to the power PC theory $\mathbf{A}=\mathbf{B}=(\mathbf{A},\mathbf{B})=0.80$.

Some New Predictions

A model should be considered as good not because it is able to explain previous data but because it allows making bold predictions about future events. For most of the cases, our model produces estimates of the causal power that are close to those provided by the power PC theory. It makes, however, a completely different prediction when, in an extreme blocking situation, the causal power of the compound is equal to the causal power of one of its components: $(\mathbf{A},\mathbf{B})=\mathbf{A}$.

Causal Judgments Under this condition, the power PC theory predicts that, independently of the values assumed by (A,B) and A, the cue B will be perceived as having a null

causal power. In this case, B is the candidate cause, and A is a background cause, as assumed in Cheng's framework. The focal set for B is constituted by all of the trials $(A, B \rightarrow O)$ and $(A \rightarrow O)$. In this set, $P(O|B)$ may be estimated as $P(O|A,B)$, and $P(O|\neg B)$ is given by $P(O|A)$. Given that, we can apply the standard equation for the causal power:

$$\mathbf{B} = \frac{P(O|A,B) - P(O|A)}{1 - P(O|A)}$$

Since $P(O|A,B) = P(O|A)$, it follows that $\mathbf{B} = 0$.[2]

On the contrary, according to our model, **B** will be given a value equal to $(\mathbf{A},\mathbf{B}) - \mathbf{A}/2$. Because, $(\mathbf{A},\mathbf{B})$ has been supposed equal to **A**, it follows that $\mathbf{B} = \mathbf{A}/2$, i.e, we expect that the causal power of B will be judged to be about half of the value of the causal power of A.

Because the model provides some details about the processes underlying causal judgment, it allows making some predictions that lie outside the scope of competitive accounts. More particularly, it predicts the following:

Confidence Ratings According to the model, trying to provide a judgment about the causal power of an experienced cue (e.g. A in the backward blocking paradigm), participants rely on an existing stored representation. On the other hand, when requested to assess the causal power of a cue that has been experienced only in a compound form (e.g., B, in the same paradigm) they cannot access a similar representation to assign a reliable value to **B**, and that constitutes an important source of uncertainty. When requested to estimate the confidence according to which they provide their causal judgments, participants should therefore trust their judgment for cue A more than that provided for B.

Latencies In the backward blocking paradigm—in which the presentation of a compound stimulus $(A, B \rightarrow O)$ is followed by the presentation of one of the its components, e.g., $(A \rightarrow O)$—the judgment concerning the causal power of A could be produced by reading off the value of its internal representation built according to associational principles. To provide a judgment for B, on the other hand, it is necessary to take into account the range of possible values that a coherent judgment could assume, i.e., it is necessary to resort to the second phase hypothesized by the model. As a consequence, the time need to provide a judgment for **B** should be longer than that spent in trying to assess the value for **A**.

The Experiment

To put these ideas under empirical testing, we carried out an experiment, using the Tanks paradigm introduced by Shanks (1985), in which the predictions of our model were directly compared with those deriving from the power PC theory.

Method

Participants The participants were 111 college students (28 males and 83 females) aged between 18 and 36 years (mean and median = 20) enrolled in an introductory Psychology course.

Design and Procedure Participants saw a series of trials in which a picture of an army tank moved across a computer screen. On every trial a weapon system fired, and the tank was hit by one or two projectiles; in some trials the tank was destroyed, in others it remained undamaged. At the moment the weapon fired, one or two colored lights went on in the lower part of the computer screen. The color of the light indicated the kind of projectile that was used. Conceptually, each light could be considered as a separate cue, and the explosion of the tank could be regarded as the outcome.

The experimental session consisted of two sets of 20 trials each. In every trial from the first set two projectiles were contemporaneously fired—this phase could be indicated by $(A, B \rightarrow O)$. In each trial from the second set one of the projectiles was fired alone $(A \rightarrow O)$. Three experimental conditions were set up, each condition differing in the probability of the tank being distroyed by the projectiles. The probabilities were equal to 0.2 (Low), 0.5 (Medium) and 0.8 (High). In the Low condition, therefore, the tank was (randomly) destroyed 4 times in 20 trials, while in the Medium and High condition it was destroyed 10 and 16 times, respectively. The probability of the tank being destroyed by two projectiles in trials from the first set was the same as the probability of being destroyed by a single projectile in the others, i.e., $P(A, B \rightarrow O) = P(A \rightarrow O)$. Finally, trials from the two sets were randomly interleaved for each participant, and participants were randomly assigned to an experimental condition.

Participants were requested to judge the efficacy of each kind of projectile on a scale ranging from 0 to 100, where 0 indicated null efficacy (i.e., the projectile never destroyed the tank) and 100 maximum efficacy (i.e., the projectile always destroyed the tank). They were also asked to indicate, on a seven point scale, the confidence with which they formulated their judgments about the causal power of each projectile. The last main dependent variable that was recorded was constituted by time needed to provide each causal judgment.

At the beginning, participants read an instructions sheet, written in Italian, that explained the task. After that, they saw four practice trials. The tank was randomly destroyed in two of the trials, and in the remaining two it was left undamaged.

At this point the experiment could start. Two colors (chosen in a set that comprised red, yellow, green, and blue) were randomly assigned to the two projectiles used in each experiment session, and the participants were exposed to the 20 trials of the first phase and 20 trials of the second one. To ensure that participants paid attention to the presentation trials, during the experiment four "control" screens appeared at randomly chosen times asking participants to indicate what they had just seen, i.e. which projectile/s was/were fired and whether the tank had been destroyed.

At the end of the presentation trials, participants were asked to provide their judgment about the efficacy of each of the two projectiles they had experienced. After that, they were requested to rate their causal judgments, i.e., to indicate how confident they were about the correctness of their answers.

Stimuli and Apparatus The experiment was performed on a PC equipped with a 15" LCD flat screen and headphones. A

[2] Cheng (1997) derived a computational analysis of the R&W model showing that, under particular conditions—that were met by our experiment—it asymptotically computes a probabilistic contrast over a focal set. It is therefore possible to conclude that, the R&W model makes here the same prediction of the power PC, i.e., $\mathbf{B} = 0$.

custom-made program written in Java was utilized to present the stimuli and to record the participants' judgments. During the presentation trials, the picture of a tank (120 x 45 pixels) moved at constant speed crossing the screen from right to left. A disk (with a diameter of 300 pixel) in the center of the screen simulated the view finder of the weapon system and displayed a desert landscape. The area of the screen outside the disk was kept blank. The tank was visible only when it crossed the disk (employing 3300 ms to cover its diameter), in the remaining time participants could only hear the engine sound through the headphones.

When the tank was approximately at half of its path, completely visible within the view finder, the weapon fired: one or two gunfire sounds were heard and one or two lights, represented by round LEDs (with diameter of 150 pixels) were lit up with the color of the projectile that had been shot.

The tank was always hit, and 1000 ms after the LEDs were brightened, it flashed for 300 ms to simulate the projectile impact. In the trials in which the tank was destroyed, an explosion sound was heard, and the tank was covered by a dust cloud that, after it dissolved, left visible only the wreck. In the trials in which the tank was left undamaged it continued its course until disappearing from the view. In both cases the LEDs remained lit. Each trial lasted approximately 7.5 s; after that, with a shutter effect, the view finder was closed and opened again, and a new trial began.

The control screen utilized to monitor the participants' attention had four LEDs placed at the vertices of an imaginary rectangle positioned at the center of the monitor, each LED associated with two radio button labeled "Yes" and "No", respectively. Participants were asked to indicate which LEDs were lit (and which projectiles were fired) in the very last trial. Moreover, they had to indicate whether the tank had been destroyed or not by choosing between two more yes/no buttons.

The judgments about the efficacy of each projectile were collected through separate screens. In each screen a colored LED was presented together with a request to provide a judgment about the projectile by setting a slider. The mark was positioned at the middle of the slider and the value for the judgment was set to "unassigned". As soon as the participant started moving the mark, an integer value appeared on screen indicating the mark position on a scale ranging from 0 to 100. The confidence rating were collected by having participants check one of seven radio buttons. The buttons at the extremes were labeled with the Italian equivalents of "No Confidence" and "Complete Confidence", respectively.

Results

To avoid considering data that did not accurately reflect the phenomena under investigation, participants that made four or more errors (over a total of 20 possible answers) in the control task were excluded from the sample.[3] The data of 19 (out of 111) participants were thus discarded, and the following analyses were carried out on the remaining ones: 28 participants in the Low condition, 30 in the Medium, and 34 in the High condition, respectively.

[3]The same criterion had been adopted in Fum & Stocco (2003).

Table 1: Mean causal judgments for **A** and **B**

	Low	Medium	High
Judgment for **A**	41.32	64.53	75.44
Judgment for **B**	23.48	39.83	48.53

Causal Judgments The causal judgments provided by participants are reported in Table 1 and illustrated in the top panel of Figure 2. A mixed-design ANOVA was carried out having Condition (Low vs Medium vs High) as a between-subjects and Judgment (A vs B) as a within-subjects variable.

The analysis showed as significant the main effects of the Condition ($F(2,89) = 34.44$, $MSE = 1125.90$, $p < .0001$) and of the Judgment ($F(2,89) = 55.29, MSE = 24357, p < .0001$) but not their interaction. Contrary to the power PC predictions, participants provided judgments for the causal power of B that were completely different from the expected zero value ($t(88) = 14.50, p < .0001$). In accordance with the predictions of our model, their judgments for the causal power of A and B differed significantly, and the value of the judgments increased with an increase in causal power of the compound stimulus (A,B). The model, however, makes a stronger prediction, i.e., that the ratio between the two stimuli should be constant. We calculated this ratio for each participant, and then computed the mean of the ratios for each condition. Results are reported in the bottom part of Figure 2: ratios remained constant across conditions, with only slight and insignificant differences among them. Values of the ratios were around 0.6, close to our estimate, i.e. 0.5. This result could be considered more than satisfactory being our model completely parameter-free.

Confidence Ratings In analyzing confidence ratings and latencies, we pooled the data of all the participants because

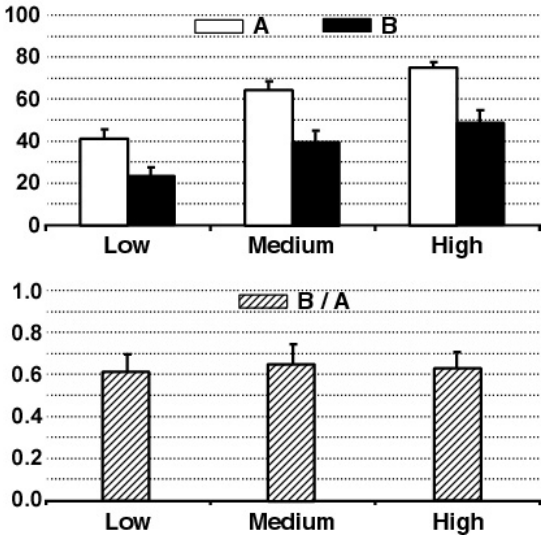


Figure 2: Mean causal judgments (*top*) and mean ratios between **B** and **A** (*bottom*) in the experimental conditions.

our model does not discriminate, under these aspects, among the different conditions.

We found that the confidence ratings for **A** (mean = 3.85) were indeed greater than those for **B** (mean = 3.52), the difference—as revealed by a two-tailed, paired *t*-test—being statistically significant ($t(91) = 2.48$, $p = 0.01$). However, to take into account the possibility that the confidence ratings—collected through a seven point Likert scale—did not conform to a normal distribution, we conducted also a Wilcoxon Matched-Paired test that confirmed the existence of the effect ($T = 676.00$, $N = 92$, $p = 0.04$).

Latencies An even strongest corroboration for our model came from the analysis of latencies. As predicted, the mean time needed to express a jugdment for *A* (17.18 s) resulted smaller than the time needed for *B* (21.31 s). A *t*-test confirmed that the difference was significant ($t(90) = 2.14$, $p = 0.04$). The difference remained significant taking into account the square root ($t(90) = -2.11$, $p = 0.03$) and the logarithm ($t(90) = 2.07$, $p = 0.04$) of the latencies. Causal induction processes seem to respect the time course we hypothesized.

Conclusions

In the paper we presented a model of high level cognitive processes in causal induction that is able to explain previous findings that resulted antithetical to some predictions of the power PC theory, and that can take into account new data that are at odds with, or beyond the scope of, that theory. The model assumes that, when required to provide a causal judgment, people recur to both associative and probabilistic processes. These processes play, however, a different role in causal cognition: associative processes contribute to the construction of an internal representation of the power of directly experienced cues, while probabilistic reasoning is required to estimate the magnitude of the non directly perceived ones.

In the paper we have gone one step further in the description of the cognitive processes underlying such judgments, and we have extended the set of data that may be taken into consideration to discriminate between different accounts. We find particularly important the fact that participants were able to express faithful subjective confidence about their own ability to estimate the causal power of different cues, an indication that these estimates lie above the subjective threshold (Dienes & Perner, 1999) and, therefore, that some kind of explicit knowledge is required to provide these judgments.

We are currently working to extend the model on paradigms other than blocking, and to provide finer estimates of human causal judgment. In particular we think that associative effects, probably reflecting an evolutionarily older, non-specific learning system, shall be further investigated. Our model might be of guidance in determing under which conditions such effects may overcome the explicit processes we described.

Acknowledgements

We thank Fabio Del Missier for some helpful suggestions and the fruitful discussions. We would also like to thank an anonymous reviewer for the inspiring comments.

References

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *17*, 837–854.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory & Cognition*, *23*, 510–524.

Cheng, P. W. (1997). From covariation to causation. A causal power theory. *Psychological Review*, *104*, 367–405.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*, 545–567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

Dickinson, A., & Burke, J. (1996). Within compound associations mediate the retrospective revaluation of causality judgments. *Quarterly Journal of Experimental Psychology*, *49B*, 60–80.

Dienes, Z., & Perner, J. (1999) A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735–755.

Dickinson, A., Shanks, D. R., & Evenden, J. L. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology*, *36A*, 29–50.

Fum, D., & Stocco, A. (2003). The role of compound cues in causal judgment: Associative and probabilistic effects. In F. Schmalhofer, R. M. Young & G. Katz (Eds.), *Proceedings of EuroCogSci 03*, Mahwah, NJ: Lawrence Erlbaum, 127–132.

Hasher, L, & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *12*, 1372–1388.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs*, *7*, 1–17.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell & R. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton.

Price, P. C., & Yates, J. F. (1993). Judgmental overshadowing: Further evidence of cue interaction in contingency judgment. *Memory & Cognition*, *21*, 561–572.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp.64–99). New York: Appleton-Century-Crofts.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology*, *37B*, 1–21.

Shanks, D. R. (1995). *The psychology of associative learning.* Cambridge, UK: Cambridge University Press.

Inferring knowledge of properties from judgments of similarity and argument strength

Sean Stromsten (sean_s@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139-4307 USA

Abstract

Psychological similarity has been invoked to explain many phenomena, including judgments of the strength of inductive arguments (Osherson et al., 1990). The present work follows the suggestion of Tenenbaum and Griffiths (2001) that judgments of similarity and judgments of argument strength cohere because they are essentially judgments of the same kind, which consult the same knowledge of properties of objects or classes. I work backward from people's judgments of argument strength and similarity to the knowledge of properties–specifically, knowledge of probable property extensions–that might explain the coherence among those judgments. I show that the knowledge inferred can be used to predict other such judgments. I then examine this knowledge for structural properties such as taxonomic organization.

Induction, or generalization from examples, is a central cognitive capacity in need of two kinds of explanation: (1) What representations and processes underly induction? (2) Why do we have those representations, and carry out those processes? That is, to the degree that they work, what relation to right reason explains their success? I focus here on the second question, and with respect to just one much-studied inductive task, category-based induction.

To illustrate this task, consider the following inductive argument (after Osherson, et al. , 1990)

Chimpanzees require biotin for hemoglobin synthesis.
Gorillas require biotin for hemoglobin synthesis.

Mammals require biotin for hemoglobin synthesis. (1)

Horizontal lines separate conclusions from their premises. The premises assert facts about categories of objects, and the conclusions do not (in general) follow deductively.

Osherson et al. collected extensive judgments of the strength of such arguments–that is, the subjective probability of the conclusions, given the premises. The arguments contained various mixtures of ten species of mammals in the premises, but all conclusions were about either 'horses' or 'all mammals' (the set of all mammals is approximated, in all models discussed here, by the set of ten mammals used in the arguments)[1].

In order to study argument strength, rather than particular knowledge of predicates, the premises and conclusion assert so-called 'blank' predicates of species, about which experimental participants will not have direct knowledge. The biological sound of the predicates, and the fact that they are asserted to be true of all members of one or more species, are clues that they are biological properties. The intention, then, is that participants have no choice but to fall back on categorical biological knowledge.

Osherson et al. propose the *similarity-coverage* model, which predicts the judged strength of these arguments as a function of judgments of pairwise similarity among the species of animals in them. The strength $g(X, Y)$ of a conclusion, according to this model, is a weighted sum of (1) the similarity of the premise categories X to the conclusion category Y, and (2) the degree to which the diversity of the premise categories 'covers' the lowest superordinate category S including both the premise categories and the conclusion category:

$$g(X,Y) = \alpha \max_i sim(X_i, Y) + (1-\alpha) \sum_j \max_i sim(X_i, S_j).$$

[1] In what follows, in addition to the 81 judgments studied by Osherson et al. , I use data on 28 additional judgments, collected by Sanjana and Tenenbaum (2003). They designed these additional generalization judgments to demonstrate effects which their Bayesian model could explain, but which the Osherson et al. model could not. Again, 'horse' was the only species in the conclusions. The innovation was repeated examples of the same species, which required a cover story that makes such examples reasonable. Participants observed a set of example animals–individual animals–with a particular disease, and were then asked to judge the probability that horses could get the disease. Trusting that participants assume that disease susceptibility is a species property, I aggregate these data with the Osherson et al. data.

Osherson et al. test their model against a number of robust qualitative patterns in the way the plausibilities people assign to such arguments relate to the similarities of the categories used. A few examples of these patterns will illustrate the utility of the similarity and coverage terms. The argument

Chimpanzees require biotin for hemoglobin synthesis.

Gorillas require biotin for hemoglobin synthesis. (2)

is stronger than the argument

Chimpanzees require biotin for hemoglobin synthesis.

Dolphins require biotin for hemoglobin synthesis. (3)

because gorillas are more like chimpanzees than dolphins are. The argument

Chimpanzees require biotin for hemoglobin synthesis.
Dolphins require biotin for hemoglobin synthesis.

Mammals require biotin for hemoglobin synthesis. (4)

is stronger than argument (1), which may be explained by the greater 'coverage' of the set of mammals by 'chimpanzees and dolphins' than by 'chimpanzees and gorillas'.

It may strike the reader that these intuitions require more than purely psychological, *ad hoc* explanations, for surely they are *correct*. If so, they require normative (Bayesian) explanation. This point has been addressed by several authors, beginning with Heit (1998).

There are a number of other reasons for dissatisfaction with an explanation of judgments of argument strength in terms of judgments of similarity, having nothing to do with the degree of predictive success of the similarity-coverage model. The most obvious, perhaps, is that similarity and argument strength are judgments of equal status, equally in need of explanation. Another objection is that the judged similarity of x to y is not a stable, context-free property of the pair (Tversky, 1977). If judgments of similarity must be computed on-the-fly, as judgments of the strength of arguments presumably are, then whatever knowledge is consulted when computing similarities could be consulted when computing argument strengths, without computing similarity as an intermediary. This is, in essence, the kind of explanation proposed in the Bayesian models of Sanjana and Tenenbaum (2003) and Kemp and Tenenbaum (2003). For purposes of direct comparison, they predicted argument strengths from similarities, just as Osherson, et al. did, but did so by way of inferring taxonomic knowledge presumed to underly both similarity and argument strength judgments.

Bayesian generalization

Before discussing the details of particular proposals, I will briefly review the notion of category-based induction as Bayesian generalization, as formulated by Tenenbaum and colleagues. We assume that:

- The premise categories are random samples from the set c of categories having the target 'blank' property.
- Prior to receipt of any examples, the generalizer has a hypothesis space H, where each hypothesis $h \in H$ is a possible extension for the target property. The generalizer also has a probability distribution over H, which represents the prior degree of belief that each candidate is the extension of the target property. This prior distribution may itself be sensitive to (conditional on) other information, for instance, about the *kind* of property being generalized.

The probability that a category y is a member of the set c, given a set of n examples $\mathbf{x}$ drawn at random from c , can be found by summing over hypotheses:

$$P(y \in c|\mathbf{x} \sim c, \xi) = \sum_h P(y \in c|c = h)P(c = h|\mathbf{x} \sim c, \xi).$$

Here $\mathbf{x} \sim c$ means that the examples $\mathbf{x}$ are random draws from c, and ξ represents background information. The first term is 1 if $y \in h$, and 0 otherwise. The second term can be re-written in an enlightening form by Bayes rule:

$$P(y \in c|\mathbf{x} \sim c, \xi) = \frac{\sum_{h \ni y} P(\mathbf{x} \sim c|c = h)P(c = h|\xi)}{\sum_{h'} P(\mathbf{x} \sim c|c = h')P(c = h'|\xi)}.$$

The terms $P(\mathbf{x} \sim c|c = h)$ represent the probability of seeing just the examples $\mathbf{x}$ in n draws from h. Assuming that items in h are drawn with equal probability, then the probability of drawing any particular item in a single draw is $1/|h|$. Then $P(\mathbf{x} \sim c|c = h)$ is $|h|^{-n}$, if h contains all the examples in $\mathbf{x}$, and zero otherwise. The likelihood term $P(\mathbf{x} \sim c|c = h)$ depends only on the examples and the contents of h, so we see now that ξ represents information we may have, prior to seeing the examples, about the probability of the various possible extensions. In what follows, I suppress this term to make the notation simpler.

Further abbreviating $P(c = h)$ to $P(h)$, we can re-write the above as

$$P(y \in c|\mathbf{x} \sim c) = \frac{\sum_{h \supset y \cup \mathbf{x}} |h|^{-n} P(h)}{\sum_{h' \supset \mathbf{x}} |h'|^{-n} P(h')}. \quad (1)$$

Note that the sum in the denominator can be broken into two sums: one is the same as that in the numerator, and the other is over those hypotheses that contain the $\mathbf{x}$ but *not* y. Generalization, then, depends on two weighted sums: one over the properties common to both $\mathbf{x}$ and y, and another over those distinctive to $\mathbf{x}$. Each summand is weighted by both its prior plausibility and its likelihood or 'fit' to the examples.

The two terms have different jobs to do. The fit for extension h–that is, $|h|^{-n}$–gives an advantage to smaller extensions, which is exponential in the number of examples. Without a likelihood term sensitive to the number of examples, we miss an important phenomenon: given that examples are consistent with two extensions, increasing the number of examples ought to shift weight to the more specific extension. For instance, suppose our prior gives high weight to the classes 'mammal' and 'rodent'. Then, given 'mouse' as an example of a species with property P, either class is quite plausible. But adding the further examples 'gerbil' and 'hamster' ought, intuitively, to give a strong advantage to 'rodent', because the selection of three rodents from the larger class is highly coincidental. The likelihood term captures this focusing effect.

Without prior preferences for some extensions over others, the likelihood or 'fit' term will always favor the extension consisting of just the examples, and will have no preference among larger extensions of the same size. For example, given 'mouse' and 'gerbil' as examples of species with some property, generalization to 'turtle' will be just as strong as that to 'hamster'. A prior favoring the natural class 'rodents' over 'rodents minus hamsters, plus turtles' prevents this bizarre behavior.

Similarity as a function of generalization probabilities Tenenbaum and Griffiths (2001) have argued that the similarity of x to y is a function of the probability of generalizing from x to y, or vice-versa, or both. This move gives the infamously slippery notion of similarity some solid footing on the ground of reason, because generalization has a normative foundation in Bayesian statistics. They also show how this view rationalizes earlier work on formalizing similarity and generalization.

For present purposes, we need not delve deeply into the question of just how generalization probabilities determine similarities. I assume, as Osherson et al. do, that similarity is symmetrical, and, further, that it has this particularly simple form:

$$sim(x, y) \equiv \frac{P(y \in c|x \sim c) + P(x \in c|y \sim c)}{2}. \quad (2)$$

Intuitively, this definition says that two items are similar to the degree that one is likely to have a property that the other exemplifies.

Previous work on Bayesian modeling of category-based induction

Various restrictions on the form of the prior could be entertained. For instance, each species might correspond to a location in a low-dimensional Euclidean 'psychological space', with higher priors assigned to sets contained by convex or connected regions. The restricted families of priors investigated by Tenenbaum and colleagues are based on binary trees, with species at the leaves. The sets with highest priors are those corresponding to single subtrees, but some probability is assigned to sets picked out by multiple subtrees. Sanjana and Tenenbaum use a generic method for assigning probabilities to disjunctions of a basis set of hypotheses (in this case, single subtrees), while Kemp and Tenenbaum define a simple 'mutation' process that can generate arbitrary hypotheses, but assigns lower probability to those that require many mutations, or mutations over short branches.

The proponents of these tree-based priors stress that taxonomic trees are not just another restricted family of priors; they are also an independently-motivated *theory* of the domain. People around the world seem to organize creatures into 'folk taxonomies' (Atran, 1995), and the geneology of species does, indeed, form a tree. This kind of theory may be applicable in domains besides biology: even artifact kinds are often the result of a process of copying and modifying earlier designs.

One obvious way to compare various proposed families of priors is to compare predictive accuracies: fit the parameters (for instance, the locations of the points in a metric-space model, or the topology and branch lengths of a tree) to subsets of the judgments and see how well each model predicts the rest.

Rather than competing with previous models on data fit, I take a complementary, 'empirical Bayes' approach (see, for instance, Gelman, et al. , 1995): I place *no* constraints on the form of the prior, find priors that do a good job predicting the data, and then examine those priors for structural properties.

This strategy has an obvious pitfall: an unrestricted search for a prior that makes the data probable may over-fit the accidental properties of the training data, especially, as in this case, when there are many more parameters than data points. Before examining the prior for interesting structural properties, therefore, I demonstrate that the model is not over-fitting so badly as to be uninformative.

Computing a prior from judgments

For any given hypothesis space and prior, Bayesian generalization yields point estimates for a set of similarities and/or argument strengths. To accommodate noisy human data, I take these point estimates to be central tendencies.

In what follows, I refer to the model's prediction of the ith judgment, given a prior, θ, as $j_i^m(\theta)$ (this is given by either equation 1 or equation 2, above). The actual human judgment I denote j_i^h. A simple noise model that respects the constraint that both generalization probabilities and similarities must be between 0 and 1 assumes that

$$log\left(\frac{j_i^h}{1-j_i^h}\right) \sim N\left(log\left(\frac{j_i^h}{1-j_i^h}\right), \sigma^2\right).$$

In words, we apply a transform to each model prediction that may (conveniently) take on any real value, and assume that the similarly-transformed human judgment is normally distributed around this transformed prediction.

A bit of work (omitted here) reveals that the log-likelihood (up to an additive constant) of a set of judgments $\mathbf{j}$ is

$$P(\mathbf{j}|\theta) = \sum_i \log\left(\frac{j_i^h}{1-j_i^h} + \frac{1-j_i^h}{j_i^h} + 2\right) + \frac{1}{2\sigma^2}\left(\log\left(\frac{j_i^h}{1-j_i^h}\right) - \log\left(\frac{j_i^m(\theta)}{1-j_i^m(\theta)}\right)\right)^2. \quad (3)$$

The log likelihood of a set of judgments has a complicated but readily-computed gradient with respect to the prior, involving only the second term in equation 3, which can therefore be optimized by off-the-shelf techniques. I used the method of conjugate gradients, stopping whenever several iterations produced less than a set increase in the log likelihood of the training data. The model was parameterized by 'soft-max' parameters z, where the prior probability of extension i is given by $\theta_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$. On each run, the z were randomly initialized such that the θ were nearly uniform.

proportion used in training		correlations of model and data on remaining	
args	sims	arguments	similarities
0	1	.50±.026	n. a.
1	0	n. a.	.88±.006
0.5	0.5	.61±.029	.77±.033
0.9	0.9	.80±.026	.72±.104
0	0.5	.29±.046	.60±.064
0.5	0	.54±.030	.67±.043
0	0.9	.41±.033	.79±.084
0.9	0	.67±.038	.82±.018

Table 1: Predictions of held-out data given various training data. All rows show averages of ten runs, with associated standard errors.

Predicting held-out judgments

Remarkably, this rather lavishly parameterized model does a reasonable job of predicting randomly held-out judgments when fit to the rest.

Tuned to the judgments of argument strength, the model's predictions of pair-wise similarity agree strongly with the actual judgments, approaching a correlation of 0.9. A number of experiments, using various proportions of each kind of judgment as training data, are reported in table 1.

This model does relatively poorly on the task that has been the focus of the previous work–predicting the argument strengths, given the similarities. A possible explanation for the deficit relative to the other published fits is that the assumptions about the form of the prior made explicitly by using a tree with mutations (and perhaps implicitly in the similarity-coverage model) are essentially correct, in which case opening up the space of priors, as I have done, can only reduce predictive accuracy. As further evidence of over-fitting, early stopping would usually have yielded better predictions, although I could not find a single stopping rule that consistently did so.

Given these results, we can expect that the priors converged to will reflect both the underlying structure of people's knowledge and, to some degree, peculiarities of the data set fit by the over-parameterized model. In the next section, I examine the priors converged on for taxonomic structure.

The 'shape' of the prior

For the purpose of examining the structure of the prior that best explains the data, I focus on results obtained by optimizing the prior over the entire set of judgments.

If we examine the hypotheses with highest priors, certain patterns can be found by eye or statistical

test. Table 2 lists the 10 sets with the highest average prior probability in a typical optimization run.

If the most probable sets are those corresponding to sub-trees of a taxonomic tree, then we should expect that most pairs of such sets will obey taxonomic constraints: either one will contain the other, or they will be disjoint. There are a suspiciously large number of these containment relations among the top-ranked sets–randomly generated collections of sets have as many containment relations between pairs as the top-ranked 100 sets only about 40 out of 1000 times. There is an even more extreme number of disjoint pairs–exceeded not even once in 1000 random sets. Forcing the random sets to match the top-ranked 100 in number of members makes no difference to these results.

However, there are also quite a few partially overlapping sets, which is not what we would expect from a single, strictly-observed tree. The overlap is notably non-arbitrary, however. For instance, the sets 'chimp, gorilla, mouse, squirrel', 'chimp, gorilla, dolphin, seal', and 'mouse, squirrel, dolphin, seal' are composed of just the three pair 'dolphin, seal', 'chimp, gorilla', and 'mouse, squirrel' ('Mouse, squirrel' is not shown here, but ranked 14th in this solution. 'Horse, cow', another pair one might expect, is not far behind.).

What this might point to is a 'mutation' process, as suggested by Kemp and Tenenbaum (2003). While there are sets above that could only be explained by mutations, if a single tree is assumed, they seem to be restricted to cases where the mutations could occur over relatively long branches; members of the very short subtrees, such as 'dolphin, seal', seem to be present or absent in tandem, as predicted by the mutation process.

Another possibility is that the prior reflects uncertainty over *several* taxonomies. Uncertainty about just which taxonomy to consult may be of two kinds: uncertainty about which taxonomy is *correct*; and uncertainty about which taxonomy is *relevant* to the property under consideration. The first is a commonplace of probabilistic modeling, and quite intuitively understandable, in this case. If I perform bottom-up, agglomerative clustering by eye, using the two-dimensional multidimensional scaling solution in figure 1, I come up with the tree topology used in both the Sanjana and Tenenbaum and the Kemp and Tenenbaum papers. But only the lowest-level clusterings are obvious. Is the 'seal, dolphin' cluster closer to the 'gorilla, chimp' cluster than the 'mouse, squirrel' cluster is? It is hard to tell.

The second kind of uncertainty is about which of several trees is relevant. Even if some properties are

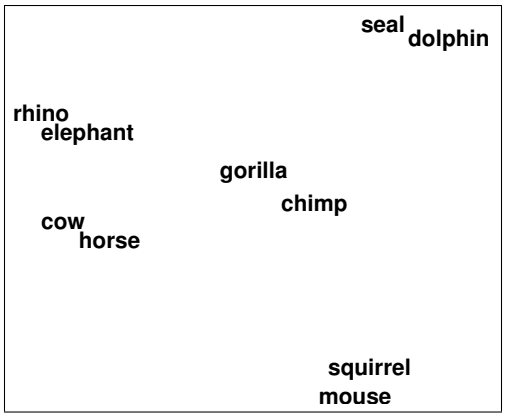


Figure 1: A two-dimensional MDS solution for the similarities of the ten mammals (Euclidean metric, variance accounted for = .81)

distributed according to a particular tree/mutation process, others are likely not to be. This is true even if we restrict attention to biological properties of the kind that are likely to be universal across a species (and which therefore are sensible fodder for the kinds of judgments we consider here). 'Deep' biological properties, such as having a certain organ or metabolic process, are quite likely to respect the 'tree of life'–that representing the genealogy of species. The distribution of other species properties, such as what and how members eat, may be quite random with respect to this tree, but might still respect a different tree.

How might people come to have these priors?

I proceeded above with no constraints on the form of the prior over possible extensions of a new predicate. People or machines asked to make these judgments, however, have no such luxury. They must assume that the extension of the new predicate is systematically related to some known predicate or predicates (and, more generally, that predicates are likely to have systematically related extensions), or have no basis for generalization.

In addition to positing coherence among new properties and old ones, real learners must learn from the kind of data available in the real world. Similarity-like data may sometimes be available, but they are not necessary; people can observe objects and their properties–for instance, that cows, horses, elephants and rhinos all eat grass. Lists of such properties are standard fodder for machine-learning methods, including agglomerative clustering or more sophisticated tree-finding methods. Several strate-

rank	contents									
1	horse	cow	chimp	gorilla	mouse	squirrel	dolphin	seal	elephant	rhino
2							dolphin	seal		
3			chimp	gorilla	mouse	squirrel				
4					mouse	squirrel	dolphin			
5			chimp	gorilla			dolphin	seal		
6	horse	cow		gorilla		squirrel			elephant	rhino
7					mouse	squirrel	dolphin	seal		
8	horse	cow	chimp	gorilla	mouse	squirrel			elephant	rhino
9			chimp	gorilla						
10	horse	cow								rhino

Table 2: The 10 sets with the highest prior probability, on a single optimization over all judgments. There are many instances of nesting, but they are not strictly compatible with any single taxonomic tree.

gies of tree-learning from such data have been applied to a number of standard machine-learning datasets in Kemp et al. (2003).

Summary and discussion

I have suggested a novel technique of general utility for fitting a Bayesian model to a set of judgments. I applied this technique to a large collection of human judgments. Without imposing a taxonomic form on the prior, the prior of a Bayesian model optimized to fit human judgments nevertheless shows significant conformity to taxonomic constraints. It seems that either participants have a bias, in the domain of animals, toward priors that respect the taxonomic constraints, or the raw facts about mammals have this structure (which would, in turn, *justify* a taxonomic bias).

The technique is not limited to the case of a structureless prior over a small set of possible extensions. Any prior that has tractable derivatives with respect to its parameters could be so optimized. In the case of a larger number of categories, whose power set is too large for enumeration, an approximate gradient could be computed using a sample from the current estimate of the prior.

A principled alternative to using held-out data to check models, and to using null-distribution hypothesis tests to look for structure in the prior, is Bayesian model comparison: compare the marginal likelihoods of various structures. For most interesting structure classes, the sums or integrals involved are intractable, but they can be approximated by Markov Chain Monte Carlo or other methods.

Acknowledgments

This work grows out of many conversations with several members of the Computational Cognitive Science lab at MIT: Charles Kemp, Tom Griffiths, and Joshua Tenenbaum. For helpful discussion, I also thank Amy Hoff, Steven Sloman, and David Sobel. This work was supported in part by NSF IGERT grant 9870676 at Brown University.

References

Atran, S. (1995). Classifying nature across cultures. In Smith, E. E. and Osherson, D. N., eds., *An Introduction to Cognitive Science*, volume 3. MIT.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., eds., *Rational Models of Cognition*, 248-274. Oxford.

Kemp, C. and Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society*.

Kemp, C., Griffiths, T. L., Stromsten, S., and Tenenbaum, J. B. (2003). Semi-supervised learning with trees. *Advances in Neural Information Processing 16*.

Osherson, D., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, *97(2)*, 185-200.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.

Sanjana, N. and Tenenbaum, J. (2003). Bayesian models of inductive generalization. In Becker, S., Thrun., S., and Obermayer, K., eds., *Advances in Neural Information Processing Systems 15*. MIT.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian Inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Detecting the Hot Hand: An Alternative Model

Yanlong Sun (Yanlong.Sun@uth.tmc.edu)

University of Texas Health Science Center at Houston
School of Health Information Sciences, 7000 Fannin Suite 600
Houston, TX 77030 USA

Abstract

The belief in the hot hand was suggested to be a "cognitive illusion" since no significant evidence was found in the basketball-shooting data to reject the simple binomial model (Gilovich, Vallone & Tversky, 1985). The present study argues that in order to evaluate the validity of human perception and cognition such as the hot hand belief, a data-driven approach is needed to compare multiple alternative models. A hot hand model with nonstationary shooting accuracy was tested and showed significantly better approximation to the data than the binomial model, indicating that the simple binomial model may not be accurate enough to serve as a normative model. This finding suggests that the hot hand might indeed have existed, and weakens the argument that the hot hand belief might be "seeing patterns out of randomness."

The Hot Hand and the Perception of Randomness

The "hot hand" in the game of basketball has received much attention in cognitive psychology because it touches an interesting topic about human perception and cognition of random and non-random events outside the psychological laboratory. A long-lasting debate about whether the hot hand exists, hence, whether the hot hand belief is a valid cognitive activity, was triggered by three articles by Gilovich, Vallone and Tversky (1985), and Tversky and Gilovich (1989a, 1989b) (later "GVT" refers to these three articles as a group, unless specified otherwise). The researchers interpreted the hot hand belief as a manifestation about statistically significant deviations from what is expected by the simple binomial model, namely, nonstationary shooting accuracy or positive dependence in basketball shooting sequences. However, no statistical evidence was found to support such belief. After a number of statistical analyses on a large set of data, the researchers found that actual basketball shooting sequences were "indistinguishable from that produced by a simple binomial model" (Gilovich et al., 1985, p. 297). They concluded, "perhaps, then, the belief in the hot hand is *merely* [italics added] one manifestation of this fundamental misconception of the laws of chance" (Tversky & Gilovich, 1989a, p. 16).

Since GVT, many studies have been carried out to investigate the hot hand in basketball or other sports such as baseball. These studies roughly fell into four categories: a) studies that conducted null hypothesis tests but failed to reject the binomial model (e.g., Adams, 1992; Albright, 1993; Chatterjee, Yilmaz, Habibullah, & Laudato, 2000), (b) studies that raised concerns about the power of significance tests conducted by Gilovich et al. (1985) and Albright (1993) (e.g., Miyoshi, 2000; Stern & Morris, 1993; Sun, 2001, 2003; Wardrop, 1999), (c) studies that proposed alternative models that may support the hot hand belief (e.g., Albert, 1993; Albert & Bennett, 2001; Larkey, Smith, & Kadane, 1989), (d) a study that addressed the adaptive value of the hot hand belief, assuming the accuracy of the binomial model (Burns, 2001).

The present paper takes a step further and examines the accuracy of the simple binomial model in a side-by-side comparison with an alternative model that assumes the existence of the hot hand. The importance of such a comparison is obvious since which model is more accurate would inevitably affect researchers' opinion about the validity of the hot hand belief. As Brunswik (1956) and Simon (1982) suggested, the environment in which human perception and cognition originate and operate must be carefully studied. On one hand, it is possible that the hot hand does not exist and the hot hand belief is another example of misperceptions of randomness outside the psychological laboratory, in addition to many previous findings when random events were clearly defined (e.g., Falk, 1981; Kahneman and Tversky, 1972; Tversky & Kahneman, 1971, 1974; Wagenaar, 1972). On the other hand, it is possible that the hot hand does exist, even with a substantial effect size (e.g., substantial changes in shooting accuracy), and traditional statistical tests are generally low in power thus not capable of detecting the effect. The fact is that a truly random process can produce seemingly non-random "patterns," but a truly non-random process can produce seemingly random events as well. Lopes and Oden (1987) demonstrated that although human subjects sometimes misidentified random events as nonrandom (i.e., false alarms), they could also correctly detect truly nonrandom signals (i.e., correct hits). Thus, it is important to find out whether the simple binomial model is accurate enough to serve as a normative model. Then, researchers might be able to answer the question whether the hot hand belief is more about signal detections, or, just "seeing something out of randomness."

Model-driven vs. Data-driven

GVT concluded that actual basketball-shooting records "may be *adequately* [italics added] described by a simple binomial model" (Gilovich et al., 1985, p. 313). However, such a conclusion was solely based on the non-significant *p* values in null hypothesis tests under the binomial model. Sun (2003) and Wardrop (1999) pointed out that GVT's statistical tests were largely redundant and generally low in power, and in many cases, GVT failed to report large deviations from the binomial process or misinterpreted the test results. In the present paper, I only address the importance of comparing multiple models and why non-significant *p* values do not necessarily suggest the accuracy of the simple binomial model.

Criticisms of null hypothesis significance testing (NHST) have been leveled for decades. Many researchers warned that when alternative hypotheses abound, misinterpretations of statistical significance could easily arise (e.g., Cohen, 1994; Lykken, 1991; Oakes, 1986). Nevertheless, many researchers tend to ignore the fact that NHST only estimates $p(D \mid H_0)$, the probability that data D could have arisen if the null hypothesis H_0 were true, not $p(H_0 \mid D)$, the probability that H_0 is true, given D. In modeling basketball shooting, the fact that no significant deviation was found to reject the binomial model, namely, $p(D \mid H_{\text{Binomial}}) > .05$, only indicates that the binomial model may not be terribly erroneous. However, not being terribly erroneous is not the same thing as being accurate or being unique. A *p* value greater than .05 only prompts researchers to retain the null, not to accept the null as if it were true or even likely to be true.

Let H_{Binomial} denote the event that the binomial model is true, $H_{\text{Hot Hand}}$ denote the event that the hot hand theory is true, and D denote the event that a certain statistic from the shooting data reaches a certain level. In order to demonstrate the adequacy of binomial model or the invalidity of the hot hand theory, given the available data, one needs to find out which hypothesis the data are in favor of, namely, to compare $p(H_{\text{Binomial}} \mid D)$ and $p(H_{\text{Hot Hand}} \mid D)$. In Bayes' theorem,

$$\frac{p(H_{Binomial} \mid D)}{p(H_{Hot\ Hand} \mid D)} = \frac{p(H_{Binomial})\,p(D \mid H_{Binomial})}{p(H_{Hot\ Hand})\,p(D \mid H_{Hot\ Hand})}. \quad (1)$$

If one is not biased toward either one of the two hypotheses before examining the data, it is reasonable to assign equal prior probabilities to both models, $p(H_{\text{Binomial}}) = p(H_{\text{Hot Hand}}) = .50$. Then, the comparison between $p(H_{\text{Binomial}} \mid D)$ and $p(H_{\text{Hot Hand}} \mid D)$ comes down to the comparison between $p(D \mid H_{\text{Binomial}})$ and $p(D \mid H_{\text{Hot Hand}})$. GVT's statistical analyses showed that in a number of statistical tests,[1] $p(D \mid H_{\text{Binomial}})$ was not significantly small. Nevertheless, such information alone cannot invalidate the hot hand theory, another piece of information, $p(D \mid H_{\text{Hot Hand}})$ is still missing.

The argument here actually calls for a data-driven approach that compares at least two rival models, rather than a model-driven approach that conducts null hypothesis tests only on one model. The distinction between these two approaches is not a clear cut but rather a difference in emphasis. The data-driven approach eventually has to come down to evaluations of a limited number of models one by one. If a certain model superior to others arises, it will be tested against further data for a need to abandon or modify the model. In this sense, the distinction between two rival models often is not an absolute dichotomy. It is true that in hypothesis testing, such as in Equation 1, two hypotheses have to be exclusive to each other. Nevertheless, in data modeling, two models might only differ in the degrees they approximate the actual process. Which model is selected would be based on which model provides a better approximation of the data, rather than some "mechanical dichotomous decisions around a sacred .05 criterion" (Cohen, 1994, p. 997).

Extracting Relevant Statistics from the Data

To compare multiple models by a data-driven approach, it is essential to extract relevant statistics from the available data. Sun (2003) pointed out that the statistical tests conducted by GVT, such as the test of serial correlation (compared to zero) and runs test were largely focused on the first moment estimate of the time series, namely, the *hit rate* (i.e., observed hitting percentage in a sequence of a certain length) as an estimate of *shooting accuracy* (i.e., the probability for any given shot to be a hit). However, by the law of large numbers, *hit rate* only provides a good approximation of *shooting accuracy* when shooting accuracy remains constant and the sample size is considerably large. Thus, assuming the hot hand is about the nonstationarity of the shooting accuracy, fluctuations of shooting accuracy would not be easily detected by fluctuations of hit rate, when a player only took a limited number of shots in each game. For instance, given a result of 5 hits in a sequence of 10 shots, a null hypothesis test *alone* cannot distinguish whether the hit rate of 50% is a result of a shooting accuracy of 40% or a shooting accuracy of 60%.

By focusing on higher moments of the shooting sequences, Sun (2003) found significant fluctuations of serial correlations in the field goal data that were originally reported by Gilovich et al. (1985). That is, a player sometimes shot in streaks (i.e., successive hits or misses), such as in {1, 1, 1, 1, 0, 0, 0, 0}, yielding a positive serial correlation, and sometimes shot alternatively (hits and misses alternated very often), such as in {1, 0, 1, 0, 1, 0, 1, 0}, yielding a negative serial correlation. The observed changes in serial correlations were unlikely to be accounted for by the simple binomial model, namely, $p(D \mid H_{\text{Binomial}}) < .05$, where D represents the event that the serial correlations changed significantly. Only when the data were *aggregated* across all periods, the *overall averaged* serial correlation was close to zero (e.g., comparing the overall

[1] Note that most of GVT's tests were mathematically redundant (see Wardrop, 1999).

serial correlation with zero, $p > .05$). This finding has at least two indications. First, the actual basketball shooting might not be a stationary process since hits and misses are *not evenly* distributed in the observed shooting sequence. Second, fluctuations of hit rates and the overall serial correlation are not sensitive enough to capture such nonstationarity. In the following, I will present an alternative model that can be distinguished from the simple binomial model by examining the fluctuations of serial correlations. Furthermore, this model may provide a better approximation to the observed data.

A Model of the Hot Hand

Model and Parameter Settings

In real basketball games, it is very possible that potentially high or low shooting accuracy ("hot hand" or "cold hand") might exist but were *interrupted* by other activities such as shot selection and defensive pressure. For example, after making one or two shots, a player may become confident and try more difficult shots, or the opposing team may intensify their defensive pressure on that player. Less frequent interruptions tend to produce shooting sequences with positive serial correlations, since the player's shooting accuracy, either high or low, remains comparatively unchanged, for example, an extreme case would be something like {1, 1, 1, 1, 1, 0, 0, 0, 0, 0}. And vice versa, more frequent interruptions tend to produce shooting sequences with negative serial correlations, for example, a resulting sequence like {1, 0, 1, 0, 1, 0, 1, 0, 1, 0}.

Figure 1 represents a Markov switching model (hence referred to as "the hot hand model"). Similar models have been used by Lopes and Oden (1987) in studying human subjects' ability of distinguishing between random and nonrandom events, and by Albert and Bennett (2001) in modeling the "streakiness" in baseball.

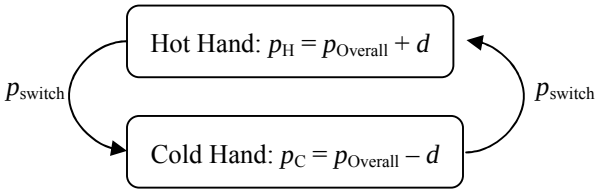


Figure 1. A Markov Model of the Hot Hand

To accommodate the hot hand theory, the major characteristic of this model is that it has two states, "hot hand" and "cold hand," representing two different levels of shooting accuracies, p_H and p_C, respectively. If a player's overall shooting percentage in the entire season was $p_{Overall}$, p_H and p_C were shifted higher or lower in the same amount of d from $p_{Overall}$. Then, this player's simulated shooting sequence will be generated as the player switches between the "hot hand" and the "cold hand." How often the player makes the switch depends on the switching probability, p_{switch}. A high p_{switch} value (e.g., $p_{switch} > .50$) means the player switches between two states very often. In an actual basketball game, this would represent the situation in which a hot hand or a cold hand is detected and a real-time adjustment is immediately deployed by either the player or the opposing team. And vice versa, a low p_{switch} value (e.g., $p_{switch} < .50$) means that the player rarely switches between two states. This would represent the situation in which a hot hand or a cold hand remained uninterrupted or real-time adjustments rarely occurred.

Actually, when $p_H = p_C = p_{Overall}$ ($d = 0$) and $p_{switch} = .50$, the hot hand model is in effect equivalent to the binomial model. If the binomial model were truly adequate and unique, one would expect that a model with dramatically different parameter settings would be less capable of describing the observed data. For this reason, I chose a set of extreme values to represent the hot hand model, in which $d = .30$ (i.e., $p_H - p_C = .60$) and p_{switch} was randomly selected from (.95 and .05) with a 50-50 percent chance for every 10 shots, whereas the binomial model only took a constant shooting accuracy $p_{Overall}$. Figure 2 illustrates the difference between two models in terms of the shooting accuracy along the time line.

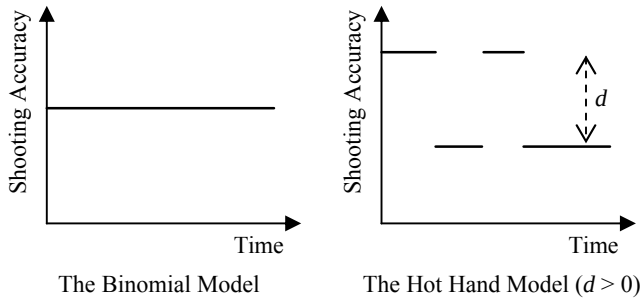


Figure 2. Two Possible Models of Basketball Shooting

Simulation Procedure

Gilovich has kindly provided the field goal data that were reported in Gilovich et al. (1985). There were 18 players in the data set, and 16 of them were included in the simulation (2 players were excluded because their shooting sequences were too short).

For each player in the simulation, I computed a statistic called "MMAC" (Max-Min Moving Autocorrelation) from his actual shooting sequence, whereas MMAC was defined as the absolute difference between the largest and smallest moving serial correlations, where the moving serial correlations were calculated as the serial correlations within a window of 100 shots, starting from the first shot then each time moving 1 shot further until the end of the sequence. The purpose for choosing such specific statistic is to capture the fluctuations of the serial correlations. In the meantime, to reduce chance errors, a large sample size is needed so that the window width of 100 shots was chosen.

For each of the 16 players, I ran 10,000 simulations with the binomial model and another 10,000 simulations with the hot hand model, each simulation generating one shooting

sequence in the same length of the player's actual shooting sequence and with the same overall shooting accuracy. The statistic MMAC was calculated from each simulated sequence, then compared to the observed MMAC from the player's actual shooting record. The probabilities for each model's simulated MMAC to include the observed MMAC were computed as $p(D \mid H_{\text{Binomial}})$ and $p(D \mid H_{\text{Hot Hand}})$. Then, given equal prior probabilities $p(H_{\text{Binomial}}) = p(H_{\text{Hot Hand}}) = .50$, posterior probabilities $p(H_{\text{Binomial}} \mid D)$ and $p(H_{\text{Hot Hand}} \mid D)$ were calculated by Equation 1. Since there were only two hypotheses considered, $p(H_{\text{Binomial}} \mid D) + p(H_{\text{Hot Hand}} \mid D) = 1$.

Simulation Results

The simulation results are listed in Table 1. Columns 2 to 5 list the probabilities $p(D \mid H_{\text{Binomial}})$, $p(D \mid H_{\text{Hot Hand}})$, $p(H_{\text{Binomial}} \mid D)$, and $p(H_{\text{Hot Hand}} \mid D)$, respectively. Column 6 lists the probabilities of detecting significance ($a = .05$, two-tailed) by runs test (Siegel, 1956) on the sequences generated by the hot hand model. The table is ordered in the ascending order of $p(D \mid H_{\text{Binomial}})$.

Considered separately, the probabilities $p(D \mid H_{\text{Binomial}})$ and $p(D \mid H_{\text{Hot Hand}})$ (Columns 2 and 3) in effect provided p values for null hypothesis significance testing, assuming either of the two models as the ture hypothesis ($a = .05$, two-tailed). For players 24, 10, and 3, the simulation results $p(D \mid H_{\text{Binomial}}) < .05$ actually provided significant p values to reject the binomial model. For players 18 and 50, $p(D \mid H_{\text{Binomial}})$ were only slightly greater than .05. (Considering the fact that there were 16 players tested, the probability of family-wise Type I errors needs to be calculated, which was found to be less than .05. see Sun, 2003) On the other hand, none of the p values in $p(D \mid H_{\text{Hot Hand}})$ reached the significance level of .05.

Assuming one is unbiased toward either of the two models prior to examining the data, so that $p(H_{\text{Binomial}}) = p(H_{\text{Hot Hand}}) = .50$, the comparisons between $p(H_{\text{Binomial}} \mid D)$ and $p(H_{\text{Hot Hand}} \mid D)$ (Columns 4 and 5) would reveal which model obtains more support from the observed data in terms of the MMAC statistic.

Table 1. Comparisons between the binomial model and the hot hand model

Player	$p(D \mid H_{\text{Binomial}})$	$p(D \mid H_{\text{Hot Hand}})$	$p(H_{\text{Binomial}} \mid D)$	$p(H_{\text{Hot Hand}} \mid D)$	Power (runs test)
24	.0178	.3380	.0500	.9500	.1911
10	.0223	.3127	.0666	.9334	.1909
3	.0232	.4116	.0534	.9466	.1891
18	.0508	.1299	.2811	.7189	.1836
50	.0690	.4709	.1278	.8722	.1824
7	.1517	.5929	.2037	.7963	.1854
25	.4084	.6539	.3844	.6156	.1836
2	.5343	.9610	.3573	.6427	.1951
11	.5446	.8983	.3774	.6226	.1795
22	.6472	.9640	.4017	.5983	.1769
53	.7094	.9766	.4208	.5792	.1936
5	.7370	.9855	.4279	.5721	.1928
4	.7625	.9953	.4338	.5662	.1918
6	.8004	.9993	.4447	.5553	.1872
1	.9393	.9993	.4845	.5155	.1871
9	.9886	.9999	.4972	.5028	.1845
Mean	.4632	.7297	.3161	.6839	.1872

Note: D represents the event that the simulated MMAC is greater than or equal to the observed MMAC calculated from each player's shooting record. Column 6 is the estimated power of runs test based on detections of significance ($a = .05$, two-tailed) on the simulated sequences by the hot hand model.

For individual cases, MMAC appeared to be substantially in favor of the hot hand model rather than the binomial model for a certain number of players (e.g., players 24, 10, 3, 18, 50, 7, 25, 2, and 11), as $p(H_{\text{Binomial}} \mid D)$ was much smaller than $p(H_{\text{Hot Hand}} \mid D)$ (see Table 1, Columns 4 and 5). One may calculate a c^2 statistic for each player to test the null hypothesis that MMAC is indifferent to either of the binomial model or the hot hand model. However, c^2 statistics tend to be over-sensitive when the expected frequency in a certain cell is too low (for example, the players 6, 1, and 9). The result that all c^2 were significant ($df = 1$, $p < .01$) for all of the 16 players might have overestimated the superiority of the hot hand model.

Taking all 16 players together, the hot hand model appeared to be substantially superior to the binomial model in accounting for the observed MMAC. On the average, $p(D \mid H_{\text{Binomial}}) = .4632$, and $p(D \mid H_{\text{Hot Hand}}) = .7297$. By the criterion of maximum likelihood, given equal priors $p(H_{\text{Binomial}}) = p(H_{\text{Hot Hand}}) = .50$, the observed data seemed to support the hot hand model rather than the binomial model: on the average, the posterior probabilities are $p(H_{\text{Binomial}} \mid D) = .3161$ and $p(H_{\text{Hot Hand}} \mid D) = .6839$.

It may be possible that the hot hand model appeared to be superior to the binomial model only in terms of the statistics of MMAC. To see whether the hot hand model was "truthful" to other observed statistics such as the number of runs, I also conducted a runs test for each simulated sequence by the hot hand model, since out of those 16 players, runs test only detected one significance at the .05 level in the observed shooting sequence (player 53, see Gilovich et al., 1985). (Note that because of the symmetrical setting of the model, there is no need to check the hitting percentage.) The results of runs test suggested that the hot hand model was largely truthful to the observed shooting sequence in the statistic of number of runs, since on the average, only 18.72% of the simulated sequences were detected as significant deviations from what is expected by the binomial model (see Table 4, last column). A further check found that during 10,000 simulations for each player with the hot hand model, the *overall* serial correlations were symmetrically distributed around the mean of zero, with a standard deviation slightly larger than the expected value ($1/\sqrt{N-3}$) assuming binomial process (N is the number of shots in each sequence). Together, these observations provided confirmations to my previous claims. That is, a nonrandom process (such as the hot hand model) can produce seemingly random sequences and may not be easily detected by traditional statistical methods (such as the runs test, or, comparing the overall serial correlation with zero).

Discussion

One might argue that the "hot hand model" fitted the data better than the binomial model simply because the former has more parameters than the latter. I have three reasons to counter this argument. First, basketball shooting is a complex process. It is very reasonable to believe that a useful model needs more parameters than just a single constant shooting accuracy. Second, the extra parameters in the hot hand model may not be counted as "free parameters" because they feasibly represent actual situations in which a player's shooting accuracy may change and real-time adjustments take place quickly (or slowly). Lastly and most importantly, as mentioned before, the hot hand model actually took parameter values that were substantially different from the simple binomial model. Yet, it provided more accurate descriptions of the observed data. This would have seriously challenged the accuracy of the simple binomial model.

It should be pointed out that the primary purpose for building the hot hand model is not to argue about its uniqueness. Nevertheless, such model may prompt researchers to consider the possibility that non-random process may easily produce seemingly random sequences and the possibility that the hot hand belief is indeed a valid cognitive activity in detecting non-random events. It is important to notice that particular statistics such as number of runs, serial correlations, including the MMAC statistic I used in this study, may not be sensitive enough to tell the difference between two different processes. Nevertheless, researchers need to consider multiple models in evaluating the validity of human perceptions, since multiple models can co-exist and provide different levels of approximations to the actual underlying process.

The simulation has shown that for a certain number of players, the hot hand model is substantially superior to the binomial model. For the other players, these two models are not easily distinguishable. By Bayes' theorem in Equation 1, if both models account for the data with the same capability so that $p(D \mid H_{\text{Binomial}}) \approx p(D \mid H_{\text{Hot Hand}})$, which model is more likely to be "perceived" from the data, namely, $p(H_{\text{Binomial}} \mid D)$ and $p(H_{\text{Hot Hand}} \mid D)$, then, is entirely determined by personal beliefs, $p(H_{\text{Binomial}})$ and $p(H_{\text{Hot Hand}})$. There is no prior reason why basketball fans and players should agree with researchers on such personal belief. In other words, the hot hand belief may not be readily dismissed as merely a misperception of randomness simply because the researchers failed to reject the binomial model by null hypothesis significance testing.

General Conclusion

The primary purpose of the present paper is not to dispute whether ordinary people misperceive probabilistic events in basketball, but to prompt further investigations of the actual process of basketball shooting. Lacking normative knowledge such as probability theory and theories of stochastic processes, ordinary people are often prone to mistakes. However, it is also possible that the hot hand belief was describing a true anomaly that was not detected by traditional statistical methods. The present study presented a case when statistical methods are applied objectively rather than subjectively toward the plausible

models, how a different point of view, regarding the validity of human perceptions of the environment, could be obtained. That is, comparing to a model-driven approach that only conducts null hypothesis testing on a single model, a data-driven approach can be more revealing by comparing multiple models. Then, it was suggested that the simple binomial model might not be accurate enough to serve as a normative model in evaluating the validity of the hot hand belief. From Brunswik's (1956) point of view, an organism and the environment in which the organism was embedded should receive equal emphasis in psychological theory and research. In this sense, the primary purpose of the present study is to serve as "a propaedeutic to functional psychology" (Brunswik, 1956, p. 119), a necessary step before psychologists can fully understand the belief in the hot hand.

Acknowledgements

This research was supported in part by the postdoctoral fellowship from the W. M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia.

References

Adams, R.M. (1992). The "hot hand" revisited: Successful basketball shooting as a function of intershot interval. *Perceptual and Motor Skills, 74*, 934.

Albert, J. (1993). A statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association, 88(424)*, 1184-1188.

Albert, J., & Bennett, J. (2001). *Curve ball: Baseball, statistics, and the role of chance in the game*. New York: Springer-Verlag.

Albright, S. C. (1993). A statistical analysis of hitting streaks in baseball. *Journal of the American Statistical Association, 88(424)*, 1175-1183.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Burns, B. D. (2001). The hot hand in basketball: Fallacy or adaptive thinking? In J. D. Moore, & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Meeting of the Cognitive Science Society* (pp. 152-157). Hillsdale, NJ: Lawrence Erlbaum.

Chatterjee, S., Yilmaz, M. R., Habibullah, M., & Laudato, M. (2000). An approximate entropy test for randomness. *Communications in Statistics: Theory and Methods, 29(3)*, 655-675.

Cohen, J. (1994). The earth is round ($p<0.05$). *American Psychologist, 49*, 997-1003.

Falk, R. (1981). The perception of randomness. In C. Comiti, & G. Vergnaud (Eds.), *Proceedings of the Fifth International Conference for the Psychology of Mathematics Education* (pp. 222-229). Grenoble, France.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Larkey, P. D., Smith, R. A., & Kadane, J. B. (1989). It's Okay to believe in the "hot hand." *Chance 2(4)*, 22-30.

Lopes, L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 13(3)*, 392-400.

Lykken, D. L. (1991). What's wrong with psychology? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology, vol. 1: Matters of public interest, essays in honor of Paul E. Meehl* (pp. 3-39). Minneapolis, MN: University of Minnesota Press.

Miyoshi, H. (2000). Is the "hot hands" phenomenon a misperception of random events? *Japanese Psychological Research, 42(2)*, 128-133.

Siegel, S. (1956). *Nonparametric statistics*. New York: McGraw-Hill.

Simon, H. A. (1982). *Models of bounded rationality*, Vol.3, Cambridge, MA: MIT Press.

Stern. H. S., & Morris, C. N. (1993). A statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association, 88(424)*, 1189-1194.

Sun, Y. (2003). *The hot hand revisited: Toward a data-driven approach.* Manuscript submitted for publication.

Sun, Y. & Tweney, R. D. (2001, November). *Detecting the "hot hand": A time series analysis of basketball*. Paper presented at the 42nd Annual Meeting of the Psychonomic Society, Orlando, FL.

Tversky, A., & Gilovich, T. (1989a). The cold facts about the "hot hand" in basketball. *Chance, 2(1)*, 16-21.

Tversky, A., & Gilovich, T. (1989b). The "hot hand": Statistical reality or cognitive illusion? *Chance, 2(4)*, 31-34.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin, 77*, 65-72.

Wardrop, R. L. (1999). *Statistical tests for the hot hand in basketball in a controlled setting* (Tech. Rep.). University of Wisconsin-Madison, Department of Statistics.

Spatial Updating in Intrinsic Frames of Reference

Yanlong Sun (Yanlong.Sun@uth.tmc.edu)
Hongbin Wang (Hongbin.Wang@uth.tmc.edu)
Todd R. Johnson (Todd.R.Johnson@uth.tmc.edu)

The University of Texas Health Science Center at Houston
Houston, TX 77030 USA

Abstract

The present study investigates the properties of the spatial updating in terms of intrinsic frames of reference. We hypothesize that the efficiency of dynamically updating object-to-object relations is based on two main factors, a relatively stable frame of reference provided by the orienting object (or object array), and the behavioral significance (salience) level of the target objects. Three experiments were conducted using tasks of direction pointing. It was found that responses were significantly slower when the orienting object was constantly rotating. Given a relatively stable frame of reference, responses to the salient objects were faster than those to the non-salient objects when the number of salient objects was limited. The salience effect disappeared and re-appeared in the absence and presence of a stable frame of reference, respectively. These findings indicated that spatial updating in intrinsic frame of reference is not automatic and is limited by the number of target objects.

Introduction

As people move through an environment, they continuously update the spatial relations between themselves and the environment and the relations between the objects in the environment. For instance, a pedestrian who is waving on a taxi may also notice that a dog is chasing the taxi from behind. In this scenario, two kinds of information have to be encoded by the pedestrian, the relation between his body and the taxi, and the relation between the taxi and the dog. In fact, this example illustrates the distinction between an egocentric reference system (body-centered) and an allocentric reference system (more specifically in this scenario, an object-centered intrinsic system). It has been generally agreed that in encoding spatial information, different reference systems can be involved. Many researchers adopted the distinction between egocentric and allocentric reference systems and conjectured that participants in their experiments used either one of such systems (e.g, Bryant & Tversky, 1999; Diwadkar & McNamara, 1997; Franklin & Tversky, 1990; Shelton & McNamara, 1997; Sholl & Nolin, 1997; Simons & R. Wang, 1998. For a recent treatment, see McNamara, 2003, and Mou & McNamara, 2002).

A large body of research has been focusing on spatial updating with respect to the egocentric system. It has been indicated that spatial memories are primarily egocentric and updating by the egocentric system is of high fidelity and automatic (e.g., Rieser, 1989; Shelton & McNamara, 1997; Simons & R. Wang, 1998; R. Wang, 1999). Nevertheless, Mou and McNamara (2002) and McNamara (2003) recently proposed that spatial information is encoded primarily of object-to-object spatial relations, and therefore is allocentric. This new theoretical framework calls for a systematic study on properties of spatial updating in intrinsic systems in dynamic situations, as compared to updating in egocentric systems. For example, Sholl and Nolin (1997) and R. Wang (1999) have suggested that egocentric self-to-object spatial relations are updated automatically as people move through an environment. It remains unclear whether updating in intrinsic frame of reference is also automatic. Furthermore, what kind of information is to be updated in intrinsic systems? Are all objects in the environment being updated with equal priorities? The present paper attempts to answer these questions by reporting three experiments.

Our working hypothesis on spatial updating in intrinsic frame of reference is that such a process involves paying attention to both the orienting objects that anchor the intrinsic frame of reference and the target objects in their relations to the orienting objects. In other words, there are two sequential components in such a process: establishing and maintaining a frame of reference, then, updating the object-to-object relations. Thus, we hypothesize that updating in intrinsic systems can be achieved dynamically only when a relatively stable frame of reference can be maintained. In the taxi example above, in order to update the relations between the taxi and the dog, the pedestrian first needs to identify the orientation of the taxi. Second, we hypothesize that updating of object-to-object relations is affected by the behavioral significance of the target objects. This hypothesis is based on previous findings that visual selection can be prioritized by the object's properties, by its specific location and background (e.g., Duncan, 1984; Wolfe, 1994), or even by cues in time (e.g., Watson, Humphreys, & Olivers, 2003). In the taxi example, a dog chasing the taxi probably is more salient than other objects on the street (say, a post stand), thus it is more likely to be

attended to continuously by the pedestrian. We will refer to this effect as the "salience effect" throughout this paper.

We conducted three experiments to test our hypotheses. The task we used was similar to the direction pointing task in the visual map condition in Hintzman, O'Dell, and Arndt (1981). Two major modifications were made to fit our specific needs. First, we added settings to test the salience effect. That is, the target objects had two different salience levels, determined by both behavioral and perceptual significance. Second, to test real-time updating, our experiments were implemented in dynamic settings, which involved continuous relative movement between the orienting object (intrinsic frame of reference) and the target objects. We tested three different movements: the translation-only movement (Experiment 1), the movement in which the orienting object rotated while the target objects remained still (Experiment 2), and the movement in which the orienting object remained still but the target objects rotated (Experiment 3).

General Method

Since all three experiments reported here shared similar settings and procedures, we summarize the common aspects of the experimental settings and data analyses in this section. The experiments were conducted on a Pentium II computer, and the stimuli were presented on a 19-inch CRT monitor. The stimuli consisted of one blue submarine image (bird's eye view) and a certain number of white dots (non-salient objects) and red dots (salient-objects) on a grey background. All three experiments used a 2x8x8x2 within-subjects design, in which there were two salience levels (salient vs. non-salient), eight submarine orientations, eight target directions, and two levels for the number of salient objects. Each participant completed 256 trials, with each trial representing one combination of all levels of all factors. The order of the 256 trials was randomly shuffled. A trial consisted of three steps. First, the submarine and the surrounding dots were presented and the submarine flashed three times to help participants identify its location and orientation. Then, depending on the specific experiment, either the submarine or the surrounding dots started to move (translating or rotating). Finally, the relative movements stopped and at the same time one of the surrounding dots flashed as the target. The instruction for all three experiments was the same. Participants were instructed to imagine being on the submarine and an enemy submarine (the target) was hiding at the location of one of those surrounding dots. The red dots were more likely to be enemy positions thus participants should pay particular attention to them. When the target flashed, participants were told to indicate the direction of the target relative to the orientation of the submarine.

The responses were made with a number keypad on a standard PC keyboard. On the keypad, the number keys 1, 2, 3, 4, 6, 7, 8, and 9 were used as response keys, with each representing one of the eight directions relative to key 5. These keys were re-labeled with drawings of arrows pointing to the corresponding directions. All other keys were removed and the key for number 5 (which was in the center of all eight response keys) was replaced with a stud that could not be pressed. Participants were instructed to use only one finger to press the response keys. At the beginning of each trial, they were told to rest the finger on the stud, and after they made the response, to put the finger back on the stud.

The primary dependent measure was the reaction times (RT) measured in milliseconds. To avoid confusion, we adopted the same labeling scheme as used in Hintzman et al. (1981). Descriptive names were used for responses (target direction relative to the submarine's orientation), such as front, right-front, right, right-back, back, left-back, left, and left-front. For submarine orientation (equivalent to the arrow orientation in Hintzman et al.'s experiment 1 and 2), we used digits 0 through 7, representing the number of steps by 45° clockwise from upright (e.g., digit 0 for the upright submarine orientation, and digit 4 represents the downright orientation).

Experiment 1

Participants Twelve college students and graduate students in the Houston medical center area participated in Experiment 1 (six males and six females, and the average age was 29.3 years with an SD of 4.81 years). Participants were paid for participation.

Procedure Figure 1 shows a typical display in Experiment 1. At the beginning of each trial, one blue submarine image and 400 dots (in which 2 or 4 of them were red and the rest were white) arranged in a 20 x 20 array were presented simultaneously. Red and white dots were the same size, with a diameter of approximately 0.40 cm. The horizontal and vertical distance between every adjacent two dots (hence referred to as one unit) was approximately 0.85 cm, and the submarine image, when upright, was approximately 0.80 cm high and 0.44cm wide. The salient objects were randomly plotted (without overlapping with each other) within a 5 by 5 array in the center of the entire array. The initial position of the submarine was 4, 5 or 6 units (randomly selected) away from the center of the array, randomly taking one of the 8 possible orientations but always approximately pointing to the center of the dot array.

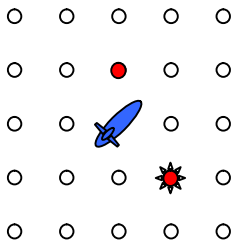


Figure 1

At the beginning of each trial, the submarine flashed three times and then started to move (translation without rotation)

toward the center of the array. The moving speed was constant for all trials, which was approximately 2136 ms per unit (0.47 units per second). When the submarine reached approximately the center of the array, it would stop and at the same time, one of the eight dots in the 3 by 3 square where the submarine was located would flash (the submarine image covered the dot in the center of the square). Participants pressed the response key to respond to the target direction relative to the submarine. The accuracy and reaction times were recorded. A regular experimental session took approximately one and a half hour (in which the training session took approximately 20 minutes).

Results The mean RT for the 12 participants was 1247.4ms with a standard deviation of 542.40ms. The mean accuracy rate was 93.5% with a standard deviation of 4.56%. RT as a function of the target direction is shown in Figure 2, and RT as a function of the submarine orientation is shown in Figure 3, with RT broken down by two salience levels, where error bars represent standard errors. The target names in Figure 2 are abbreviated (F for Front, FR for Right Front, etc.). To emphasize the symmetry and continuity, the direction F in Figure 2 and the orientation 0 in Figure 3 were represented twice. This convention is used in other similar figures throughout this paper.

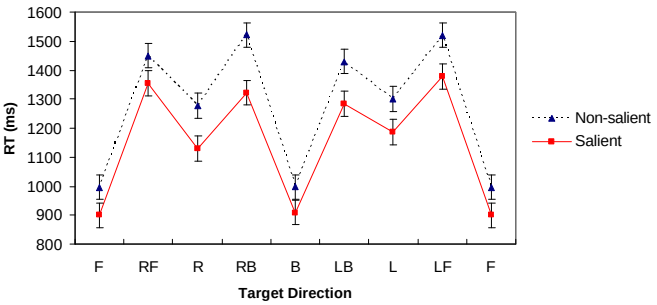


Figure 2. Experiment 1, RT as a function of target direction.

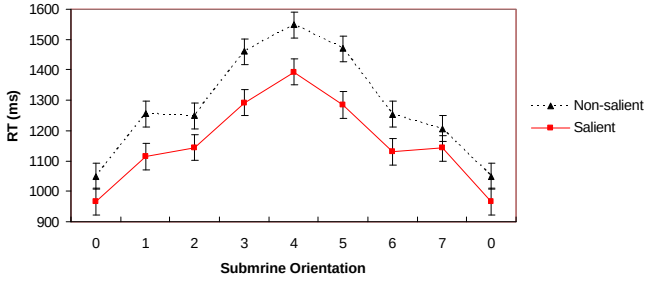


Figure 3. Experiment 1, RT as a function of submarine orientation.

A significant orientation effect was observed. RT as a function of the target direction showed an "M" shaped profile, fastest on the front and back directions and slowest on the right-back and left-back directions ($F_{7, 77}$ = 16.14, p < .001, estimated effect size = .595). As a function of submarine orientation, RT was fastest when the submarine was upright (orientation 0), and slowest when the submarine was pointed down (orientation 4) ($F_{7, 77}$ = 16.74, p < .001, estimated effect size = .603).

Of primary interests to us was the salience effect. It is clear from both Figures 2 and 3 that the salience level had a significant effect on RT. In all target directions and submarine orientations, responses to salient objects were faster than that of non-salient objects, with an average difference of 128.6 ms ($F_{1, 11}$ = 27.95, p < .001, estimated effect size = .718). Moreover, the salience effect appeared to be very stable in size across all target directions and submarine orientations: the two curves in each figure are essentially parallel, and both interactions (salience by target direction, and salience by submarine orientation) were not significant ($F_{7, 77}$ = 1.185, p = .321; and $F_{7, 77}$ = 1.471, p = .190, respectively).

The number of salient objects (hence abbreviated as NOS) appeared to have a small effect on RT. On average, RT was faster when NOS = 2 than when NOS = 4 (mean difference = 26.5 ms), which was marginally significant ($F_{1, 11}$ = 4.57, p = .056). The effect of NOS would be observed more clearly through the interaction between NOS and salience, which was statistically significant ($F_{1, 11}$ = 6.09, p = .031, estimated effect size = .356). It appeared that NOS had little effect on RT when the target was a non-salient object (1314.9 ms when NOS = 2, compared to 1308.3 ms when NOS = 4), but the effect was considerably larger when the target was a salient object (1153.3 ms when NOS = 2, compared to 1212.9 ms when NOS = 4, mean difference = 59.6 ms).

Other factors remaining constant, faster RT for the salient objects in Experiment 1 suggests that spatial information about salient objects was updated with a higher priority thus retrieved more quickly than the information about the non-salient objects. Moreover, the orientation dependence was presented in responses to both salient objects and non-salient objects: both main effects of target direction and submarine orientation were significant but none of the interactions (target direction and salience, submarine orientation and salience, respectively) reached significance, implying the important role the orientation plays in spatial updating.

Furthermore, it is interesting to note that the salience effect remained but reduced in size when the number of salient objects increased (from 161.6 ms when NOS = 2, to 95.4 ms when NOS = 4). One explanation is that more than four salient objects were prioritized but the retrieval of the corresponding information was achieved in a serial fashion. Another explanation is that the capacity of such prioritization was already exceeded when there were four salient objects. Then, participants might randomly choose, say, two of the salient objects for particular attention. As a result, the averaged salience effect in the four salient objects condition was reduced, compared to the two salient objects condition. In either case, it appears the salience effect would eventually disappear when the number of salient objects exceeds a certain level. It would be interesting to conduct

experiments to further investigate the capacity of such a "salience buffer."

Experiment 2

Participants Twelve college students and graduate students in the Houston medical center area participated in Experiment 2 (four males and eight females, and the average age was 26.3 years with an SD of 4.49 years). Participants were paid for participation.

Procedure The procedure was essentially the same as in Experiment 1. The following were the major differences. The potential targets were 8 dots aligned on a circle at 45° intervals, with the submarine located in the center. When the submarine is aligned upright, these eight dots are on the front, right-front, right, right-back, back, left-back, left, and left-front, respectively (see Figure 4). For the two levels of the number of salient objects, instead of 2 salient objects vs. 4 salient objects, Experiment 2 compared the conditions of 1 salient object vs. 2 salient objects.

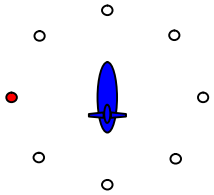


Figure 4

At the beginning of each trial, one blue submarine image and 8 dots (in which 1 or 2 of them were red and the rest were white, on a circle with a diameter approximately of 5.93 cm) were presented simultaneously. All dots were in the same size with a diameter approximately of 0.80 cm. The submarine image, when upright, was approximately 1.98 cm high and 0.88 cm wide. The positions of salient objects were randomly selected (without overlapping with each other). The initial orientation of the submarine was always upright. After flashing 3 times, the submarine started to rotate around the center of the circle. The rotation speed was constant for all trials, which was approximately 0.033° per ms (approximately 2763ms for every 90°). When the submarine reached a certain orientation, it would stop and at the same time one of the eight dots would flash. The rotating distance was determined by the trial settings on the submarine orientation, with a minimum of 45° and a maximum of 360°. Among 256 trials, each 32 trials had the same rotating angle ranging from 45° to 360° in the step of 45°. The order of the trials was randomly shuffled before presentation.

Results The mean RT for the 12 participants was 1772.7 ms with a standard deviation of 663.54ms. The mean accuracy rate was 94.7% with a standard deviation of 6.13%. RT as a function of the target direction is shown in Figure 5, with RT broken down by two salience levels. (RT as a function of the submarine orientation showed the same overlapping pattern. The figure is omitted here to save space.)

The most obvious observation in Figure 5 is the absence of the salience effect: RT for the salient objects was almost identical to that for the non-salient objects. Overall, there was little difference between RT for salient objects and RT for non-salient objects (mean difference = 18.4 ms, $F_{1, 11} = 0.841$, $p = .379$). Moreover, both the main effect of NOS and the interaction of salience and NOS were not significant ($F_{1, 11} = 0.436$, $p = .522$; $F_{1, 11} = 4.190$, $p = .065$, respectively). Though statistically it is impossible to prove the null hypothesis (i.e., the salience effect did not exist), compared to the magnitude of the salience effect observed in Experiment 1, we are confident that the salience effect was at least largely reduced in Experiment 2.

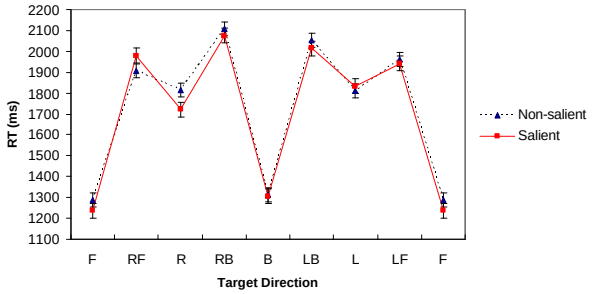


Figure 5. Experiment 2, RT as a function of target direction

Another interesting finding was that RT in Experiment 2 was much slower than that in Experiment 1 for all target directions and submarine orientations. (Figure 6 shows the comparison between all three experiments.) The average difference in RT between Experiment 1 and Experiment 2 was 525.3ms. We suspect that the absence of a stable frame of reference played a major role. Compared to Experiment 1, the major difference in Experiment 2 was that the submarine was rotating constantly before the target was presented. As a result, no stable frame of reference was provided in terms of a fixed submarine orientation. In such a situation, there were two possible strategies of establishing and maintaining a frame of reference. One was that participants could first establish a frame of reference as the submarine was initially presented, then update (i.e., rotate) that frame of reference along with the submarine as it rotated. The other was that participants just waited until the submarine stopped then re-established a frame of reference. We had two reasons to believe that the second strategy was preferred and actually utilized by participants. First, it would take much less effort to re-establish a frame of reference when the submarine stopped rotating, than to maintain a frame of reference by constantly updating it with the rotating submarine. In extreme cases, the submarine would have rotated 360° before it stopped. It would make little sense to update the frame of reference if it would return to its initial position. Second, if the first strategy was actually applied and our participants indeed were updating a frame of reference along with the rotating submarine, Experiment 2 would have had similar RTs as in Experiment 1.

Nevertheless, one may raise the question whether RTs in Experiment 1 and Experiment 2 were directly comparable

since there were confounding factors such as the size of the stimuli and the difference between translation and rotation. For example, it was found that imagined rotation was more difficult than imagined translation (e.g., Presson & Montello, 1994). Due to this consideration, we conducted Experiment 3 with these factors controlled.

Experiment 3

Participants Twelve college students and graduate students in the Houston medical center area participated in Experiment 1 (four males and eight females, and the averaged age was 27.7 years with an SD of 5.55 years). Participants were paid for participation.

Procedure The procedure and device was essentially the same as in Experiment 2. The only difference was that in Experiment 3, the eight surrounding dots were rotating simultaneously while the submarine remained still. Other factors such as the image sizes, the arrangement of the display, and the relative rotation speed, remained the same. The initial orientation of the submarine was randomly selected as one of eight possible orientations. After flashing the submarine three times, the surrounding dots started to rotate around the center of the circle. When they reached a certain location (determined by the trial settings), they would stop and at the same time one of the eight dots would flash.

Results The mean RT for the 12 participants was 1373.1 ms with a standard deviation of 401.68ms. The mean accuracy rate was 93.8% with a standard deviation of 4.52%.

Similar to in Experiments 1 and 2, we observed significant effects of target direction and submarine orientation in Experiment 3 ($F_{7, 77} = 31.590$, $p < .001$, estimated effect size = .742; $F_{7, 77} = 41.075$, $p < .001$, estimated effect size = .789, respectively). Similar to Experiment 1, we observed a significant salience effect. Through all target directions and submarine orientations, responses to salient objects were faster than that to non-salient objects. The average difference in RT for salient objects and non-salient objects was 88.8ms, which was statistically significant ($F_{1, 11} = 16.546$, $p < 0.01$, estimated effect size = .601). (RT as a function of the target direction and a function of the submarine orientation showed the similar split patterns as in Figures 2 and 3. The figures are not shown here.) The salience effect appeared to be very stable across all target directions and submarine orientations: both interactions (salience by target direction, and salience by submarine orientation) were not significant ($F_{7, 77} = 1.123$, $p = .358$; and $F_{7, 77} = 0.999$, $p = .439$, respectively).

The number of salient objects (NOS) showed significant effect on RT. On average, RT was faster when NOS = 1 than when NOS = 2 (mean difference = 79.4 ms, $F_{1, 11} = 32.588$, $p < .001$, estimated effect size = .748). The interaction between NOS and salience was statistically significant ($F_{1, 11} = 42.025$, $p < .001$, estimated effect size = .793). As a result, it appeared that NOS had little effect on RT when the target was a non-salient object (1415.8 ms when NOS = 1, compared to 1419.2 ms when NOS = 2), but the effect was considerably large when the target was a salient object (1251.1 ms when NOS = 1, compared to 1406.4 ms when NOS = 2, mean difference = 155.3 ms). In addition, the salience effect here was larger when there was only one salient object (1415.8 ms compared to 1251.1 ms, with a difference of 164.7 ms), but essentially disappeared when there were two salient objects (1419.2 ms compared to 1406.4 ms, with a difference of 12.8 ms). The diminished salience effect could be due to participants' limited capacity in prioritizing salient objects, or due to the conflicting relations between the two target objects and the orienting submarine (for example, the two salient objects could be on the opposite sides of the submarine). We will leave this question to future investigations.

On average, RT in Experiment 3 was faster than in Experiment 2 (average difference was 399.6ms), but still slower than in Experiment 1 (average difference was 125.7ms) (see Figure 6). This observation confirmed the previous hypothesis that rotation was indeed more difficult than translation. However, it also confirmed the hypothesis that faster reaction times can be produced by a relatively stable frame of reference.

General Discussion

We summarize the general findings in the present study by comparing all three experiments. The experiments had similar task instructions (direction pointing by paying specific attentions to the salient objects), but differed mainly in the forms of relative movements between the orienting object and the target objects. By manipulating the relative movement, we obtained different response times. Figure 6 shows the comparison.

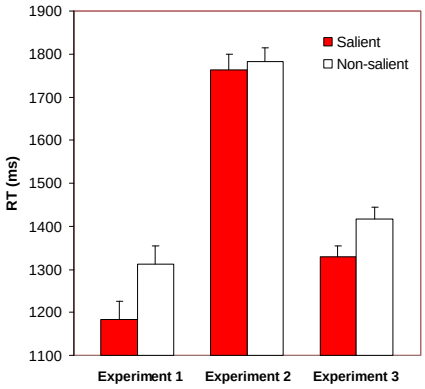


Figure 6. Comparison of RT among three experiments

From a computational point of view, updating of the object-to-object relations in intrinsic frame of reference depends mainly upon two factors: a frame of reference, and a potential target. The different reaction times in the three experiments suggest that a stable frame of reference is critical when the object-to-object relations are to be updated. When the orienting object was rotating, it appeared that the intrinsic frame of reference based on that object was not continuously updated: responses tended to take longer as if a frame of reference had to be re-established. Previous

studies have suggested that egocentric self-to-object spatial relations are updated continuously as people move through an environment (e.g., Sholl & Nolin, 1997; R. Wang, 1999). Our results indicate that maintaining an intrinsic frame of reference is not automatic. The difference might be due to the fact that an egocentric system is relatively easier to be maintained.

Furthermore, we found that given a relatively stable frame of reference, responses for salient objects were significantly faster than those for non-salient objects (Experiments 1 and 3). In addition, the salience effect was largely reduced when a fixed frame of reference was removed (Experiment 2), and re-appeared when a fixed frame of reference was provided (Experiment 3). This observation confirmed our hypothesis that updating of spatial relations can take place dynamically with different priorities when a relatively stable frame of reference is maintained.

Responses to both salient objects and non-salient objects manifested the same orientation dependence in all three experiments, similar to the orientation dependence found in the experiments where egocentric systems were used (e.g., Hinzman, et al.). This similarity might provide an interesting link between the egocentric systems and intrinsic systems. Either participants were imposing an egocentric frame of reference on an external object (e.g., imagine themselves on the orienting submarine), or, as suggested by McNamara (2003), people could in effect treat their bodies as just another object in the space.

Overall, the present study identified several properties of spatial updating in intrinsic frames of reference. In the real world situations, the surrounding environment is constantly changing and people have to adaptively and efficiently prioritize and organize necessary spatial information. Therefore, salient spatial entities, determined by both behavioral and perceptual significance, would receive higher priorities in processing and updating. Furthermore, the current study supports the general claim that multiple reference systems can co-exist in the brain and in the mind to represent space, with each supporting a different class of spatial tasks (H. Wang, Johnson, and Zhang, 2001). For example, while egocentric systems (body-centered) are more convenient for directly supporting motor actions, allocentric systems are more important for representing object-to-object relations in the environment. When a stable allocentric frame of reference is not available, the spatial information will have to be inferred from egocentric information.

Acknowledgements

This study was supported in part by the postdoctoral fellowship from the W. M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia for the first author and the Grant N00014-01-1-0074 from the Office of Naval Research Cognitive Science Program for the second author. We would like to thank Drs. Mike Byrne and Jiajie Zhang for their comments and suggestions at different stages of the research.

References

Bryant, D. J., & Tversky, B. (1999). Mental representations of spatial relations from diagrams and models. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 137-156.

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science, 8(4)*, 302-307.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*, 501-507.

Franklin, N., & Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General, 119*, 63-76.

Hintzman, D. L., O'Dell, C. S., & Arndt, D. R. (1981). Orientation in cognitive maps. *Cognitive Psychology, 13*, 149-206.

McNamara, T. P. (2003). How are the locations of objects in the environment represented in memory? In C. Freksa, W. Brauer, C. Habel, & K. Wender (Eds.), *Spatial cognition III: Routes and navigation, human memory and learning, spatial representation and spatial reasoning* (pp. 174-191). Berlin: Springer-Verlag.

Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *Journal Experimental Psychology: Learning, Memory, and Cognition, 28*, 162-170.

Presson, C. C., & Montello, D. R. (1994). Updating after rotational and translational body movements: Coordinate structure of perspective space. *Perception, 23*, 1477-1455.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 15*, 1157-1165.

Shelton, A. L., & McNamara, T. P. (1997). Multiple views of spatial memory. *Psychonomic Bulletin & Review, 4*, 102-106.

Sholl, M. J., & Nolin, T. L. (1997). Orientation specificity in representations of place. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1494-1507.

Simons, D. J., & Wang, R. F. (1998). Perceiving real-world viewpoint changes. *Psychological Science, 9*, 315-320.

Wang, R. F. (1999). Representing a stable environment by egocentric updating and invariant representations. *Spatial Cognition and Computation, 1*, 431-445.

Wang, H., Johnson, T. R., & Zhang, J. (2001). The mind's views of space. In *Proceedings of the Fourth International Conference of Cognitive Science*, Beijing, China.

Watson, D. G., Humphreys, G. W., & Olivers, C. N. L. (2003). Visual marking: Using time in visual selection. *Trends in Cognitive Sciences, 7*, 180-186.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin Review, 1*, 202-238.

Probabilistic Judgment by a Coarser Scale: Behavioral and ERP Evidence

Yanlong Sun (Yanlong.Sun@uth.tmc.edu)[1]
Hongbin Wang (Hongbin.Wang@uth.tmc.edu) [1]
Yingrui Yang (yangyri@rpi.edu) [2]
Jiajie Zhang (Jiajie.Zhang@uth.tmc.edu) [1]
Jack W. Smith (Jack.W.Smith@uth.tmc.edu) [1]

[1] The University of Texas Health Science Center at Houston, Houston, TX 77030 USA
[2] Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

We hypothesize that statistically unsophisticated people perceive event likelihood with a coarser scale with a limited number of categories, before they report exact numerical probability values. Refinement of the scale beyond a certain level would therefore not improve overall judgmental accuracy and consistency but just impose a heavier burden on their limited computational capacity. An experiment of probabilistic judgment was conducted to test this hypothesis. Results from both behavioral data and event-related potentials in EEG recordings supported our hypothesis.

Introduction

Assessing the likelihood of uncertain events is an essential aspect of human reasoning and decision-making. In the absence of adequate formal models for computing the probabilities, people often rely on intuitions and heuristics to assess uncertainty. The question of how lay people and experts evaluate the probabilities of uncertain events has attracted enormous research interest. (For a historical review, see, e.g., Goldstein & Hogarth, 1996). Proponents of the "heuristics and biases" program argued that intuitive probabilistic judgment is often systematically biased and error-prone (Kahneman, Slovic, & Teversky, 1982). Various violations of normative models, including overconfidence, base-rate neglect, and the conjunction fallacy, have been attributed to applications of a small number of distinctive judgmental heuristics. However, others argued that the human mind has evolved to deal with the structure of the social and physical environment rather than to solve abstract probability problems (Chase, Hertwig, & Gigerenzer, 1998; Gigerenzer et al., 1999; Wang, Johnson, & Zhang, in press). When external information is presented effectively, people can be good intuitive statisticians (e.g., Cosmides & Tobby, 1996). For instance, Gigerenzer and colleagues found that when problems were stated in terms of frequencies instead of probabilities, the stable errors of judgments disappeared (Gigerenzer, 1991; Sedlmeier & Gigerenzer, 2001). For a recent collection of different perspectives on interpreting human intuitive probabilistic judgment, see Gilovich, Griffin, & Kahneman, 2002.

The theoretical claim that the human mind is not adapted to process probabilities has been a magnet for controversy (e.g., Gigerenzer, 1994; Gigerenzer & Hoffrage, 1995; Kahneman & Tversky, 1996). In spite of the centrality of the question, the internal representations and functional neural foundations underlying human probabilistic judgment are poorly understood. Until recently, most research has relied on observations and interpretations of behavioral experiments. In this paper we report a study that investigates the internal representations of human intuitive probabilistic assessment, from both behavioral and neurological perspectives.

Uncertainty Assessment by Approximation

A common scheme in psychological experiments of probabilistic judgment is to ask participants to report probabilities in numerical values, such as the chance of breast cancer in percentage given a positive test result (e.g., Sedlmeier & Gigerenzer, 2001). One question with this scheme, however, is that participants' response of numerical probability values may not genuinely reflect the true internal representation of their likelihood assessment. The following example illustrates this point. When asked to give a numerical estimate of the probability of a certain event provided with probabilities of other events, people often give incoherent answers (e.g., Osherson et al., 2001):

(a) Prob (Clinton is re-elected to the Senate in 2006) = .75
(b) Prob (Giuliani runs for the Senate in 2006) = .5

Participants' response:
(c) Prob (Clinton is re-elected to the Senate in 2006 and Giuliani runs for the Senate in 2006) = .1

Apparently, the numerical estimate of .1 is incoherent with the other two given probabilities, and the correct answer should be at least .25. However, it is possible that participants did not distinguish small increments on the continuous scale of probabilities. The estimate of .1 probably is only an approximation by the idea that the chance for Clinton may drop dramatically if Giuliani joins in the competition, rather than the result of exact

calculations. In other words, the incoherence is likely to be produced by the application of a coarser scale, rather than by some systematically biased heuristic.

We propose two hypotheses on the internal representations when people intuitively assess the event likelihood without exact calculations. First, the task of judgment of probabilities can be partitioned into two separate phases: the internal representation and the response (see Figure 1). The internal representation of event likelihood is the result of an approximate estimation of presented information (cues) on a coarser scale (internal scale). Only when subjects need to report probability values (e.g., in a typical psychological experiment), the internal scale is projected onto a finer continuous scale (response scale). Second, the internal scale has only a limited number of categories that represent different magnitudes of the perceived event likelihood. Together, these two hypotheses allow us to distinguish two fundamentally distinctive types of internal representations underlying human probabilistic judgment. We illustrate each of them in detail next.

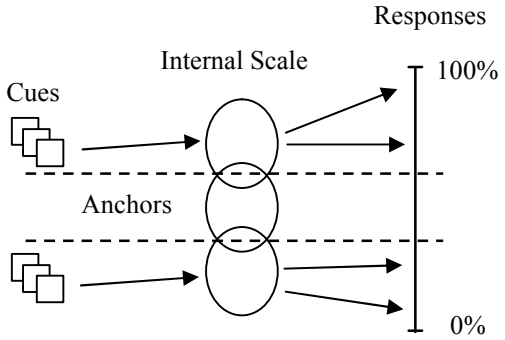


Figure 1. Internal representation on a coarser scale

Internalization by Anchoring and Adjustment

The hypothesis that people's perception of the event likelihood may not be infinitely refined is based on the notion that human beings have limited computational capacity. Miller (1956) suggested that the number of levels of any variable that can be internalized is not only finite but also small. Miller's finding has been widely cited in psychometric research on whether there is an optimal number of response alternatives in designing a scale. Many researchers believe that a further refinement of scales is meaningless if it is beyond human information processing capacity (for a review, see Cox, 1980).

In assessing the event likelihood, the number of categories on the internal scale is determined by the way these categories are formed. Previous studies show that people often use the "anchoring and adjustment" heuristic in judgments of belief and value (Tversky & Kahneman, 1974). Tversky and Kahneman presented anchoring as a process in which "people make estimates by starting from an initial value that is adjusted to yield a final answer" (1974, p. 1128). Chapman & Johnson (2002) provided a review on recent development in understanding the mechanisms of anchoring. One important implication of anchoring is that the number of categories on the internal scale is limited, rather than infinitely refined, since there are only a limited number of anchors that can be processed at the same time (see Figure 1). Note that it is not to suggest that people can not distinguish two very alike events with subtle difference. However, such distinction can only be achieved based on a side-by-side comparison, as one of the two events serves as an anchor.

The anchors can be directly provided by external cues, predefined action thresholds, or by knowledge retrieved from memory (for example, availability heuristic, Tversky & Kahneman, 1974). In the example of incoherent estimate mentioned before, the anchors would be likely provided by the probabilities stated in the first two statements, .75 and .5. Thus, there would be only three categories of likelihood divided by these two numbers, more likely than .75, less likely than .5, and a category in the between. When asked to report a numerical probability estimate to the third statement, the category of the lowest possibility was projected onto a continuous scale As a result, a small number such as .1 was likely to be generated. It is very conceivable, however, that in a different occasion the same person would give an answer of .2 instead of .1 to the same question without changing his or her perception of the event.

In everyday life situations, it is often the case that one of a few options is chosen based on predefined action thresholds rather than on exact probability values. It was found that physicians often significantly overestimated the probability of disease given a positive test result. In one study, 95 out of 100 physicians estimated the probability of breast cancer between 70% and 80% given a positive mammography while the correct answer is merely 7.8% (Eddy, 1982). Nevertheless, such large deviations do not necessarily mean that they are poor physicians. Probably to a physician, what matters most is a dichotomy whether a test result is positive. Consequently, a two-category internal representation (for example, "more likely" and "less likely") would be adopted until further diagnosis is conducted. When the physicians were asked to report an exact numerical value, they simply just obtained a number that would represent the category of the highest likelihood on the internal scale. Thus, a large number was likely to be reported.

Coarser Scales versus Finer Scales

The above argument actually suggests the need to partition the errors of intuitive probabilistic judgment. By distinguishing the internal representation from the numerical responses, we in effect partition the errors into two sources: systematic errors when the external information is internalized onto a coarser scale, such as overconfidence, availability bias, and conjunction fallacy (e.g., Kahneman, Slovic, & Teversky, 1982), and "random errors" when the internal representation are projected onto the continuous scale of numerical values. From Figure 1, it can be seen that a large portion of errors is elicited by the projection of a

coarser scale onto a continuous scale. Thus, a direct comparison between human perception of event likelihood and criteria derived from a normative model very likely has exaggerated the irrationality of human intuitive judgment. In the previous example of incoherent estimate, incoherence would be greatly reduced if a coarser scale is used to evaluate participants' responses.

We reason that in assessing uncertainty, people prefer a coarser scale (e.g., fewer categories) to a finer scale (e.g., more categories) as the internal representation, since the former may function effectively and demand less computational load in a variety of cognitive tasks. People assess the event likelihood only to the extent that it is adequate to reach a conclusion or choose an action. For example, when a person needs to decide whether or not to bring an umbrella to work based on the chance of rain, a chance of 70% and a chance of 80% probably would not make much difference in such a decision. Sometimes a finer judgment can be obtained but such refinement might have little effect to improve the choice. For example, Kareev and colleagues suggested that the limited capacity of working memory could actually help the early detection of covariation (Kareev, 1995; Kareev, Lieberman & Lev, 1997). In Sun & Tweney (2002), researchers found that participants chose their actions based on a heuristic of using small samples, and their performance in the task was very close to that obtained by the optimal strategy.

Our hypothesis that there exists a coarse representation of uncertainty for intuitive probabilistic judgment is consistent with the large body of literature on mental presentations of quantity and numbers. Dehaene and colleagues (Dehaene, 1997; Dehaene et al, 1998) have suggested that there is a coarse and analog mental number line, which is the foundation of a "number sense" and shared by humans and animals. Recent studies using brain-imaging techniques have provided further support for the existence of such coarse scale representations. It has been found that the approximation and exact calculation tasks of large numbers (as compared to rote arithmetic operations) put heavy emphasis on the left and right parietal cortices, which may encode numbers in a non-verbal quantity format (e.g., Dehaene et al., 1999; Pesenti et al., 2000; Stanescu-Cosson et al., 2000).

To test our hypotheses, we conducted an experiment in which participants performed a task of probability estimation. The task was to estimate the winning probabilities of poker hands in a standard "draw poker" game when presented with the highest two cards out of the five cards on a hand. The reason we selected this task is that it provides an objective criterion for evaluating participants' estimates. Most importantly, this task offers a distribution of probabilities ranging from zero to close to 100% with extremely small increments. This feature allows us to look into the refinement of the internal scale when people need to make estimates intuitively. We compared two conditions in the experiment. The coarser-scale condition required probability estimation within an increment of 30% (e.g., less than 30%, 30% ~ 60%, and greater than 60%). The finer-scale condition required probability estimation within an increment of 10%. Both behavioral data and the event related potentials (ERP) in EEG recordings were collected. The experiment results supported our hypotheses. First, estimates by the finer scale showed significantly worse performance in terms of overall accuracy and consistency. Second, significant ERP difference was found over the parietal area between two conditions, indicating that different levels of effort were involved in the application of different internal scales.

Method

Participants Six graduate students in the Houston medical center area participated in the experiment. All participants were right-handed males. The averaged age was 27 years old. None of the participants reported having any in-depth knowledge of probabilistic theories. All participants reported having some experience of playing poker games but only at a novice level.

Procedure The stimuli in the experiment (two-card poker hands) consisted of 18 levels of winning probabilities, ranging from 5% to 90% with a step of approximately 5% (as in a one-deck and two-person game). With suit variations, there were 36 different hands used in the experiment. (Note that in this experiment, the suit variation does not change the winning probability since the highest hand is a one-pair of aces.) Participants were introduced with the rules of hand ranking before the experiment (e.g., a pair is higher than single cards; the Ace is higher than the King, and etc.) without being informed of the corresponding probability values.

Figure 2 shows the procedure of the experiment. Stimuli were displayed on a 17-inch CRT monitor. To reduce eye movements, the viewing angle of the displays on the screen was limited to 2%. After the fixation display, two poker cards were presented. Participants were instructed to form an estimate of winning probability in their mind immediately. After the second fixation, a number was displayed and the participants needed to compare their estimate to the displayed number as quickly as possible. A Microsoft PC mouse was used to collect the responses (the left button for "less than," and the right button for "greater than").

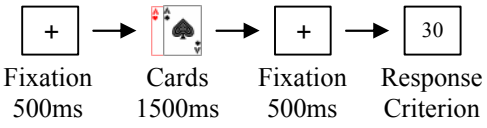


Figure 2. Experiment Procedure

There were two experimental conditions: the coarser-scale condition and the finer-scale condition. In the coarser-scale condition, participants were instructed to estimate probabilities in three categories (less than 30%, 30% ~ 60%, and more than 60%). The number displayed as the

comparison criterion is either 30 or 60, randomly selected. In the finer-scale condition, participants were instructed to form the estimate accurate to the single digit. The comparison criterion was either 5% plus or 5% minus (randomly selected) the true winning probability. For example, a pair of aces has a winning probability of 90%. Then, participants needed to compare their estimates with either 85% or 95%, depending on which number was displayed. Each participant completed both conditions. The order of two conditions was balanced among participants. Each condition consisted of 72 trials in 2 trial list cycles, and each of the 36 poker hands was displayed once in each cycle. The order of trials was randomly shuffled. Response time was recorded as the latency after the onset of comparison criterion. During the experiment, ERPs were sampled at 250 Hz with a 128-electrode geodesic sensor net (GSN) reference to the vertex (Tucker, 1993).

Results

Behavioral Results Table 1 shows the comparison between two experimental conditions on participants' probabilistic estimates in three categories, accuracy, consistency, and response time. Accuracy was calculated as the percentage when participants made the comparison correctly. For example, if the true winning probability was 70 percent, and the number displayed as the comparison criterion was 60, the correct response should be "greater than." (In case of identical numbers, either response was counted as correct.) All participants showed higher levels of accuracy in the coarser-scale than in the finer-scale condition. On average, the accuracy was 76.4% in the coarser-scale condition, and 56.0% in the finer-scale condition, with a difference of 20.4% ($t(5) = 4.576$, $p < 0.01$, two-tailed). Furthermore, we compared participants' accuracy in two trial list cycles in each condition and found little improvement in accuracy over the trials. The low accuracy level in the finer-scale condition was not significantly different from random guess (comparing to the expected value of 50%, $t(5) = 1.836$, $p = .12$, two-tailed), indicating that participants were not able to distinguish winning probabilities of poker hands in the increment of 10%.

To evaluate participants' judgmental consistency, we calculated Pearson correlations in their responses over two identical sets of stimuli (36 poker hands in each set) in each condition. All six participants in the coarser-scale condition and only one participant in the finer-scale condition showed correlations significantly different from zero ($n = 36$). The averaged correlation was .472 in the coarser-scale condition, and .076 in the finer-scale condition, with a difference of .396 ($t(5) = 7.906$, $p = .001$, two-tailed). This finding was consistent with the difference in accuracy between two conditions, supporting the speculation that participants were more likely to make a random guess in the finer-scale condition.

Table 2 Comparing Coarser-scale and Finer-scale

	Coarser-scale	Finer-scale	Difference
Accuracy	76.4% (8.61)	56.0% (8.05)	20.4%
Consistency	.472 (.060)	.076 (.094)	.396
RT	664.1ms (233.75)	798.4ms (298.93)	-123.2ms

N = 6. Standard deviations were listed in parentheses. All three comparisons were significant $p < .05$ (two-tailed).

The response time was also significantly different between two conditions. The averaged RT was 664.1ms in the coarser-scale condition and 798.4ms in the finer-scale condition, and the former was 123.2ms faster that the latter ($t(5) = -3.118$, $p < 0.05$, two-tailed). Since the response time was recorded as the latency after the onset of comparison numbers rather than the onset of poker cards, it is not clear whether probability estimation or number comparison, or both, produced the difference. Previous studies found that the time to make magnitude comparisons decreases linearly as the numerical distance between two numbers (e.g., Moyer & Landauer, 1967). Nevertheless, the large RT difference in our experiment indicated that the task of probability estimation and comparison as a whole might take more efforts in the finer-scale condition than in the coarser-scale condition.

ERP Results We rejected trials with voltages exceeding ± 100 μV. The remaining trials were segmented then averaged in synchrony with stimulus onset (display of poker cards) in a window of 1100ms (100ms before and 1000ms after stimulus onset), digitally transformed to an average reference, band-pass filtered (0.5 to 20 Hz), and corrected for baseline over 100ms before stimulus onset. Experimental conditions (Coarser vs. Finer) were compared on the 10 central-parietal electrodes by a repeated-measure ANOVA. We found significant ERP difference between two conditions at 400 ± 20ms (peak values) following stimulus onset (Greenhouse-Geisser $F(1,5) = 12.511$, $p <0.05$), where the finer-scale condition yielded more positive voltages over parietal electrodes. Figure 3 shows the wave forms of electrode CP1 (GSN 38) and the voltage difference map (finer-scale minus coarser scale) by spherical spline interpolation. The comparison between the left and right hemispheres over the parietal area was not significant (Greenhouse-Geisser $F(1,5) = 5.339$, $p = .127$). At the current stage of the study, we have not found significant voltage differences over other brain areas. Furthermore, examinations on latency did not reveal any significant differences between two experimental conditions.

Examination of the waveforms showed that the ERP difference occurred after the P300 component, when participants were viewing identical displays and had not yet received the comparison stimuli. The P300 and its sub-components p3a and p3b have been considered as a process that indexes the ensuing memory storage operations, as

P300 amplitudes were found related to memory of previous stimulus presentations (e.g., Fabiani, Karis, & Donchin, 1990; Johnson, 1995; Paller, McCarthy, & Wood, 1988; for a recent review, see Polich, 2003). Our results are also consistent with those of Dehaene and colleagues (Dehaene et al., 1999; Naccache & Dehaene, 2001), who showed that 200-400ms after the stimulus onset is critical to distinguish among different numerical operations with distinctive semantic implications. Based on this observation, we speculate that the difference in ERPs in our experiment probably was due to different working memory load. Specifically, when judging the winning chance of a certain poker hand, other poker hands (either from previous trials or by temporary construction) were used as anchors to build categories on the internal representation. Fewer anchors were needed in the coarser-scale condition. On the contrary, the finer-scale condition demanded more effort because more hands needed to be considered at the same time. This speculation is consistent with the participants' oral report after the experiment. For example, one participant reported that he had to think of more hands with "nearby" rankings in the finer-scale condition.

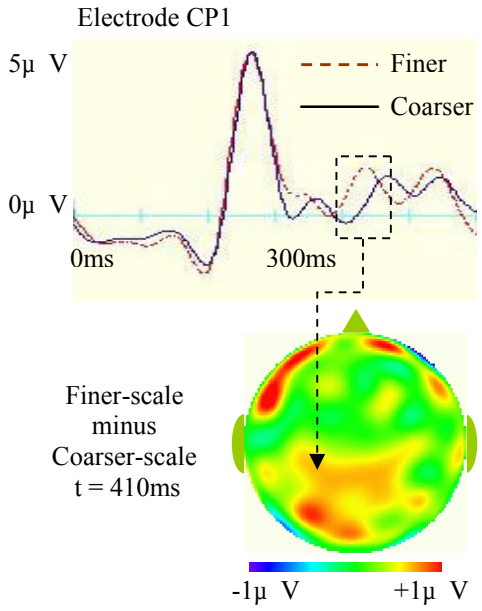


Figure 3

General Discussion

In sum, the findings in our experiment were consistent with our hypotheses of the internal representation of a coarser scale in intuitive judgment of event likelihood. The analyses of behavioral data suggested that if participants were only able to distinguish event likelihood with a coarser scale, a large portion of errors would be produced when they were forced to estimate probabilities with a finer scale. This kind of errors was manifested in both judgmental accuracy and consistency. The low accuracy level (close to random guess) and the low consistency level (close to zero) in the finer-scale condition indicated that participants were not able to distinguish winning probabilities of poker hands in the increment of 10%. Therefore, if the internal scale indeed had a limited number of categories, most likely this number was not greater than 10. It is interesting to point out that in our experiment, there were 36 different hands at 18 equally distributed levels of winning probabilities. If these hands were presented *externally* at the same time in the order of their rankings, it is reasonable to assume that any participant can report probability values accurate to the 5% increment by anchoring and adjustment. Nevertheless, their poor performance in the finer-scale condition indicated that the number of anchors that can be processed *internally* was quite limited. Furthermore, the ERP difference occurred when participants were viewing identical displays, indicating that participants followed the experimental instruction in forming their estimates at different levels of refinement. And estimating by a finer scale appeared to require more computational effort.

Note that the present study is still at its preliminary stage and there are many questions left to be answered. For example, more replications are needed and the ERP analyses can be extended such as comparisons over other brain areas and source localization. It would also be interesting to test someone who is an expert at poker and to see whether the performance would be better, especially in the finer-scale condition. Another example is that the model of a coarser internal scale can be further examined by manipulating the external representations, as previous studies indicated the important roles of the interaction between internal and external representations in human numerical cognition (e.g., Zhang & Norman, 1995; Zhang & Wang, in press). Upon further experiments and analyses, we believe that the present study will provide a better understanding of human intuitive probabilistic judgment.

Acknowledgements

This research was supported in part by the postdoctoral fellowship from the W. M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia for the first author.

References

Chapman, G. B., & Johnson, E. J., (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120-149). New York: Cambridge.

Chase, V. M., Hertwig, R., Gigerenzer, G., (1998). Visions of rationality, *Trends in Cognitive Sciences, 2(6)*, 206-214.

Cosmides, L. Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.

Cox, III, E. P., (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407-422.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L., (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neuroscience, 21*, 355-361.

Dehaene, S., Spelke, E., Pinel, P. Stanescu, R., Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284*, 970-974.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 249-267). New York: Cambridge University Press.

Fabiani, M., Karis, D., & Donchin, E. (1990). Effects of mnemonic strategy manipulation in a von Restorff paradigm. *Electroencephalography and Clinical Neurophysiology, 75*, 22-35.

Johnson, R. (1995). Event related potential: Insights into the neurobiology of memory systems. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology*, Vol. 10, (pp. 135-163). Amsterdam: Elsevier.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582-591.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56*, 263-269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126(3)*, 278-287.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology*, Vol.2, (pp. 83-115). Chichester, England: Wiley.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability*. New York: Wiley

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gigerenzer, G., Todd, P. M. & the ABC Research Group (Eds.). (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). Heuristics and biases: *The psychology of intuitive judgment*. New York: Cambridge.

Goldstein, W. M, & Hogarth, R. M. (1996). Judgment and decision research: Some historical context. In Goldstein, W. M & Hogarth, R. M. (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (pp. 3-65). New York: Cambridge University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*, 1519-1520.

Naccache, L., & Dehaene, S. (2001). The priming method: Imaging unconscious repetition priming reveals an abstract representation of number in the parietal lobes. *Cerebral Cortex, 11*, 966-974.

Osherson, D., Lane, D., Hartley, P., & Batsell, R. (2001). Coherent probability from incoherent judgment. *Journal of Experimental Psychology: Applied, 70(1)*, 3-12.

Paller, K. A., McCarthy, G., & Wood, C. C. (1988). ERPs predictive of subsequent recall and recognition performance. *Biological Psychology, 26*, 269-276.

Pesenti, M., Thioux, M., Seron, X., Volder, A. (2000). Neuroanatomical substrates of Arabic number processing, numerical comparison, and simple addition: A PET study. *Journal of Cognitive Neuroscience, 12(3)*, 461-479.

Polich, J. (2003). Theoretical overview of P3a and P3b. In J. Polich (Ed.), *Detection of change: Event-related potential and fMRI findings* (pp. 83-98). New York: Kluwer Academic Publishers.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General, 130(3)*, 380-400.

Stanescu-Cosson, R., Pinel, P., Moortele, P., Bihan, D. L., Cohen, L., & Dehaene, S. (2000). Understanding dissociations in dyscalculia: A brain imaging study of the impact of number size on the cerebral networks for exact and approximate calculation. *Brain, 123*, 2240-2255.

Sun, Y. & Tweney, R. D. (2002). Detecting the local maximum: The adaptive significance of a simple heuristic. In C. Schunn & W. Gray (Eds.), *Proceedings of the twenty-fourth annual conference of the Cognitive Science Society*, (pp. 856-860). Mahwah, NJ: Lawrence Erlbaum Associates.

Tucker, D. M. (1993). Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalography and Clinical Neurophysiology, 87*, 154-163.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wang, H., Johnson, T. R., & Zhang, J. (in press). The order effect in human abductive reasoning: An empirical and computational study. *Journal of Experimental and Theoretical Artificial Intelligence*.

Zhang, J., & Norman, D. A. (1995). A representational analysis of numeration systems. *Cognition, 57*, 271-295.

Zhang, J., & Wang, H. (in press). The effects of external representations on numerical tasks. *Quarterly Journal of Experimental Psychology*.

Accounting for Similarity-Based Reasoning within a Cognitive Architecture

Ron Sun (rsun@rpi.edu)
Cognitive Sciences Department
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Xi Zhang (xzf73@mizzou.edu)
Department of CS
University of Missouri
Columbia, MO 65211, USA

Abstract

This work explores the importance of similarity-based processes in human everyday reasoning, beyond purely rule-based processes prevalent in AI and cognitive science. A unified framework encompassing both rule-based and similarity-based reasoning may provide explanations for a variety of human reasoning data.

The paper implements this analysis in a cognitive architecture CLARION, which has previously succeeded in capturing a variety of human learning data in simulations. The exploration of similarity-based reasoning in this architecture leads to a more complete and more comprehensive framework of human reasoning and learning. The simulation within this architecture accurately captures human reasoning data, including numerical measures and verbal protocols. This work demonstrates the significant role played by similarity-based reasoning. Furthermore, it demonstrates how such a reasoning process falls out of the existing structure in the cognitive architecture CLARION.

Introduction

What is human everyday reasoning like? Is it suitably captured by formal models developed by logicians and AI researchers? Or is it different? What are its similarities and differences to these models? After all, computationally speaking, what are the essential patterns in such reasoning?

In this paper, we will attempt to describe some data of human everyday (i.e., mundane or "commonsense") reasoning in computational terms. We will instantiate our analysis in the form of a computational model implemented in a generic cognitive architecture — CLARION (Sun 2002).

A little background is in order here. Sun (1991) proposed a theory of human everyday reasoning based on a combination of rule-based reasoning and similarity-based reasoning, implemented with a mixture of localist and distributed connectionist models. This theory was further developed and elaborated in Sun (1995). The basic tenet of this theory is that, to a significant extent, human everyday reasoning may be described by a combination of rule-based and similarity-based reasoning. Human everyday reasoning may be reduced to these two types of processes. The intermixing of rule-based and similarity-based reasoning can lead to complex patterns of inferences as commonly observed in human everyday reasoning. And these two types of processes may be captured within a unified connectionist model; that is, they fall out of the very same model (albeit with a combination of localist and distributed representations).

The theory was backed up by psychological evidence in the form of verbal protocols as in Collins (1978) and Collins and Michalski (1989). In Sun (1995), these protocols were analyzed based on two mechanisms: rules and similarity (Tversky 1977, Hahn and Chater 1998). The analysis showed that vast majority of the protocol data might be easily captured by the intermixing of these two mechanisms. This theory was crystallized into a two-component model whereby rule-based reasoning was carried out in one component with localist representation, and similarity-based reasoning in another with distributed representation (Sun 1995). Relevant to this approach, Sloman (1993) published a set of experiments, which provided support to the hypothesis of Sun (1991) (see also Sun 1995). He found that similarity played a significant role in determining outcomes of inductive reasoning and similarity might be characterized by feature overlapping (as in Sun 1991). Five years later, Sloman (1998) described further experiments that again supported the hypothesis that there were two parallel mechanisms at work in human everyday reasoning (Sun 1991).

In the remainder of this paper, we first describe the three pertinent experiments of Sloman (1998), which were consistent with the theory advanced in Sun (1991) and Sun (1995). We then describe the generic cognitive architecture, CLARION, used in capturing human everyday reasoning. Next, the particular setup of the architecture for capturing this set of human experiments is described. We then describe the results from simulating the experiments of Sloman (1998) using CLARION. Finally, some general discussion completes the paper.

The Categorical Inference Task

Let us examine some human reasoning data that illustrates combinations of similarity-based and rule-based reasoning (SBR and RBR, respectively). We will look into the data from experiments 1, 2, 4, and 5 of Sloman (1998), which are most relevant to this issue.

Among them, according to our interpretation, although experiment 1 used forced choice while experiment 2 used rating of argument strength, both involved SBR to a very significant extent. Experiment 4 involved explicit use of categorical relations, and thus mainly RBR. Experiment 5 involved more of SBR, as well as RBR.

Specifically, in experiment 1, subjects were given pairs of arguments, either in the form of *premise specificity*:

a. All flowers are susceptible to thrips. $\Longrightarrow$ All roses are susceptible to thrips.

b. All plants are susceptible to thrips. $\Longrightarrow$ All roses are susceptible to thrips.

or in the form of *inclusion similarity*:

a. All plants contain bryophytes. $\Longrightarrow$ All flowers contain bryophytes.

b. All plants contain bryophytes. $\Longrightarrow$ All mosses contain bryophytes.

Subjects were to pick the stronger of the two arguments from each pair. 73 subjects were tested and each was given 18 pairs of arguments (among other things not related to this task).

The results showed that the more similar argument from each pair of arguments was chosen 82% of times (for inclusion similarity) and 91% of times (for premise specificity). t tests showed that these percentages were significantly above chance, either by subjects ($t(72) = 18.64$ and $t(72) = 33.09$ for premise specificity and inclusion similarity, respectively; $p < 0.0001$) or by argument pairs ($t(8) = 6.97$ and $t(8) = 15.61$ respectively; $p < 0.0001$). We note that, if only RBR had been used, then similarity should not have made a difference, because the conclusion category was contained in the premise category and thus both arguments in each pair should have been equally, perfectly strong. Therefore, the data suggest that SBR was involved to a significant extent.

In experiment 2, subjects were instead asked to rate the likelihood ("conditional probability") of each argument. Ratings could range from 0 to 1. 18 subjects were tested.

The mean rating was 0.89 for inclusion similarity and 0.86 for premise specificity. Both were significantly below 1, both by subjects ($t(17) = 2.75$ and $t(17) = 3.23$ respectively; $p < 0.01$), and by arguments ($t(17) = 8.87$ and $t(17) = 6.14$ respectively; $p < 0.0001$). Again we note that it would have been the case that the outcome was 1 if only RBR had been used (because the conclusion category was contained in the premise category). Thus, SBR was significantly present here too. Indeed, ANOVA showed that across subjects, there was a significant main effect of similarity (low vs. high; $F(1, 17) = 18.90, p < 0.001$). So was the case across argument pairs ($F(1, 16) = 12.64, p < 0.001$).

In experiment 4, subjects were asked to rate the likelihood of each argument. Ratings could range from 0 to 1. However, in this case, each category inclusion relation was specifically presented as part of each argument. For example,

All plants contain bryophytes. All mosses are plants. $\Longrightarrow$ All mosses contain bryophytes.

The results showed that the mean judgment was 0.99. 23 out of 27 subjects gave all 1's. 32 out of 36 arguments received judgments of all 1's (excluding one individual who gave 0.99 throughout). In other words, the similarity-based phenomena almost disappeared. Instead, an explicit RBR mode based on category inclusion relations was used.

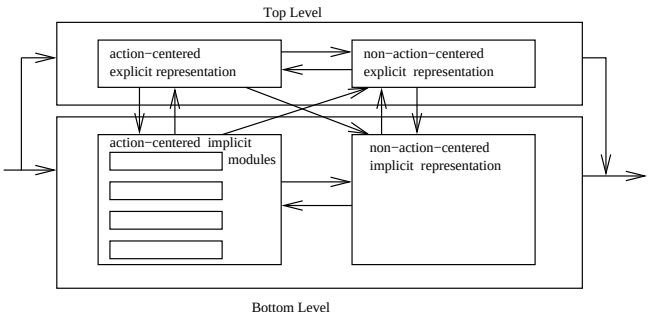


Figure 1: The CLARION architecture.

Experiment 5 was similar to experiment 2, in that ratings were obtained. However, before any ratings were done, subjects were asked to make category inclusion decisions. Thus, in this case, subjects were reminded of rule-based reasoning explicitly involving category inclusion relations. Therefore, they were more likely to use RBR, although probably not as much as in experiment 4, due to the separation of category inclusion judgment and argument likelihood rating in the experiment procedure (unlike that of experiment 4).

The results showed that no one of the 18 subjects gave a likelihood judgment of 1 for every argument, indicating SBR was probably at work. Compared with experiment 2, having subjects make category inclusion judgments increased the likelihood ratings. The mean judgment for experiment 5 was 0.92 as opposed to 0.87 for experiment 2. [1] This increase might reflect the increased involvement of RBR. Nevertheless, ANOVA showed a significant effect of similarity (low vs. high), across subjects ($F(1, 17) = 9.33, p < 0.01$), and across argument pairs ($F(1, 16) = 11.42, p < 0.01$).

Below, we will utilize this task of categorical inference for the further testing of cognitive architecture CLARION. The simulation shows indications of the significance of similarity-based reasoning (as opposed to probabilistic or Bayesian reasoning; cf. Anderson and Lebiere 1998).

The CLARION Model

CLARION is an integrative model with a dual representational structure (Sun et al 2001, Sun 2002). It consists of two levels: the top level captures *explicit* processes and the bottom level captures *implicit* processes. See Figure 1.

First, the inaccessible nature of implicit knowledge is suitably captured by subsymbolic distributed representations provided by a backpropagation network. This is because representational units in a distributed representation are capable of accomplishing tasks but are

[1] However, the difference was not statistically significant by subjects, although significant by arguments ($t(35) = 3.81, p < 0.0001$).

subsymbolic and generally not individually meaningful (see Smolensky 1988, Sun 1995). This characteristic of distributed representation accords well with the (direct) inaccessibility of implicit knowledge.

In contrast, explicit knowledge may be captured in computational modeling by a symbolic or localist representation (Clark and Karmiloff-Smith 1993), in which each unit is more easily interpretable and has a clearer conceptual meaning. This characteristic captures the property of explicit knowledge being (directly) more accessible and more manipulable (Smolensky 1988, Sun 1995).

This radical difference in the representations of the two types of knowledge leads to a two-level model whereby each level using one kind of representation captures one corresponding type of process, either implicit or explicit. The model may select to use one level or the other, based on current circumstances (e.g., experimental conditions; see Sun 2002 for details). When both levels are used, the outcome from the two levels may be combined in some ways, which may be partially domain specific (Sun 2002).

At each level of the model, there may be multiple modules, both *action-centered* modules and *non-action-centered* modules (Schacter 1990, Moscovitch and Umilta 1991). The reason for having both action-centered and non-action-centered modules (at each level) is because, as it should be obvious, action-centered knowledge (roughly, procedural knowledge) is not necessarily inaccessible (directly), and non-action-centered knowledge (roughly, declarative knowledge) is not necessarily accessible (directly). Although it was argued by some that all procedural knowledge is inaccessible directly and all declarative knowledge is directly accessible, such a clean mapping of the two dichotomies is untenable in our view. We will refer to these two sets of modules as the *action-centered subsystem* (the ACS) and the *non-action-centered subsystem* (the NACS), respectively. There are also other components, such as working memory, episodic memory, etc., which are not important to this work.

In this work, we will focus on the NACS, due to the declarative nature of the task. This subsystem, as stated earlier, consists of (1) a top level (known as the GKS, or the general knowledge store), which is made up of a set of chunks and a set of explicit associative rules linking chunks, and (2) a bottom level (known as the AMNs, or the associative memory networks), which is made up of implicit associative memories (Sun 2002).

At the top level of the NACS, the essential elements are *chunks*, each of which is specified by a set of dimension-value pairs (i.e., attribute-value pairs) that describes an entity (or an object), along with a chunk label. Each chunk is represented by a chunk node, which is linked to the nodes at the bottom level (the AMNs) representing the individual dimension-value pairs involved.

The support for the conclusion of an associative rule, which is a chunk, is calculated as follows (Sun 1994):

$$S_j^a = \sum_i S_i^c * W_i^a \qquad (1)$$

where j indicates the jth rule at the top level, S_j^a is the support for associative rule j, S_i^c is the strength of the ith chunk in the condition of the rule, i ranges over all the chunks in the condition of rule j, W_i^a is the weight of the ith chunk in the condition of rule j (which, by default, is $W_i^a = 1/n$, where n is the number of chunks in the condition of the rule).

The conclusion chunk has a strength level that is determined by the maximum of all the support from all the relevant rules:

$$S_{c_k}^c = \max_{j:\text{all associative rules leading to } c_k} S_j^a \qquad (2)$$

where $S_{c_k}^c$ is the strength of chunk c_k (resulting from associative rules), and j ranges over all the associative rules pointing to c_k.

In addition, similarity-based reasoning falls out of knowledge encoding with chunks (i.e., with sets of dimension-value pairs). A known (given or inferred) chunk is automatically compared with another chunk. If their similarity is high enough, then the other chunk is inferred. The strength of a chunk c_i as the result of similarity-based reasoning is:

$$S_{c_i}^c = \max_j (S_{c_j \sim c_i} \times S_{c_j}^c)$$

where $S_{c_j \sim c_i}$ measures the similarity from c_j to c_i (Tversky 1977), $S_{c_j \sim c_i} \times S_{c_j}^c$ measures the support to c_i from the similarity, and j ranges over all the chunks.

The default similarity measure (Sun 1995, Tversky 1977) is:

$$S_{c_1 \sim c_2} = \frac{N_{c_1 \cap c_2}}{f(N_{c_2})}$$

where $S_{c_1 \sim c_2}$ denotes the similarity from c_1 to c_2. $N_{c_1 \cap c_2}$ is the weighted sum of the identically valued dimensions in c_1 and c_2 (among all the specified dimensions of c_2 — the dimensions that have specified values). That is, $N_{c_1 \cap c_2} = \sum_{i \in c_2 \cap c_1} W_i^{c_2} \times A_i$, where A_i is the strength of the value of dimension i in chunk c_1, which is normally 1 (representing full strengths). The weights ($W_i^{c_2}$) in the weighted sum are specified with respect to c_2 (the target of similarity, not the source of it). Normally, these weights are the same and equal to 1. N_{c_2} is the weighted sum of the specified dimensions (the dimensions that have specified values) of c_2. That is, $N_{c_2} = \sum_{i \in c_2} W_i^{c_2} \times A_i$, where normally $A_i = 1$ and $W_i^{c_2} = 1$. f is a super-linear, but close to linear, function (such as $f(x) = x^{1.0001}$ as in our simulation of this task). [2] For further details, see Sun (1995).

Similarity is automatically computed whenever reasoning involves multiple chunks that are similar to one another. Therefore, there is no dedicated representation of similarity between any two chunks.

Similarity-based and rule-based reasoning can be inter-mixed. When both SBR and RBR are employed, we have:

$$S_{c_i}^c = \max(c_{14} \times \max_{j:\text{all rules leading to } c_i} S_j^a,$$

[2] Similarity is thus limited to $[0, 1)$.

$$c_{15} \times \max_{j:\text{all chunks similar to } c_i} (S_{c_j \sim c_i} \times S^c_{c_j}))$$

where c_{14} and c_{15} are two constants that balance the two measures (rule versus similarity), and $S_{c_j \sim c_i}$ is the similarity measure.

As a result of mixing SBR and RBR, complex patterns of reasoning can emerge. As explicated in Sun (1995), the conclusion from one step of reasoning can be used as the starting point of the next step. The iterative process of combined rule-based and similarity-based reasoning allows all possible conclusions to be reached (including "inheritance" reasoning; Sun 1995). These different sequences together capture essential patterns of human everyday reasoning (see Sun 1995 for details).

Note that all of the operations of the non-action-centered subsystem are under the control of the action-centered subsystem, which makes action decisions each step of the way. To do so, the top level of the ACS consists of a set of explicit action rules, either externally given or extracted from the bottom level (from implicit knowledge), while the bottom level consists of implicit decision networks (trained with reinforcement learning algorithms, negligible in this task). For details regarding the ACS and its parameters, see Sun et al (2001) and Sun (2002). We will not get into these details here, as they are not directly relevant to this work.

It is worth noting that CLARION has been successful in simulating a variety of cognitive tasks. These tasks include serial reaction time tasks, artificial grammar learning tasks, process control tasks, alphabetical arithmetic tasks, and the Tower of Hanoi task (Sun 2002, Sun and Zhang 2004). In addition, we have done extensive work on a complex minefield navigation task (Sun et al 2001, Sun and Peterson 1998). We are now in a good position to extend the effort to the capturing of a wide range of human reasoning and memory processes, through simulating reasoning and memory task data. This paper is but one aspect of this effort.

Simulation Setup

At the top level of the NACS (i.e., the GKS), all relevant category inclusion relations, such as "flowers are plants" or "mosses are plants", were encoded as associative rules. Chunk nodes in the GKS were used to represent the concepts involved, such as "flowers" and "plants". The dimensional values of these chunks were represented as separate nodes in the AMNs, and thus the chunk nodes were linked to the AMNs.

For simulating various experimental settings, the following manipulations were used: For simulating settings where SBR was dominant, RBR was de-emphasized. For simulating settings where RBR was dominant, RBR was emphasized. The relative emphasis of the two methods was accomplished through the *balancing* parameters. We set $c_{14} = 0.5$ and $c_{15} = 1.0$ for experiments 1 and 2, because of the heavy reliance on SBR as opposed to RBR as suggested by the analysis of the human data (see the earlier discussion of the human data). For simulating experiment 4, they were set at $c_{14} = 1.0, c_{15} = 1.0$, because this setting prompted more reliance on RBR as indicated by the human data. For simulating experiment 5, they were set at $c_{14} = 0.88, c_{15} = 1.0$, because the experiment involved an intermediate level of reliance on RBR as suggested by the human data. In all, these values were set in accordance with our interpretations of what happened under these different experimental conditions respectively.

At the bottom level of the NACS (the AMNs), although the associative memories were present, they were not very relevant for the performance of this task, because there was no sufficient prior training of the network with any data directly relevant to this task. [3]

Training of the model, before the simulation of the experimental test, consisted of presenting categorical features (dimension-value pairs) along with the category labels, to both levels of the NACS. The features (dimension-value pairs) captured similarities between entities. That is, if A was more similar to C than B was, then A would have more features in common with C than B would. And so on. Note that repeated presentations were not required. The one-pass presentation enabled the formation of chunks and associative rules in the GKS, but not much implicit knowledge in the AMNs. With a proper process of chunk encoding and associative rule encoding as in CLARION, one-pass presentation was sufficient for the GKS.

During test, when a category name was given, the category name was matched with a corresponding chunk label. The matching chunk was activated to the full extent (i.e., 1). Then, through associative rules as well as through similarity-based processes, conclusion chunks were also activated (to varying extents). Conclusion chunks were retrieved along with their strengths, combining SBR and RBR according to the balancing parameters.

For simulating ratings of conclusions (as in experiments 2, 4, and 5), the strengths of chunks derived from a proper combination of the results of SBR and RBR (as determined by the balancing parameters) were directly used. However, for simulating forced choices (as in experiment 1), a stochastic decision process based on the Boltzmann distribution was used to select between two possible outcomes.

Simulation Results

We simulated the data from experiments 1, 2, 4, and 5 of Sloman (1998) as described earlier. For each experiment, a set of simulation runs (i.e., simulated "subjects") equal to the number of the human subjects involved were used. The results and the statistical analysis of the results were as follows.

As described before, in experiment 1, subjects were to pick the stronger of the two arguments from each pair. The simulation of experiment 1 showed, the same as the human data, that the more similar argument from each

[3] For the associative memory network, the number of input units was 1800 (for representing all chunks specifiable with 60 dimensions of 30 possible values each), the number of hidden units was 500, and the number of output units was 1800. The learning rate was 0.2 and the momentum was 0.1.

pair of arguments was chosen more often: 82% of times (for inclusion similarity) and 83% of times (for premise specificity). t tests showed that these percentages were significantly above chance, either by subjects ($p < 0.001$) or by argument pairs ($p < 0.001$), the same as in the human data. In our simulation setup, there was a significant involvement of SBR (with $c_{14} = 0.5, c_{15} = 1.0$). If only RBR had been used, then similarity could not have made a difference, and thus both arguments in a pair should have been equally strong. This simulation demonstrated that the conjecture of the involvement of SBR in producing the human data in this experiment was a reasonable interpretation (see the earlier exposition of the human experiments), given the close match with the human data.

In experiment 2, subjects were instead asked to rate the likelihood of each argument. In this simulation, the mean rating was 0.86 for inclusion similarity and 0.87 for premise specificity. Both were significantly below 1, different from what would have been predicted if only RBR had been used, both by subjects ($p < 0.001$) and by arguments ($p < 0.001$), the same as in the human data. ANOVA also showed that across subjects and across argument pairs, there was a significant main effect of similarity (low vs. high; $p < 0.001$). With the same setup as the previous simulation, this simulation again demonstrated the same pattern of significant involvement of SBR in the human performance.

In experiment 4, subjects were asked to rate the likelihood of each argument, right after being presented relevant category inclusion relations. The simulation produced the mean judgment 0.99, exactly the same as the human data. Compared with experiment 2, explicit RBR based on category inclusion was much more prominent in this case, as specified in our simulation setup ($c_{14} = 1.0, c_{15} = 1.0$), which captured the human data accurately.

In experiment 5, ratings were obtained after subjects were asked to make category inclusion decisions. In this case, subjects were reminded of RBR involving category inclusion relations and therefore they were more likely to use RBR (compared with experiment 2), although not exclusively (unlike experiment 4). In the simulation, the mean judgment for experiment 5 was 0.91 for both inclusion similarity and premise specificity, as opposed to 0.86 and 0.87 for the two cases in experiment 2. ANOVA also showed a significant main effect of similarity (low vs. high), across subjects ($p < 0.001$), and across argument pairs ($p < 0.001$). This simulation replicated the human data well, which showed that our interpretation as embodied in the simulation setup ($c_{14} = 0.88, c_{15} = 1.0$), that is, less involvement of RBR compared with experiment 4 but more compared with experiment 2, was a reasonable one.

In all, simulation of this task successfully validated the interpretation and the analysis of human performance in this task and, to some extent, our framework in general.

Concluding Remarks

Overall, the simulation accurately captured the human reasoning data from Sloman (1998). The simulation was conducted based on our framework of mixed rule-based reasoning and similarity-based reasoning, which, along with other simulations published elsewhere (e.g., Sun 1995, 2002, Sun et al 2001, Sun and Zhang 2004), showed the cognitive plausibility of the CLARION architecture to some extent.

This simulation demonstrates the importance of similarity-based reasoning in human everyday reasoning. This similarity-based process is quite distinct from probabilistic reasoning as implemented in other existing cognitive architectures, such as ACT-R (see Anderson 1993 or Anderson and Lebiere 1998). Let us compare the two different approaches. ACT-R as described in Anderson and Lebiere (1998) tries to capture all inferences in a probabilistic framework. In so doing, it lumps together all forms of weak inferential connections in a unified way. Although this approach leads to uniformity, it has shortcomings as well. All similarity relations between any pair of any two objects must be explicitly represented with all the associated parameters, which specify probabilistic computation used to capture similarity-based reasoning (along with other inexact inferences). The problem is the complexity of representing all similarity pairings. This complexity is very high in ACT-R but in contrast is avoided in CLARION.

The limitations of probabilistic reasoning (Pearl 1988) in general include its neglect of many heuristics, simplifications, and rules of thumb (Tversky and Kahneman 1983, Sun 1995, Yang and Johnson-Laird 2001) useful in reducing the computational complexity of formal mathematical models. As a result, it suffers from higher computational complexity (Sun 1995).

We should also look into the framework of Collins and Michalski (1989), which apparently incorporated "similarity-based" reasoning through explicitly representing similarity in a complex logical formalism. Similarity was explicitly represented as a logical operator: That is, for almost any pair of any two objects, there would be a logical relation explicitly represented, denoting their similarity. Inferences could be performed on the basis of similarity operators, using a search process. The complexity of this representational framework was extremely high.

In general, logic-based models suffer from a number of well known shortcomings, including their restrictiveness concerning pre-conditions, consistency, and correctness, and their inability in dealing with inexactness (see, e.g., Israel 1987, Sun 1995). Their restrictiveness renders such models costly, difficult to specify, and difficult to use.

In a different vein, psychological work on reasoning is relevant also. Such work mostly centers around either mental logic (Rips 1994, Braine and O'Brien 1998) or mental models (Yang and Johnson-Laird 2001). Neither approach deals with similarity-based reasoning as captured in CLARION. Their focuses are elsewhere.

In sum, this line of work, combining similarity-based reasoning and rule-based reasoning (Sun 1995, Sloman

1998, Hahn and Chater 1998), offers a new approach for capturing some essential patterns of human everyday reasoning (albeit not all patterns of human reasoning). It complements logic-based "commonsense" reasoning models prevalent in AI, which is very much centered on logic and thus limited by logic. This work also points to new avenues of cognitive modeling, beyond the current psychology of reasoning (which largely focuses on various logics and mental models) and beyond existing cognitive architectures (Anderson and Lebiere 1998). In addition, this approach may well be extended to case-based and/or analogical reasoning (e.g., Sun 1995a).

Acknowledgment

This work has been supported in part by Army Research Institute contract DASW01-00-K-0012 to Ron Sun and Bob Mathews.

References

J. R. Anderson, (1993). *Rules of the Mind*. Lawrence Erlbaum Associates. Hillsdale, NJ.

J. Anderson and C. Lebiere, (1998). *The Atomic Components of Thought*, Lawrence Erlbaum Associates, Mahwah, NJ.

M. Braine and D. O'Brien (eds.), (1998). *Mental Logic*. Lawrence Erlbaum Associates, Mahwah, NJ.

A. Clark and A. Karmiloff-Smith, (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language*. 8 (4), 487-519.

A. Collins, (1978). Fragments of a theory of human plausible reasoning. In: D.Waltz (ed.), *Theoretical Issues in Natural Language Processing II*, 194-201. Ablex, Norword, NJ.

A. Collins and R. Michalski, (1989). The logic of plausible reasoning. *Cognitive Science*, 13(1), 1-49.

E. Davis, (1990). *Representations of Commonsense Knowledge*. Morgan Kaufman, San Mateo, CA.

U. Hahn and N. Chater, (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, 65, 197-230.

D. Israel, (1987). What's wrong with non-monotonic logic? In: Ginsberg (ed.), *Readings in Non-monotonic Reasoning*, pp.53-55, Morgan Kaufman, San Mateo, CA.

M. Moscovitch and C. Umilta, (1991). Conscious and unconscious aspects of memory. In: *Perspectives on Cognitive Neuroscience*. Oxford University Press, New York.

J. Pearl, (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, CA.

L. Rips, (1994). *The Psychology of Proof*. MIT Press, Cambridge, MA.

D. Schacter, (1990). Toward a cognitive neuropsychology of awareness: implicit knowledge and anosagnosia. *Journal of Clinical and Experimental Neuropsychology*. 12 (1), 155-178.

S. Sloman, (1993). Feature based induction. *Cognitive Psychology*, 25, 231-280.

S. Sloman, (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35, 1-33

P. Smolensky, (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11 (1), 1-74.

R. Sun, (1991). Connectionist models of rule-based reasoning. *Proceedings of the 13th Cognitive Science Conference*, pp.437-442. Lawrence Erlbaum Associates, Hillsdale, NJ.

R. Sun, (1995). Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence*. 75, 2. 241-296.

R. Sun, (1995a). A microfeature based approach toward metaphor interpretation. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*. Montreal, Canada. pp.424-430, Morgan Kaufmann, San Francisco, CA.

R. Sun, (2002). *Duality of the Mind*. Lawrence Erlbaum Associates, Mahwah, NJ.

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*. Vol.25, No.2, 203-244.

R. Sun and T. Peterson, (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, Vol.9, No.6, pp.1217-1234.

R. Sun and X. Zhang, (2004). Top-down versus bottom-up learning in cognitive skill acquisition. *Cognitive Systems Research*, Vol.5, No.1, pp.63-89.

Y. Yang and P. Johnson-Laird, (2001). Mental models and logical reasoning problems in the GRE. *Journal of Experimental Psychology: Applied*, 7 (4), 308-316.

A. Tversky, (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

A. Tversky and D. Kahneman, (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 439-450.

Temporal Characteristics of Categorical Perception of Emotional Facial Expressions

Atsunobu Suzuki (suzuki@bayes.c.u-tokyo.ac.jp)
Department of Cognitive and Behavioral Science,
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Susumu Shibui (shibui@bayes.c.u-tokyo.ac.jp)
Department of Cognitive and Behavioral Science,
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Kazuo Shigemasu (kshige@bayes.c.u-tokyo.ac.jp)
Department of Cognitive and Behavioral Science,
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Abstract

Perceptual processing of emotional facial expressions occurs very quickly, even without awareness. To determine whether the fast processing of facial expressions is categorical, we studied temporal characteristics of categorical perception (CP) of facial expressions. We investigated the effect of shortening stimulus duration on participant performance with respect to identifying and discriminating morphed facial expressions. The results of two experiments showed that CP was attenuated or even disappeared when facial stimuli were presented for as briefly as 50-75 milliseconds. These findings indicate that CP is irrelevant to the fast perceptual processing of facial expressions.

Facial expressions of emotion are processed very quickly, even without awareness (Morris, Ohman, & Dolan, 1998; Whalen et al., 1998). It has been proposed that this fast emotional processing provides a 'dirty' image of the external world, enabling organisms to detect salient stimuli immediately (Adolphs, 2002; LeDoux, 1996). What is still not understood, however, is how this fast and dirty processing of facial expressions works.

A hallmark of facial expression recognition is its categorical nature, that is, facial expressions are recognized as belonging to discrete categories of emotion, so-called *basic emotions* (e.g., Ekman, 1992; Izard, 1992). Reports on categorical perception (CP) of facial expressions are thought to provide strong evidence that people process facial expressions categorically (Calder et al., 1996; DeGelder, Teunisse, & Benson, 1997; Etcoff & Magee, 1992; Young et al., 1997). Previous studies have confirmed that recognizing facial expressions fulfills the following features of CP: (a) in identifying a stimulus within a continuum extending from one category to another, the rate of categorizing the stimulus changes abruptly at a boundary (category boundary); and (b) in discriminating a pair of stimuli that differ by a constant physical amount, discrimination is superior for pairs straddling the category boundary (between-category pairs), as compared to pairs falling within one category (within-category pairs). It has been argued that the perceptual system transforms the information continuously received from a given facial expression into categorical information corresponding to the most likely emotion (Etcoff & Magee, 1992).

In previous research, however, facial stimuli were presented for a relatively long period ($\geqq$750 ms), so it remains unclear whether the fast perceptual processing of facial expressions is categorical in nature. Our goal was to investigate the effects of shortening stimulus duration on CP of facial expressions and to examine whether the fast perceptual processing is categorical.

Experiment 1

In Experiment 1, participants identified facial expressions from three continua extending from anger to happiness, from happiness to fear, and from fear to anger. We investigated how shortening stimulus exposure duration (750 ms, 150 ms, 50 ms) affected the abrupt change in the rate of identification at the category boundary.

Method

Participants Twelve undergraduates and postgraduates participated (8 men, 4 women; 20-24 years old).

Facial Stimuli Three sets of eight gray-scale images of facial expressions were used. Each set consisted of a continuum extending either from anger to happiness, from happiness to fear, or from fear to anger. The original (endpoint) facial expressions were posed photographs of a Japanese woman, and interpolated (morphed) expressions between the two continuum endpoints were created with software for facial expression processing (Information-technology Promotion Agency, Japan (IPA), 1998).

The morph transformation started with the delineation of two endpoint expressions; landmarks were placed manually on critical positions of each image. There were 759 landmarks in total, 88 points of which were placed manually on corresponding positions of each image: for the head, 4 points; for the outline, 28 points; for the eyes, 5 x 2 points; for the eyebrows, 4 x 2 points; for the nose, 4 points; for the

mouth, 6 points; for the neck, 13 points; and finally for the hairline, 15 points. An intermediate face was then created with linear interpolation between point-to-point pixel intensity values, yielding a weighted blend of both facial configuration and texture of the two endpoint faces. Six intermediate faces were generated for each continuum, so that eight images from one endpoint to the other were spaced with a 14.3% gap.

Procedure and Design Participants performed the identification task (three sessions) by viewing stimuli presented on a 17-inch CRT monitor. Each trial began with an 800-ms presentation of a fixation point, followed by a blank interval of 300 ms, and then a facial stimulus from one continuum for 750 ms, 150 ms, or 50 ms. After a 300-ms blank interval, participants were asked to decide which endpoint category the face expressed. The continuum was manipulated across sessions: from anger to happiness (AH session), from happiness to fear (HF session), and from fear to anger (FA session). The order of the three sessions was counterbalanced across participants. Each session began with 24 training trials, followed by five blocks of 48 experimental trials, corresponding to two repetitions of the 24 combinations of faces (8) and stimulus durations (3). The order of the face-duration sets was fully randomized in each repetition.

Results & Discussion

Figure 1 displays the overall percentage with which a given endpoint expression was identified at each stimulus duration, for each continuum. As an estimate of the category boundary, the point at which two endpoints were identified with equal probability was computed by applying a logit model to each identification rate curve (Table 1). In general, the category boundary lay near the 50% morph. More importantly, Table 1 also indicates the slope of the logistic curve at the category boundary as a measure of the identification rate change at the boundary. Multiple comparisons (z tests) with Bonferroni's method (α=0.017) revealed that the slope at 50 ms was significantly, or marginally significantly, smaller than at 150 ms and at 750 ms for the anger-happiness and happiness-fear continua (AH session, 50 ms vs. 150 ms, z=2.37, p=0.018, 50 ms vs. 750 ms, z=3.67, p<0.001, 150 ms vs. 750 ms, z=1.48, p=0.139; HF session, 50 ms vs. 150 ms, z=2.69, p=0.007, 50 ms vs. 750 ms, z=4.07, p<0.001, 150 ms vs. 750 ms, z=1.74, p=0.082). No slope change was observed for the fear-anger continuum (all ps>0.35).

The decreased slopes of the identification rate curves for the anger-happiness and happiness-fear continua at 50 ms indicate that the CP of facial expressions is attenuated with the briefest stimulus exposure duration. The lack of slope change for the fear-anger continuum may be due to the originally weak categorical perception between fear and anger, which reflects their remarkable similarity with respect to the dimensional information of pleasure and arousal (Russell, 1997). Indeed, the slope for the fear-anger continuum was the smallest of the three continua at 750 ms. Likewise, there are previous reports that have failed to detect clear CP for the fear-anger continuum (Calder et al., 1996; DeGelder et al., 1997).

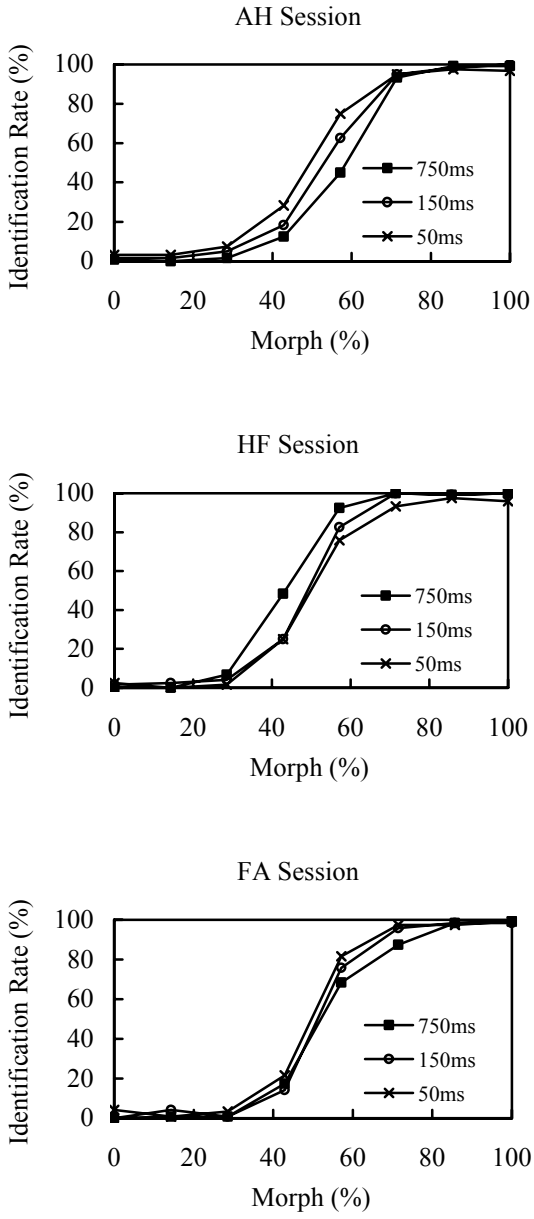


Figure 1: Overall percentage with which a given endpoint expression (top: happiness, middle: fear, bottom: anger) was identified at each stimulus duration, for each session. *Morph* represents a percentage of happiness (top), fear (middle), or anger (bottom) for a given face.

Table 1: Estimates of the position of the category boundary and the slope at the boundary.

Session	Duration	Boundary (%)	Slope
AH	750 ms	56.9	14.5
	150 ms	52.5	12.5
	50 ms	49.0	10.0
HF	750 ms	43.3	18.1
	150 ms	47.9	14.9
	50 ms	51.2	11.3
FA	750 ms	53.9	13.3
	150 ms	51.8	13.8
	50 ms	49.2	12.6

Note *Boundary* represents a percentage of happiness (AH Session), fear (HF Session), or anger (FA Session) for a morphed face.

Experiment 2

Experiment 1 revealed that the abrupt change in identification rate at the category boundary was attenuated at the briefest stimulus exposure duration, indicating that CP of facial expressions did not reflect fast perceptual processing. To further examine temporal characteristics of CP, Experiment 2 investigated whether superior discrimination for between-category pairs could be observed with the brief stimulus exposure duration when compared to discrimination for within-category pairs. Because a more accurate discrimination performance for between-category pairs is regarded as the key indicator of CP (Harnad, 1987), its disappearance at the brief stimulus duration provides crucial evidence that categorical perception does not occur rapidly. In Experiment 2, participants engaged in both discrimination and identification tasks of facial expressions from the happiness-fear continuum. We selected the happiness-fear continuum from the three continua used in Experiment 1 because it did not incorporate any gross changes in physical features (e.g., from open to closed mouth), which might obscure CP (Calder et al., 1996).

Method

Participants Twenty-three undergraduates participated (12 men, 11 women; 18-24 years old).

Facial Stimuli One set of 11 gray-scale images of facial expressions was used. The set consisted of a continuum extending from happiness to fear. The endpoint expressions and the morphing procedure to create interpolated faces were the same as in Experiment 1. Nine intermediate faces were generated so that the 11 images from happiness to fear were spaced with a 10% gap.

Procedure and Design Participants performed two successive tasks (two sessions each) by viewing stimuli presented on a 17-inch CRT monitor.

XAB discrimination task The sequence of the stimuli presented in each trial was as follows: (1) a fixation point for 450 ms; (2) a blank interval for 300 ms; (3) a facial stimulus 'X' for 150 ms or 75 ms (target); (4) a black-and-white checker-pattern for 150 ms (backward mask); (5) a blank interval for 750 ms; (6) and finally, two facial stimuli 'A' and 'B', positioned horizontally to the right and left of center, displayed until the participants responded (reference). Participants were asked to decide which reference was identical to the target.

The backward mask was used to restrict visual access to the target over the controlled stimulus exposure duration (Enns & DiLollo, 2000). Because the backward mask might degrade target perception, we used 75 ms as the brief exposure duration, which is somewhat longer than the 50 ms tested in Experiment 1. We also used 150 ms as the sole longer exposure duration interval, because we speculated that task difficulty might differ too much between 75 ms and 750 ms. Reference stimuli were spaced with a 20% gap, resulting in nine possible pairs. For each pair, there were four presentation orders; (X,A,B) = (A,A,B), (A,B,A), (B,A,B), (B,B,A). Target duration was manipulated across sessions, and the order of the two durations was counterbalanced across participants. Each session contained 10 training trials and two blocks of 36 experimental trials representing all combinations of pairs (9) and presentations (4). The order of the pair-presentation sets was fully randomized in each block.

Identification task Stimulus sequence in each trial was the same as in the XAB discrimination task, with the exception that reference stimuli were removed. Participants were asked to decide whether the target stimulus expressed fear or happiness. The order of the two sessions across which the target duration was manipulated was the same as in the XAB discrimination task. Each session contained 10 training trials and two blocks of 44 experimental trials corresponding to four repetitions of the 11 faces. The order of the faces was fully randomized in each repetition.

Results & Discussion

Identification Task Figure 2 presents the overall percentage of trials in which "fear" was identified at each stimulus exposure duration in the identification task. Application of a logit model to each identification rate curve revealed that the category boundary lay at 52.4% (150 ms) and 55.2% (75 ms). Contrary to the results of Experiment 1, logit models also showed that the slope at 150 ms (17.9) was somewhat smaller than at 75 ms (19.4), though this difference did not reach significance (z=1.03, p=0.303). As the slope was attenuated in proportion to the decreased duration in Experiment 1, these data suggest that the duration difference between 150 ms and 75 ms was insufficient to cause a significant slope change.

XAB Discrimination Task Figure 3 presents the overall correct response rate for each stimulus exposure duration in

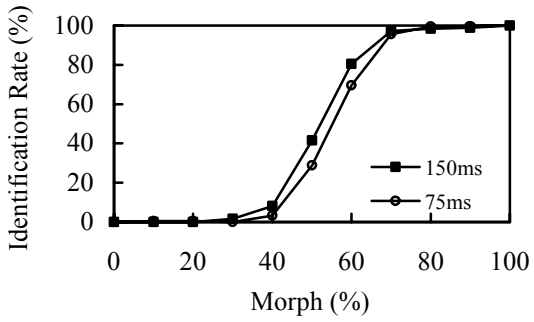


Figure 2: Overall percentage with which "fear" was identified at each stimulus duration. *Morph* represents a percentage of "fear" for a given face.

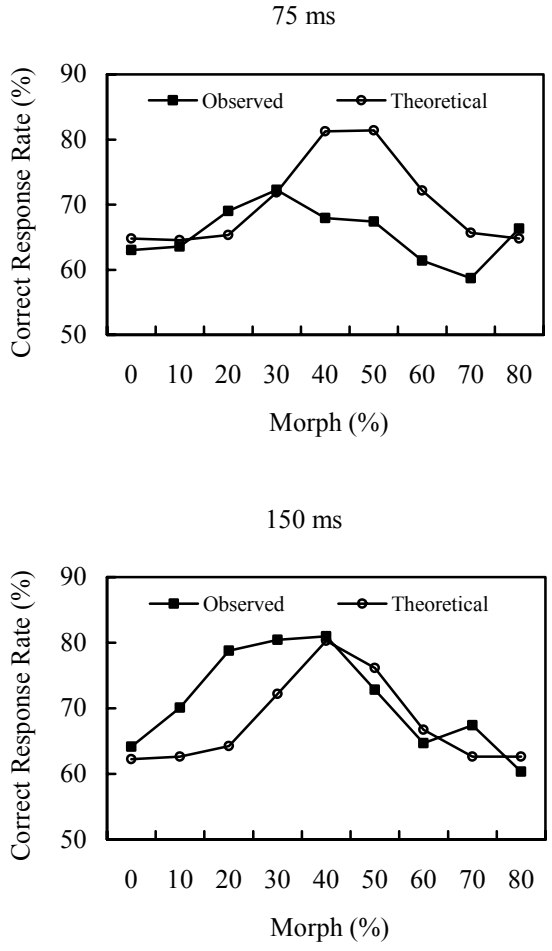


Figure 3: Overall correct response rate in the XAB discrimination task for each duration. *Observed*=observed rate. *Theoretical*=theoretical rate predicted from CP. *Morph* represents a percentage of "fear" for the less fearful face of a given pair.

the XAB discrimination task. The data from the identification task revealed that identifying between-category pairs was 40%-60% and 50%-70% at both durations. The peak in discrimination performance at the category boundary was found only at the 150-ms stimulus exposure duration.

We calculated the mean correct rate for between-category pairs (40%-60%, 50%-70%) and within-category pairs (the remaining seven pairs) for each duration, and conducted a 2 x 2 ANOVA of the mean correct rate with the two factors of pair (between-category or within-category) and stimulus duration (150 ms or 75 ms). A significant main effect of pair was found ($F(1,22)=7.79$, $p=0.011$, $MSE=0.008$), indicating better discrimination for between-category pairs (72%) than for within-category pairs (67%). The main effect of duration was also significant ($F(1,22)=9.74$, $p=0.005$, $MSE=0.011$), indicating worse performance at 75 ms (66%) than at 150 ms (73%). There was no significant interaction between the two factors ($p>0.25$), suggesting that discrimination of between-category pairs was more accurate than that for within-category pairs, at both durations. However, post-hoc paired *t* tests revealed that superior discrimination for between-category pairs was significant only at the 150-ms exposure duration (150 ms, $M=8\%$, $t(22)=2.42$, $p=0.024$; 75 ms, $M=3\%$, $t(22)=1.18$, $p=0.252$). Post-hoc comparisons also revealed that poorer discrimination at 75 ms was significant for both between-category and within-category pairs (between-category, $M=9\%$, $t(22)=2.36$, $p=0.027$; within-category, $M=4\%$, $t(22)=2.66$, $p=0.014$).

To indicate CP in the discrimination task, previous studies (Calder et al., 1996; Young et al., 1997) have reported correlations between observed and theoretical performances, predicted from CP. Calder and colleagues (Calder et al., 1996) assumed that discrimination between two facial expression stimuli depends on two cues: first, the physical difference between the stimuli, which is constant for any pair, regardless of their expressions; and second, the expression categories of the stimuli. To estimate the contribution of the first non-categorical cue, they computed the mean observed discrimination for the two pairs, placed at both ends of the continuum. To estimate the contribution of the second categorical cue, they calculated the difference in identification rates for the two relevant stimuli observed in the identification task, and multiplied the obtained difference by 0.25 (a constant). Calder et al. (1996) argued that summing the two estimates measures theoretical performance in the discrimination task.

Figure 3 also presents theoretical performance, based on the same formula as Calder et al. (1996), along with observed performance. The fit between observed and theoretical performance looks better at 150 ms than at 75 ms; indeed, their correlation was significant only at the 150-ms exposure duration (150 ms, $r=0.667$, $t(7)=2.37$, $p=0.050$; 75 ms, $r=0.362$, $t(7)=1.03$, $p=0.338$).

Analyses revealed that the superior discrimination performance for between-category pairs disappeared at 75 ms but was present at 150 ms, providing decisive evidence of null CP at the brief stimulus duration. In response to objections claiming that observing only one continuum is insufficient, studies using neuroimaging (Morris et al.,

1996; Morris, Friston, et al., 1998; Whalen et al., 1998; Wright et al., 2001) and developmental data (Kotsoni, DeHaan, & Johnson, 2001) have indicated that happiness and fear are the most distinct categories, and that the lack of CP between the two expressions is important. Critics might also insist that the disappearance of CP at 75 ms is a floor effect; however, discrimination for any stimulus pair at 75 ms was above chance (all $ps<0.01$), indicating that residual perception of differences in some visual properties was present. Moreover, the 75-ms duration was sufficient for accurate identification of near-endpoint facial expressions (see Figure 2). Thus, categorical perception did not occur, even with such well-preserved visual processing.

General Discussion

The present experiments revealed that categorical perception of facial expressions was attenuated and even disappeared with brief stimulus exposure durations. The abrupt change in identifying facial expressions at the category boundary was weakened at the 50-ms stimulus duration (Experiment 1), and the more accurate discrimination observed for between-category pairs than for within-category pairs was eliminated at the 75-ms stimulus duration (Experiment 2). Our findings indicate that the fast perceptual processing of emotional facial expressions is not categorical.

It has been postulated that the fast processing of facial expressions involves subcortical structures, specialized in the detection of *salient* stimuli (Adolphs, 2002). However, categorical perception refers to the processing of *ambiguous* stimuli or the transformation of an indistinctive facial configuration into a distinctive emotion category. Such fine analysis of emotional information may involve cortical structures, implying slower processing (LeDoux, 1996).

Etcoff and Magee (1992) have claimed that their observation of CP in facial expression recognition rejects the possibility that categorical processing of facial expressions is performed by higher conceptual and linguistic systems. There are data, however, demonstrating that verbal interference eliminated CP of colors and facial expressions, suggesting the essential role of verbal coding over visual coding (Roberson & Davidoff, 2000). Medin and Barsalou (1987) have argued that it is important to distinguish between *sensory perception categories* (SP categories), categories related to low-level perceptual experiences, and *generic knowledge categories* (GK categories), categories related to high-level knowledge representation. Shibui and colleagues (Shibui, Yamada, Sato, & Shigemasu, 2001) found that discrimination accuracy for within-category pairs of facial expressions was proportional to their semantic distance, and they inferred that facial expressions belong to cognitive GK categories. Our findings are consistent with the view that categorical processing of facial expressions is performed by higher cognitive systems.

Some researchers have postulated that CP of facial expressions supports the notion of basic emotions (DeGelder et al., 1996; Ekman, 1994), such that each emotion possesses an innate and specialized neuro-cognitive system. However, categorical perception is also known to occur for non-emotional visual categories such as identity (Beale & Keil, 1995; Levin & Beale, 2000), race (Levin & Angelone, 2002), and even familiar objects (Newell & Bulthoff, 2002). Categorical perception of facial expressions may thus reflect visual information processing that is common to general object recognition rather than specific to emotional content.

References

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology, 12,* 169-177.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition, 57,* 217-239.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3,* 81-117.

DeGelder, B., Teunisse, J. P., & Benson, P. J. (1997). Categorical perception of facial expressions: categories and their internal structure. *Cognition and Emotion, 11,* 1-23.

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99,* 550-553.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin, 115,* 268-287.

Enns, J. T., & DiLollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4,* 345-352.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44,* 227-240.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: a critical over view. In S. Harnad (Ed.), *Categorical perception: the groundwork of cognition* (pp. 1-25). Cambridge: Cambridge University Press.

Information-technology Promotion Agency, Japan. (1998). Software for Facial Image Processing System for Human-like 'Kansei' Agent [Computer Software]. Retrieved from http://www.tokyo.image-lab.or.jp/aa/ipa/

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review, 99,* 561-565.

Kotsoni, E., DeHaan, M., & Johnson, M. H. (2001). Categorical perception of facial expressions by 7-month-old infants. *Perception, 30,* 1115-1125.

LeDoux, J. (1996). *The emotional brain: the mysterious underpinnings of emotional life.* New York: Simon & Schuster Inc.

Levin, D. T., & Angelone, B. L. (2002). Categorical perception of race. *Perception, 31,* 567-578.

Levin, D. T., & Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception and Psychophysics, 62,* 386-401.

Medin, D. L., & Barsalou, L. W. (1987). Categorization processes and categorical perception. In S. Harnad (Ed.), *Categorical perception: the groundwork of cognition* (pp. 1-25). Cambridge: Cambridge University Press.

Morris, J. S., Friston, K. J., Buchel, C., Frith, C. D., Young, A. W., Calder, A. J., & Dolan, R. J. (1998). A neuromodulatory role for the human amygdala in

processing emotional facial expressions. *Brain, 121,* 47-57.

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A. W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383,* 812-815.

Morris, J. S., Ohman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature, 393,* 467-470.

Newell, F. N., & Bulthoff, H. H. (2002). Categorical perception of familiar objects. *Cognition, 85,* 113-143.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: the effect of verbal interference. *Memory and Cognition, 28,* 977-986.

Russell, J. A. (1997). Reading emotions from and into faces: resurrecting a dimensional-contextual perspective. In J. A. Russell, & J. M. Fernandez-Dols (Ed.), *The psychology of facial expression* (pp. 295-320). Cambridge: Cambridge University Press.

Shibui, S., Yamada, H., Sato, T., & Shigemasu, K. (2001). The relationship between the categorical perception of facial expressions and semantic distances. *The Japanese Journal of Psychology, 72,* 219-226.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience, 18,* 411-418.

Wright, C. I., Fischer, H., Whalen, P. J., McInerney, S., Shin, L. M., & Rauch, S. L. (2001). Differential prefrontal cortex and amygdala habituation to repeatedly presented emotional stimuli. *NeuroReport, 12,* 379-383.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition, 63,* 271-313.

Making Sense of Embodiment: Simulation Theories and the Sharing of Neural Circuitry Between Sensorimotor and Cognitive Processes

Henrik Svensson (svensson@ida.his.se)
University of Skövde, School of Humanities and Informatics
Box 408, 54128 Skövde, Sweden

Tom Ziemke (tom@ida.his.se)
University of Skövde, School of Humanities and Informatics
Box 408, 54128 Skövde, Sweden

Abstract

Although an increasing number of cognitive scientists are convinced that cognition is *embodied*, there still is relatively little agreement on what exactly that means. Notions of what it actually means for a cognizer to be embodied range from simplistic ones such as 'being physical' or 'interacting with an environment' to more demanding ones that consider a particular morphology or a living body prerequisites for embodied cognition. Based on experimental evidence from a range of disciplines, we argue that one of the keys to understanding the embodiment of cognition is the sharing of neural mechanisms between sensorimotor processes and higher-level cognitive processes. The latter are argued to be embodied in the sense that they make use of (partial) simulations or emulations of sensorimotor processes through the re-activation of neural circuitry also active in bodily perception and action.

Introduction

Although an increasing number of cognitive scientists are convinced that cognition is *embodied* (e.g. Varela et al., 1991; Clancey, 1997; Clark, 1997; Lakoff & Johnson, 1999; Ziemke, 2002), there still is relatively little agreement on what exactly that means. Notions of what it actually means for a cognizer to be embodied range from simplistic ones such as 'being physical' or 'interacting with an environment' to more demanding ones that consider a particular morphology or a living body prerequisites for embodied cognition (cf., e.g., Chrisley & Ziemke, 2002; Wilson, 2002; Anderson, 2003; Ziemke, 2003).

This lack of agreement or coherence, after more than a decade of research on embodied cognition, has unfortunate consequences. Firstly, critics commonly argue that the only thing that embodied cognitive theories have in common is in fact the rejection of traditional, computationalist and supposedly disembodied cognitive science. Secondly, there is a certain trivialization of embodiment, not least among many AI researchers who consider as embodied any physical system, or in fact any agent that interacts with some environment, such that the distinction between computationalist and embodied cognitive theories disappears since, in some sense, all systems are embodied, and thus cognitive science has always been about embodied cognition (Chrisley & Ziemke, 2002). Thirdly, there is the 'misunderstanding' that perhaps embodiment is only relevant to sensorimotor processes directly involving the body in perception and action, while higher-level cognition might very well be computational in the traditional sense and only dependent on the body in the sense that mental representations ultimately need to be grounded in sensorimotor interaction with the physical environment.

This paper, on the other hand, argues that one of the keys to understanding the embodiment of cognition, in an important, non-trivial sense, is to understand the sharing of neural mechanisms between sensorimotor processes and higher-level cognitive processes. Based on experimental evidence from a range of disciplines, we argue that many, if not all, higher-level cognitive processes are body-based in the sense that they make use of (partial) *simulations* or *emulations* of sensorimotor processes through the re-activation of neural circuitry that is also active in bodily perception and action (cf. Clark & Grush, 1999; Grush, in press; Hesslow, 2002). As Barsalou et al. (2003) put it, the main point is that "simulations of bodily states in modality specific brain areas may often be the extent to which embodiment is realized".

The next section elaborates the key idea of this paper, i.e., cognition as body- and simulation-based in the above sense, in more detail. In the following sections then supporting empirical evidence from a range of disciplines is presented. The final section then presents a brief summary as well as some open questions and directions for future work.

Cognition as body-based simulation

The idea that even higher-level cognitive processes are in a strong sense grounded in bodily activity and experience is, of course, hardly new, but was developed already in the 1980s, most influentially by Maturana and Varela (1980, 1987) from a neurobiological perspective, and by Lakoff and Johnson (1980, 1999) from a linguistic perspective. Lakoff (1988) summarized the basic idea as follows:

> Meaningful conceptual structures arise from two sources: (1) from the structured nature of bodily and social experience and (2) from our innate capacity to imaginatively project from certain well-structured aspects of bodily and interactional experience to abstract conceptual structures.

Back in the 1980s, however, relatively little was known about exactly how such an *imaginative projection* from bodily experience to abstract concepts might work. In recent years more detailed accounts of how the sensorimotor structures of the brain are involved in cognition have been developed in several disciplines, often taking into account data from neurophysiological and neuroimaging studies. These accounts show that the traditional strong division between perception and action, as well as between sensorimotor and cognitive processes, needs to be revised.

A particular kind of "embodiment" theory that has emerged in different contexts are so-called *emulation* or *simulation theories*[1] (e.g., Barsalou et al., 2003: Decety, 1996; Frith & Dolan, 1996; Grush, in press; Hesslow, 2002; Jeannerod, 1994, 2001). The basic idea is that neural structures that are responsible for action and/or perception are also used in the performance of various cognitive tasks. As Hesslow (2002) pointed out, this idea is not entirely new; e.g. Alexander Bain suggested in 1896 that thinking is basically a covert form of behavior that does not activate the body and thus remains invisible to external observers. Today simulation theories, based partly on data from neuroscience, can further clarify the possible role of simulation in cognition, thus explaining in a more concrete way than before the embodiment of cognition.

One of the more comprehensive descriptions of the idea has been presented by Grush (in press; see also Clark & Grush, 1999). Based on the control theoretic concept of forward models (emulators), previously used to account for motor control (e.g., Wolpert & Kawato, 1998), Grush developed an emulation theory for several types of cognitive processes, including perception, imagery, reasoning and language. In a nutshell, he argued that emulation circuits are able to calculate a forward mapping from control signals to the (anticipated) consequences of executing the control command. For example, in goal-directed hand movements the brain has to plan parts of the movement before it starts. To achieve a smooth and accurate movement proprioceptive/kinesthetic (and sometimes visual) feedback is necessary, but sensory feedback per se is too slow to affect control appropriately (Desmurget & Grafton, 2000). The 'solution' is an emulator/forward model that can predict the sensory feedback resulting from executing a particular motor command.[2]

The following section summarizes a number of the many empirical studies that support the idea that cognition is body-based, especially as predicted by simulation theories.

Empirical Evidence

Several sources of evidence support the basic tenet of the simulation account, viz., that perceptual and motor areas of the brain can be covertly activated either separately or in sequence for use in cognitive processes. In particular, several studies have indicated that there are extensive similarities between the neural structures activated during preparation (and execution) of an action and mentally simulating an action (i.e., motor imagery), as well as between visual perception and visual imagery. The similarities are so striking that some have argued that internally activated actions and perceptions are the same as overt ones, except that the overt execution or sensory input is missing (e.g., Hesslow, 2002; Jeannerod, 2001).

The following subsections review some of the empirical evidence that suggest that sensorimotor structures of the brain are deeply involved in the generation of cognitive phenomena, such as imagery and problem solving.

Motor imagery

There is an extensive literature on the neural and behavioral similarities between actions and motor imagery (e.g., Decety, Jeannerod, & Prablanc, 1989; Jeannerod & Decety, 1995; Jeannerod & Frak, 1999; for reviews see Decety, 1996, 2002; Jeannerod, 1994, 2001).[3]

Motor imagery is the recreation of an experience of actually performing an action, e.g., the person should feel as if he/she was actually walking (Decety, 1996; Jeannerod, 1994). The evidence cited in support for the equivalence of performing an action and simulating an action comes mainly from three different sources: *mental chronometry*, *autonomic responses*, and *measurements of brain activity*.[4]

In *mental chronometry* experiments, it has been found that the time to mentally execute actions closely corresponds to the time it takes to actually perform them (Jeannerod & Frak, 1999). For example, Decety and Jeannerod (1996) found that Fitt's law (i.e., the finding that execution times increase with task difficulty) also holds for motor imagery. Decety et al. (1989) compared the durations of walking towards targets (with blindfolds) placed at different distances and mental simulation of walking to the same targets. In both conditions times were found to increase with the distance covered.

Autonomic responses, such as the adaptation of heart and respiratory rates, which are beyond voluntary control have been shown to be activated by motor imagery to an extent proportional to that of actually performing the action, and as a function of mental and actual effort (Decety, 1996; Jeannerod, 1994; Jeannerod & Decety, 1995).

[1] The terms simulation and emulation are used somewhat interchangeably in this paper, as in much of the literature, but it should be noted that they are sometimes used differently (e.g., Grush, in press).

[2] According to Blakemore, Frith and Wolpert (1999), this is also why it is not so easy to tickle oneself: the forward model produces predicted sensory feedback that 'prepares' the agent.

[3] There are also similarities between actions and other cognitive tasks, such as observing an action and prospective action judgments, which differ from motor imagery in that they do not produce a conscious motor image of performing an action and the brain activation is not as similar to overt actions as in motor imagery (cf., Jeannerod, 2001).

[4] Further evidence comes from effects on action performance after mental training using motor imagery and also from lesion studies.

Since the first study that investigated motor imagery using regional cerebral blood flow (rCBF) to indicate active brain areas (Ingvar & Philipson, 1977) there have been many neuroimaging experiments that confirm the first study's indication that similar brain areas are activated during action and motor imagery. The general conclusion is that there is a functional equivalence between performing an action and mentally simulating it (Decety, 2002; Jeannerod, 2001).

Together, these three types of evidence point to an explanation of motor imagery that implicates sensorimotor structures.[5] Although the focus above has been on actions, it should not be forgotten that the motor system also integrates sensory information when planning and executing an action (e.g., Grush, in press; cf. Desmurget & Grafton, 2000; Jeannerod, 1997). Thus, simulating an action might also involve an emulator mechanism that predicts the sensory feedback that would have resulted from the executed action (Decety, 2002; Grush, in press; Jeannerod, 1997, 2001).

Mental visual imagery

The discussion of motor imagery can be extended to the visual modality in that the same type of studies report that perceptual structures can and are internally reactivated when, e.g., visually recreating a previous perception. Many studies in cognitive psychology have found similarities between visual perception and visual imagery (Farah, 1988; Finke, 1989). For example, Shepard and Metzler (1971) performed a number of mental chronometry type experiments, where they compared the concrete manipulation of physical objects and corresponding manipulations performed mentally. In their experiments, the time between the two conditions was closely correlated, which suggests that mental imagery uses the same mechanisms as the visual system. Although alternative explanations are difficult to rule out, neuropsychological and neuroimaging studies offer more conclusive evidence of the involvement of sensorimotor areas of the brain in mental visual imagery (Farah, 1988; Hesslow, 2002).

Note that these findings do not necessarily influence debates on representational format, since both perception and imagery may be said to use the same format (Block, 1983). On the other hand, explaining cognition as reactivation of sensorimotor structures does not (at least in some cases) rely on the computer metaphor of symbol manipulation, and thus may offer a novel view that does not see the vehicle and the content of representations as separate entities but as constitutive of each other (Gallese, 2003; cf. Dreyfus, 2002; Thomas, 1999). Gallese argued that canonical neurons in the monkey brain illustrate how the interaction between an agent and its environment provides an example of such representations.

Canonical neurons

In the macaque monkey, neurons located in the rostral part of the inferior premotor cortex (area F5) of the monkey brain discharge during goal directed movements, such as grasping, holding, or tearing. However, they do not respond to similar movements, but only actions that have the same "meaning" (di Pellegrino et al. 1992; Rizollatti et al., 1996; Rizzolatti et al., 2002), which is why they are often interpreted as internal representations of actions, rather than motor or movement commands (Jeannerod, 1994; Rizollatti et al. 1996; Rizzolatti et al., 2002). Gallese (2003) emphasized seeing them as coding not physical parameters of movement, but a relationship between agent and object.

Some of the neurons in area F5, so-called *canonical neurons*, also have sensory properties and discharge both during the action they code and when an object that *affords* that action in the Gibsonian sense is perceived. Canonical neurons have a strict congruence between the type of grasping action and the size or shape of the object they respond to (Gallese, 2003). This implies that they implement affordances, e.g. code things that are graspable-in-a-certain-way, specifying not only perceptual and action aspects but a particular relationship between agent and environment (cf. Gallese, 2003, cf. also Dreyfus, 2002).

Problem solving using covert perception and action

Instead of adverting to symbol manipulation, a flexible inner world, in which an agent might try out possible action sequences, can be explained by internal activation of perceptions and actions (Clark & Grush, 1999, Grush, in press; Hesslow, 2002). That means, an agent can sustain such an inner world by letting an internally activated action elicit through an anticipatory mechanism internally generated perceptions that would be the likely result of executing that action in the particular external situation. As discussed above, the internal activation of sensorimotor structures is well supported, but the neural underpinnings of the anticipation mechanism are still an open issue. Some suggest the involvement of the cerebellum (e.g., Hesslow, 2002).

Some support for the involvement of this type of simulation of behavioral chains comes from the problem solving and planning involved in the Tower of London (ToL) problem (Shallice, 1982). Dagher et al. (1999) found that planning and problem solving activated higher motor areas (premotor cortex, prefrontal cortex) and the basal ganglia, and that they seemed to interact with visual and posterior parietal areas (cf. Schall et al., 2003). This gives some support to the idea that the subjects solved the problem by simulating the action of moving one ball to another location through the use of reactivated perceptions and actions (Hesslow, 2002).

[5] There are a number of unanswered questions concerning motor imagery worth mentioning here, which, however, do not affect the paper's main point concerning the embodiment of cognition. To what degree do actions and mental simulations of actions engage executive motor structures (such as, the primary motor cortex) (cf., e.g., Decety, 2002; Jeannerod & Frak, 1999), and how is the overt movement hindered (Jeannerod, 2001; Hesslow, 2002)? Although there may not be a complete overlap between the neural structures involved in real and mentally simulated action, the evidence suggests that they are not different in nature, but only in degree.

Social cognition and language

The use of internally reactivated sensorimotor structures has also been suggested to play a crucial role in social cognition, especially emotive states (cf. Barsalou et al., 2003; Nielsen, 2002). The evidence that is reported in this subsection emphasizes that perceptual and motor processes are not different in nature at the neural and behavioral level, but seem to be intimately linked in social cognition possibly through simulation mechanisms.

Embodiment effects Barsalou et al. (2003) argued that there are at least four types of well known phenomena in social cognition which can be explained by simulation of bodily states. The types of effects they mention are: a) perceived social stimuli produce bodily states, b) perceiving bodily states produce bodily mimicry, c) bodily states can produce and affect emotion states, and d) compatibility between bodily states and emotional states increases performance (see also Nielsen, 2002). These effects seem to arise automatically without any conscious mediating knowledge structures (Barsalou et al., 2003; Nielsen, 2002).

Firstly, social stimuli do not only produce cognitive responses, but at the same time automatically and unconsciously cause bodily responses (e.g., movement patterns, facial expressions). For example, subjects induced with an elderly stereotype (e.g., by watching elderly people) perform in a manner more similar to the elderly stereotype than control subjects (Barsalou et al., 2003).

Secondly, bodily responses sometimes are the same or similar to the eliciting social stimuli as in the many cases of facial mimicry. For example, watching somebody yawning often causes oneself to yawn too (cf. Barsalou et al., 2003).

Thirdly, bodily states can directly induce effects on affective states (Barsalou et al., 2003). One such effect is the facial feedback hypothesis, which states that a person's own facial expressions can (either directly from the brain areas responsible for the facial expression or through processing of proprioceptive feedback) produce or modify his/her experience of the emotional state (Nielsen, 2002).

Finally, when there is a mismatch between bodily state and affective state cognitive performance is degraded. Barsalou et al. (2003) argued that this is perhaps the most important effect because it indicates that embodied states are directly involved in higher cognition. That is, the affective states are thought to involve sensorimotor states and when there is an incompatible bodily state co-present it produces a competing sensorimotor state, which reduces performance. For example, Wells and Petty (1980) showed that head movements were faster when compatible with the message (e.g., nodding vertically to an agreeable message) than when incompatible (cf. Barsalou et al., 2003).

The findings discussed in this section suggest that bodily states are involved in social cognition and that they might constitute the very foundations of the particular social cognitive phenomena in question. An example of how perception, action, and social cognition come together at the level of single neurons is so-called *mirror neurons* in macaque monkeys (Decety & Sommerville, 2003).

Observation execution matching system Besides canonical neurons, area F5 of the monkey brain contains so called *mirror neurons*, which are neurons with sensory properties that become activated both when performing a specific action and when observing the same goal-directed hand (and mouth) movements of an experimenter (di Pellegrino et al., 1992; Rizzolatti et al., 1996). Mirror neurons provide a key example of sensorimotor brain structures also involved in (social) cognitive processes.

Although different hypotheses exist, many of the theories of the function of mirror neurons emphasize their role in social cognition (e.g., Decety & Chaminade, 2003; Gallese & Goldman, 1998; Rizzolatti & Arbib, 1998; Rizzolatti et al. 2002). These researchers acknowledge that area F5 and mirror neurons can be interpreted as a kind of observation-execution mechanism or resonance mechanism, which links the observed actions to actual actions of the subject's own behavioral repertoire. That is, it enables the monkey to understand the meaning of the observed action. Thus, mirror neurons can be interpreted as representations of actions, used both for performing and understanding actions (e.g., Rizzolatti et al., 1996; Rizzolatti & Arbib, 1998). [6]

Gallese and Goldman (1998) hypothesized that mirror neurons might be a basic mechanism necessary for "mind-reading", i.e., attributing mental states in others. They further argued that such mechanisms can explain how an agent determines what mental states of another agent have already occurred. When mirror neurons are externally activated by observing a target agent executing an action (allowing the subject to evaluate the meaning of the other's action), the subject knows (visually) that the observed target is currently performing this very action and thereby "tags" the "experienced" action as belonging to the target. However, how the subject can distinguish its own actions from those performed by others is relatively unknown (cf. Blakemore, Wolpert & Frith, 2002).

Language Some researchers have argued that conceptualization and language understanding cannot be achieved through the manipulation of amodal, arbitrary symbols alone, but has to be grounded in bodily interaction with the environment. In particular, Glenberg and Kaschak (2003) have outlined an explanation of language in line with the idea of cognition as body-based simulation as expressed in this paper, suggesting that language is partly achieved through the same neural structures used to plan and guide action.

[6] For practical and ethical reasons it is so far not possible to investigate the existence of mirror neurons (and canonical neurons) at the single neuron level in humans. However, many researchers have presented strong arguments for the existence of a similar system in humans (e.g., Fadiga et al., 1995; Grafton et al., 1996; Rizzolatti, & Arbib, 1998; Rizzolatti et al., 1996).

Under the heading of the *indexical hypothesis* they developed an account of language comprehension partly based on simulation of action. They argued that the meaning of a sentence is achieved by a process that indexes words to perceptual symbols which in turn retrieves the available affordances in the situation and determines their relevance through the particular sentence construction. Thus, the understanding of a sentence is essentially achieved through a simulation of action using the same neural systems active in overt behavior.

An empirical result that supports the close coupling between language and action is the "action-sentence compatibility effect" (Glenberg & Kaschak, 2002). It was found that the sensibility of a sentence is modified by physical actions. Reaction times increased when subjects read "toward sentences" that implied action toward the reader, such as "Open the drawer" *and* had to give the answer through an incongruent action, i.e., moving the hand away from the body. Conversely, when subjects answered through an action congruent with the sentence, reaction times decreased. Glenberg and Kaschak interpreted the result as indicating that understanding a sentence is dependent on structures usually used for action. Readers interested in more comprehensive reviews of the coupling between language and action/perception are referred to Glenberg and Kaschak (2003) or Zwaan (2004).

Summary and conclusions

This paper has presented an emerging framework of simulation based on terminology and ideas from control theory and data from neurophysiological and neuroimaging studies that explain higher-level cognitive processes as at least partly based on reactivations of sensorimotor structures of the brain. By reactivating mechanisms used in perception and action together with a predictive mechanism a flexible inner world emerges that can be used for many different higher-level cognitive tasks (cf. Grush, in press; Hesslow, 2002). Crucial to the embodiment of cognition, according to this account, is perhaps not so much the physical nature of a cognizer's body, or its interaction with the environment as such, but the relation between sensorimotor and higher-level cognitive processes, more specifically, the way that the latter are fundamentally based on and rooted in the former.

Although corroborating evidence comes from several disciplines, the simulation account is not yet a well established or coherent theory of cognition in general, and there are many questions still to be answered. For example, in current accounts it is unclear exactly what constitutes the difference between an executed, overt action and a simulated/imagined, covert one. Can this be accounted for in terms of simulation theories or are other, presumably higher-level, mechanisms required after all to selectively trigger one or the other?

Moreover, there is a *level of granularity* problem (cf. Meltzoff & Prinz, 2002) that seems to apply to many simulation accounts. That is, at what level of abstraction does the simulation occur? During imagery it seems that the simulation occurs on a low-level including very many of the aspects of actually perceiving or acting, as indicated by neuroimaging studies (e.g., Jeannerod, 2001), whereas in problem solving, as in the ToL task, more abstract aspects of actions may be employed, which might be supported by the finding that problem solving activity in the ToL seem to activate only higher motor centers, such as prefrontal and premotor cortex (Dagher et al., 1999). However, this is a speculative interpretation of the neuroimaging results.

The level of granularity is also an important issue in robotic models of simulation theories (e.g. Ziemke, Jirenhed & Hesslow, in press). Previous work in our lab has dealt with internal simulations at the lowest level, but current work also addresses simulation at more abstract levels of granularity. We believe that robotic models offer a fruitful approach to further investigating this and other open questions that need to be answered in the future development of simulation theories.

References

Anderson, M. L. (2003). Embodied cognition: a field guide. *Artificial Intelligence, 149*, 91-130.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., & Ruppert, J. M. (2003). Social embodiment. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43). San Diego: Academic Press.

Blakemore, S.-J., Frith, C. D. & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience, 11*, 551-559.

Blakemore, S.-J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences, 6*, 237-242.

Block, N. (1983). Mental pictures and cognitive science. *Philosophical Review, 93*, 499-542.

Chrisley, R., & Ziemke, T. (2003). Embodiment. In: *Encyclopedia of Cognitive Science*. London: Macmillan.

Clancey, W. J. (1997). *Situated Cognition.* Cambridge: Cambridge University Press.

Clark, A. (1997). *Being There*. Cambridge, MA: MIT Press.

Clark, A. & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior, 7*(1), 5-16.

Dagher, A., Owen, A. M., Boecker, H. & Brooks, D. J. (1999). Mapping the network for planning. *Brain, 122*, 1973-1987.

Decety, J. (1996). Do imagined and executed actions share the same neural substrate? *Cognitive Brain Res., 3*, 87-93.

Decety, J. (2002). Is there such a thing as functional equivalence between imagined, observed, and executed action. In M. A. Meltzoff & W. Prinz (Eds.), *The Imitative Mind.* Cambridge: Cambridge University Press.

Decety, J, and Chaminade, T. 2003. Neural correlates of feeling sympathy. *Neuropsychologia, 41*, 127-138.

Decety, J., & Jeannerod, M. (1996). Mentally simulated movements in virtual reality. *Behavioral Brain Research, 72*, 127-134.

Decety, J., Jeannerod, M., Prablanc, C. (1989). The timing of mentally represented actions. *Behavioral Brain Research, 34*, 35-42.

Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other. *Trends in Cognitive Sciences, 7,* 527-533.

Desmurget, M., & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences, 4*, 423-431.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G. (1992). Understanding motor events. *Experimental Brain Research, 91*, 176-180.

Dreyfus, H. L. (2002). Intelligence without representation. *Phenomenology and the Cognitive Sciences, 1*, 367-383.

Fadiga, L., Fogassi, L., Pavesi, G. & Rizzolatti, G (1995) Motor facilitation during action observation. *Journal of Neurophysiology, 73*, 2608-2611.

Farah, M. J. (1988). Is visual imagery really visual? *Psychological Review, 95*(3), 307-317

Finke, R. A. (1989). *Principles of mental imagery.* Cambridge, MA: MIT Press.

Frith, C. & Dolan, R. (1996). The role of the prefrontal cortex in higher cognitive functions. *Cognitive Brain Research, 5*, 175-181.

Gallese, V. (2003) A neuroscientific grasp of concepts. *Phil. Trans. Royal Soc. London, B., 358,* 1231-1240.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind reading. *Trends in Cognitive Sciences, 2*, 493-501.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, 9, 558-565.

Glenberg, A. M. & Kaschak, M. P. (2003) The body's contribution to language. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43). San Diego: Academic Press.

Grafton, S. T., Arbib, M. A., Fadiga, L. & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography - 2. Observation compared with imagination. *Experimental Brain Research, 112*, 103-111.

Grush, R. (in press). The emulation theory of representation. *Behavioral and Brain Sciences*, to appear.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences, 6*, 242-24.

Ingvar, D. H. & Philipsson, L. (1977). Distribution of the cerebral blood flow in the dominant hemisphere during motor ideation and motor performance. *Annals of Neurology, 2*, 230-237.

Jeannerod, M. (1994). The representing brain. *Behavioral and Brain Sciences, 17*(2), 187-245.

Jeannerod, M. (1997). *The cognitive neuroscience of action.* Cambridge, MA: Blackwell Publishers.

Jeannerod, M. (2001). Neural simulation of action. *NeuroImage, 14*, S103-S109.

Jeannerod, M. & Decety, J. (1995). Mental motor imagery. *Current Opinion in Neurobiology, 5*, 727-732.

Jeannerod, M. & Frak, V. (1999). Mental imagining of motor activity in humans. *Current Opinion in Neurobiology, 9*, 735-739.

Lakoff, G. (1988). Cognitive Semantics. In U. Eco *et al.* (Eds.), *Meaning and Mental Representations*. Bloomington: Indiana University Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh.* New York: Basic Books.

Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and Cognition.* Dordrecht: D. Reidel Publishing.

Maturana, H. R. & Varela, F. J. (1987). *The Tree of Knowledge*. Boston, MA: Shambhala.

Meltzoff, A. N., & Prinz, W. (2002). An introduction to the imitative mind and brain. In M. A. Meltzoff & W. Prinz (Eds.), *The Imitative Mind.* Cambridge: Cambridge University Press.

Nielsen, L. (2002). The simulation of emotion experience. *Phenomenology and the Cognitive Sciences,* 1, 255-286.

Rizzolatti, G. & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188-194.

Rizzolatti, G., Fadiga, L., Fogassi, L., & Gallese, V. (2002). From mirror neurons to imitation. In M. A. Meltzoff & W. Prinz (Eds.), *The imitative mind.* Cambridge: Cambridge University Press.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131-141.

Schall, U., Johnston, P., Lagopoulos, J., Jüptner, M., Jentzen, W., Thienel, Dittmann-Balcar, A., Bender, S., & Ward, P. B. (2003). Functional brain maps of tower of London performance. *NeuroImage, 20*, 1154-1161.

Shallice T. (1982). Specific impairments of planning. *Phil. Trans. Royal Soc. London, B, 298*, 199-209.

Shepard, R. N. & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Thomas, N. J. T. (1999). Are theories of imagery theories of imagination? *Cognitive Science, 23*, 207-245.

Varela, F. J., Thompson, E. & Rosch, E. (1991). *The embodied mind.* Cambridge, MA: MIT Press.

Wells, G. L. & Petty, R. E. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility responses. *Basic and Applied Social Psychology, 1*, 219-230.

Wilson, M. (2002) Six views of embodied cognition. *Psychonomic Bulletin and Review, 9*(4), 625-636.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*, 1317-1329.

Ziemke, T. (Ed.) (2002). *Situated and embodied cognition* (special issue). *Cognitive Systems Research*, 3(3).

Ziemke, T. (2003). What's that thing called embodiment? In: *Proc. of the 25th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Ziemke, T., Jirenhed, D.-A. & Hesslow, G. (in press) Internal simulation of perception: A minimal neuro-robotic model. *Neurocomputing*, to appear.

Zwaan, R.A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In: B.H. Ross (Ed.), *The Psychology of Learning and Motivation*, Vol. 44. New York: Academic Press.

Computationally Recognizing Wordplay in Jokes

Julia M. Taylor (tayloj8@email.uc.edu)
Lawrence J. Mazlack (mazlack@uc.edu)
Electrical & Computer Engineering and Computer Science Department
University of Cincinnati

Abstract

In artificial intelligence, researchers have begun to look at approaches for computational humor. Although there appears to be no complete computational model for recognizing verbally expressed humor, it may be possible to recognize jokes based on statistical language recognition techniques. This is an investigation into computational humor recognition. It considers a restricted set of all possible jokes that have wordplay as a component and examines the limited domain of "Knock Knock" jokes. The method uses Raskin's theory of humor for its theoretical foundation. The original phrase and the complimentary wordplay have two different scripts that overlap in the setup of the joke. The algorithm deployed learns statistical patterns of text in N-grams and provides a heuristic focus for a location of where wordplay may or may not occur. It uses a wordplay generator to produce an utterance that is similar in pronunciation to a given word, and the wordplay recognizer determines if the utterance is valid. Once a possible wordplay is discovered, a joke recognizer determines if a found wordplay transforms the text into a joke.

Introduction

Thinkers from the ancient time of Aristotle and Plato to the present day have strived to discover and define the origins of humor. Most commonly, early definitions of humor relied on laughter: what makes people laugh is humorous. Recent works on humor separate laughter and make it its own distinct category of response. Today there are almost as many definitions of humor as theories of humor; as in many cases, definitions are derived from theories (Latta, 1999). Some researchers say that not only is there no definition that covers all aspects of humor, but also humor is impossible to define (Attardo, 1994).

Humor is an interesting subject to study not only because it is difficult to define, but also because sense of humor varies from person to person. The same person may find something funny one day, but not the next, depending on the person's mood, or what has happened to him or her recently. These factors, among many others, make humor recognition challenging.

Although most people are unaware of the complex steps involved in humor recognition, a computational humor recognizer has to consider all these steps in order to approach the same ability as a human being.

A common form of humor is verbal, or "verbally expressed, humor" (Ritchie 2000). Verbally expressed humor involves reading and understanding texts. While understating the meaning of a text may be difficult for a computer, reading it is not.

One of the subclasses of verbally expressed humor is the joke. Hetzron (1991) defines a joke as "a short humorous piece of literature in which the funniness culminates in the final sentence." Most researchers agree that jokes can be broken into two parts, a setup and a punchline. The setup is the first part of the joke, usually consisting of most of the text, which establishes certain expectations. The punchline is a much shorter portion of the joke, and it causes some form of conflict. It can force another interpretation on the text, violate an expectation, or both (Ritchie, 1998). As most jokes are relatively short, it may be possible to recognize them computationally.

Computational recognition of jokes may be possible, but it is not easy. An "intelligent" joke recognizer requires world knowledge to "understand" most jokes.

Theories of Humor

Raskin's (1985) *Semantic Theory of Verbal Humor* has strongly influenced the study of verbally expressed humor. The theory is based on assumption that every joke is compatible with two scripts, and those two scripts oppose each other in some part of the text, usually in the punch line, therefore generating humorous effect.

Another approach is Suls' (1972) two-stage model, which is based on false expectation. The following algorithm is used to process a joke using two-stage model (Ritchie, 1999):

- As a text is read, make predictions
- While no conflict with prediction, keep going
- If input conflicts with prediction:
  - If not ending – PUZZLEMENT
  - If is ending, try to resolve:
    - No rules found – PUZZLEMENT
    - Cognitive rules found –HUMOR

There have been attempts at joke generation (Attardo, 1996; Binsted, 1996; Lessard and Levison, 1992; McDonough, 2001; McKay, 2002; Stock and Strapparava, 2002) and pun recognizers (Takizawa, et al. 1996; Yokogawa, 2002) for Japanese. However, there do not appear to be any theory based computational humor efforts. This may be partly due to the absence of a theory that can be expressed as an unambiguous computational algorithm. In the cases of Raskin and Suls, the first does not offer any formal algorithm, and the second does not specify what a cognitive rule is, leaving one of the major steps open to interpretation.

Wordplay Jokes

Wordplay jokes, or jokes involving verbal play, are a class of jokes depending on words that are similar in sound, but are used in two different meanings. The difference between the two meanings creates a conflict or breaks expectation,

and is humorous. The wordplay can be created between two words with the same pronunciation and spelling, with two words with different spelling but the same pronunciation, and with two words with different spelling and similar pronunciation. For example, in $Joke_1$ the conflict is created because the word has two meanings, while the pronunciation and the spelling stay the same. In $Joke_2$ the wordplay is between words that sound nearly alike.

$Joke_1$: "Cliford: The Postmaster General will be making the TOAST.
Woody: Wow, imagine a person like that helping out in the kitchen!"
$Joke_2$: "Diane: I want to go to Tibet on our honeymoon.
Sam: Of course, we will go to bed."[1]

Sometimes it takes world knowledge to recognize which word is subject to wordplay. For example, in $Joke_2$, there is a wordplay between "Tibet" and "to bed." However, to understand the joke, the wordplay by itself is not enough, a world knowledge is required to "link" honeymoon with "Tibet" and "to bed."

A focused form of wordplay jokes is the Knock Knock joke. In Knock Knock jokes, wordplay is what leads to the humor. The structure of the Knock Knock joke provides pointers to the wordplay.

A typical Knock Knock (KK) joke is a dialog that uses wordplay in the punchline. Recognizing humor in a KK joke arises from recognizing the wordplay. A KK joke can be summarized using the following structure:

$Line_1$: "Knock, Knock"
$Line_2$: "Who's there?"
$Line_3$: any phrase
$Line_4$: $Line_3$ followed by "who?"
$Line_5$: One or several sentences containing one of the following:
$Type_1$: $Line_3$
$Type_2$: a wordplay on $Line_3$
$Type_3$: a meaningful response to $Line_3$.

$Joke_3$ is an example of $Type_1$, $Joke_4$ is an example of $Type_2$, and $Joke_5$ is an example of $Type_3$.

$Joke_3$: Knock, Knock
Who's there?
Water
Water who?
Water you doing tonight?
$Joke_4$: Knock, Knock
Who's there?
Ashley
Ashley who?
Actually, I don't know.
$Joke_5$: Knock, Knock
Who's there?
Tank
Tank who?
You are welcome.[2]

From theoretical points of view, both Raskin's (1985) and Suls' (1972) approaches can explain why $Joke_3$ is a joke. Following Raskin's approach, the two belong to different scripts that overlap in the phonetic representation of "water," but also oppose each other. Following Suls' approach, "what are" conflicts with the prediction. In this approach, a cognitive rule can be described as a function that finds a phrase that is similar in sound to the word "water," and that fits correctly in beginning of the final sentence's structure. This phrase is "what are" for $Joke_3$.

N-grams

A joke generator has to have an ability to construct meaningful sentences, while a joke recognizer has to recognize them. While joke generation involves limited world knowledge, joke recognition requires a much more extensive world knowledge.

To be able to recognize or generate jokes, a computer should be able to "process" sequences of words. A tool for this activity is the N-gram, "one of the oldest and most broadly useful practical tools in language processing" (Jurafsky and Martin, 2000). An N-gram is a model that uses conditional probability to predict N^{th} word based on N-1 previous words. N-grams can be used to store sequences of words for a joke generator or a recognizer.

N-grams are typically constructed from statistics obtained from a large corpus of text using the co-occurrences of words in the corpus to determine word sequence probabilities (Brown, 2001). As a text is processed, the probability of the next word N is calculated, taking into account end of sentences, if it occurs before the word N.

"The probabilities in a statistical model like an N-gram come from the corpus it is trained on. This training corpus needs to be carefully designed. If the training corpus is too specific to the task or domain, the probabilities may be too narrow and not generalize well to new sentences. If the training corpus is too general, the probabilities may not do a sufficient job of reflecting the task or domain" (Jurafsky and Martin, 2000).

A bigram is an N-gram with N=2, a trigram is an N-gram with N=3, etc. A bigram model will use one previous word to predict the next word, and a trigram will use two previous words to predict the word.

Experimental Design

A further tightening of the focus was to attempt to recognize only $Type_1$ of KK jokes. The original phrase, in this case $Line_3$, is referred to as the *keyword*.

There are many ways of determining "sound alike" short utterances. The only feasible method for this project was computationally building up "sounds like" utterances as needed.

The joke recognition process has four steps:

$Step_1$: joke format validation
$Step_2$: generation of wordplay sequences
$Step_3$: wordplay sequence validation
$Step_4$: last sentence validation

Once $Step_1$ is completed, the wordplay generator generates utterances, similar in pronunciation to $Line_3$. $Step_3$ only checks if the wordplay makes sense without touching the rest of the punchline. It uses a bigram table for validation. Only meaningful wordplays are passed to $Step_4$ from $Step_3$.

[1] $Joke_1$, $Joke_2$ are taken from TV show "Cheers"
[2] http://www.azkidsnet.com/JSknockjoke.htm

If the wordplay is not in the end of the punchline, $Step_4$ takes the last two words of the wordplay, and checks if they make sense with the first two words of text following the wordplay in the punchline, using two trigram sequences. If the wordplay occurs in the end of the sentence, the last two words before the wordplay and the first two words of the wordplay are used for joke validation. If $Step_4$ fails, go back to $Step_3$ or $Step_2$, and continue the search for another meaningful wordplay.

It is possible that the first three steps return valid results, but $Step_4$ fails; in which case a text is not considered a joke by the Joke Recognizer.

The punchline recognizer is designed so that it does not have to validate the grammatical structure of the punchline. Moreover, it is assumed that the $Line_5$ is meaningful when the expected wordplay is found, if it is a joke; and, that $Line_5$ is meaningful as is, if the text is not a joke. In other words, a human expert should be able to either find a wordplay so that the last sentence makes sense, or conclude that the last sentence is meaningful without any wordplay. It is assumed that the last sentence is not a combination of words without any meaning.

The joke recognizer is to be trained on a number of jokes; and, tested on jokes, twice the number of training jokes. The jokes in the test set are previously "unseen" by the computer. This means that any joke, identical to the joke in the set of training jokes, is not included in the test set.

Generation of Wordplay Sequences

Given a spoken utterance A, it is possible to find an utterance B that is similar in pronunciation by changing letters from A to form B. Sometimes, the corresponding utterances have different meanings. Sometimes, in some contexts, the differing meanings might be humorous if the words were interchanged.

A repetitive replacement process is used for generation of wordplay sequences. Suppose, a letter a_1 from A is replaced with b_1 to form B. For example, in $Joke_3$ if a letter 'w' in a word 'water' is replaced with 'wh', 'e' is replaced with 'a', and 'r' is replaced with 're', the new utterance, 'what are' sounds similar to 'water'.

A table, containing combinations of letters that sound similar in some words, and their similarity value was used. The purpose of the Similarity Table is to help computationally develop "sound alike" utterances that have different spellings. In this paper, this table will be referred to as the Similarity Table. Table 1 is an example of the Similarity Table. The Similarity Table was derived from a table developed by Frisch (1996). Frisch's table contained cross-referenced English consonant pairs along with a similarity of the pairs based on the natural classes model. Frisch's table was heuristically modified and extended to the Similarity Table by "translating" phonemes to letters, and adding pairs of vowels that are close in sound. Other phonemes, translated to combinations of letters, were added to the table as needed to recognize wordplay from a set of training jokes.

The resulting Similarity Table approximately shows the similarity of sounds between different letters or between letters and combination of letters. A heuristic metric indicating how closely they sound to each other was either taken from Frisch's table or assigned a value close to the average of Frisch's similarity values. The Similarity Table should be taken as a collection of heuristic satisficing values that might be refined through additional iteration.

Table 1: Subset of entries of the Similarity Table, showing similarity of sounds in words between different letters

a	e	0.23
e	a	0.23
e	o	0.23
en	e	0.23
k	sh	0.11
l	r	0.56
r	m	0.44
r	re	0.23
t	d	0.39
t	z	0.17
w	m	0.44
w	r	0.42
w	wh	0.23

When an utterance A is "read" by the wordplay generator, each letter in A is replaced with the corresponding replacement letter from the Similarity Table. Each new string is assigned its similarity with the original word A.

All new words are inserted into a heap, ordered according to their similarity value, greatest on top. When only one letter in a word is replaced, its similarity value is being taken from the Similarity Table. The similarity value of the strings is calculated using the following heuristic formula:

similarity of string = number of unchanged letters +
sum of similarities of each replaced entry from the table

Note, that the similarity values of letters are taken from the Similarity table. These values differ from the similarity values of strings.

Once all possible one-letter replacement strings are found, and inserted into the heap, according to the string similarity, the first step is complete.

The next step is to remove the top element of the heap. This element has the highest similarity with the original word. If this element can be decomposed into an utterance that makes sense, this step is complete. If the element cannot be decomposed, each letter of the string, except for the letter that was replaced originally, is being replaced again. All newly constructed strings are inserted into the heap according to their similarity. Continue with the process until the top element can be decomposed into a meaningful phrase, or all elements are removed from the heap.

Consider $Joke_3$ as example. The joke fits a typical KK joke pattern. The next step is to generate utterances similar in pronunciation to 'water.'

Table 2 shows some of the strings received after one-letter replacements of 'water' in $Joke_3$. The second column shows the similarity of the string in the first table with the original word.

Suppose, the top element of the heap is 'watel,' with the similarity value of 4.56. Watel cannot be decomposed into a meaningful utterance. This means that each letter of 'watel' except for 'l' will be replace again. The newly formed strings will be inserted into the heap, in the order of

their similarity value. The letter 'l' will not be replaced as it not the 'original' letter from 'water.' The string similarity of newly constructed strings will be most likely less than 4. (The only way a similarity of a newly constructed string is greater than 4 is if the similarity of the replaced letter is above 0.44, which is unlikely.) This means that they will be placed below 'wazer.' The next top string, 'mater,' is removed. 'Mater' is a word. However, it does not work in the sentence 'Mater you doing.' (See Sections on *Wordplay Recognition* and *Joke Recognition* for further discussion.) The process continues until 'whater' is the top string. The replacement of 'e' in 'whater' with 'a' will result in 'whatar'. Eventually, 'whatar' will become the top string, at which point 'r' will be replaced with 're' to produce 'whatare'. 'Whatare' can be decomposed into 'what are' by inserting a space between 't' and 'a'. The next step will be to check if 'what are' is a valid word sequence.

Table 2: Examples of strings received after replacing one letter from the word 'water' and their similarity value to 'water'

New String	String Similarity to 'Water'
watel	4.56
mater	4.44
watem	4.44
rater	4.42
wader	4.39
wather	4.32
watar	4.23
wator	4.23
whater	4.23
wazer	4.17

Generated wordplays that were successfully recognized by the wordplay recognizer, and their corresponding keywords are stored for the future use of the program. When the wordplay generator receives a new request, it first checks if wordplays have been previously found for the requested keyword. The new wordplays will be generated only if there is no wordplay match for the requested keyword, or the already found wordplays do not make sense in the new joke.

Wordplay Recognition

A wordplay sequence is generated by replacing letters in the keyword. The keyword is examined because: if there is a joke, based on wordplay, a phrase that the wordplay is based on will be found in $Line_3$. $Line_3$ is the keyword. A wordplay generator generates a string that is similar in pronunciation to the keyword. This string, however, may contain real words that do not make sense together. A wordplay recognizer determines if the output of the wordplay generator is meaningful.

A database with the bigram table was used to contain every discovered two-word sequence along with the number of their occurrences, also referred to as *count*. Any sequence of two words will be referred to as *word-pair*. Another table in the database, the trigram table, contains each three-word sequence, and the count.

The wordplay recognizer queries the bigram table. The joke recognizer, discussed in section on *Joke Recognition*, queries the trigram table.

To construct the database several focused large texts were used. The focus was at the core of the training process. Each selected text contained a wordplay on the keyword ($Line_3$) and two words from the punchline that follow the keyword from at least one joke from the set of training jokes. If more than one text containing a given wordplay was found, the text with the closest overall meaning to the punchline was selected. Arbitrary texts were not used, as they did not contain a desired combination of wordplay and part of punchline.

To construct the bigram table, every pair of words occurring in the selected text was entered into the table.

The concept of this wordplay recognizer is similar to an N-gram. For a wordplay recognizer, the bigram model is used.

The output from the wordplay generator was used as input for the wordplay recognizer. An utterance produced by the wordplay generator is decomposed into a string of words. Each word, together with the following word, is checked against the database.

An N-gram determines for each string the probability of that string in relation to all other strings of the same length. As a text is examined, the probability of the next word is calculated. The wordplay recognizer keeps the number of occurrences of word sequence, which can be used to calculate the probability. A sequence of words is considered valid if there is at least one occurrence of the sequence anywhere in the text. The count and the probability are used if there is more than possible wordplay. In this case, the wordplay with the highest probability will be considered first.

For example, in $Joke_3$ 'what are' is a valid combination if 'are' occurs immediately after 'what' somewhere in the text.

Joke Recognition

A text with valid wordplay is not a joke if the rest of the punchline does not make sense. For example, if the punchline of $Joke_3$ is replaced with "Water a text with valid wordplay," the resulting text is not a joke, even though the wordplay is still valid. Therefore, there has to be a mechanism that can validate that the found wordplay is "compatible" with the rest of the punchline and makes it a meaningful sentence.

A concept similar to a trigram was used to validate the last sentence. All three-word sequences are stored in the trigram table.

The same training set was used for both the wordplay and joke recognizers. The difference between the wordplay recognizer and joke recognizer was that the wordplay recognizer used pairs of words for its validation while the joke recognizer used three words at a time. As the training text was read, the newly read word and the two following words were inserted into the trigram table. If the newly read combination was in the table already, the count was incremented.

As the wordplay recognizer had already determined that the wordplay sequences existed, there was no reason to re-validate the wordplay.

To check if wordplay makes sense in the punchline, the last two words of the wordplay, $\mathbf{w_{wp1}}$ and $\mathbf{w_{wp2}}$, are used, for the wordplay that is at least two words long. If the punchline is valid, the sequence of $\mathbf{w_{wp1}}$, $\mathbf{w_{wp2}}$, and the first word of the remainder of the sentence, $\mathbf{w_s}$, should be found in the training text. If the sequence $<\mathbf{w_{wp1}}\ \mathbf{w_{wp2}}\ \mathbf{w_s}>$ occurs in the trigram table, this combination is found in the training set, and the three words together make sense. If the sequence is not in the table, either the training set is not accurate, or the wordplay does not make sense in the punchline. In either case, the computer does not recognize the joke. If the previous check was successful, or if the wordplay has only one word, the last check can be performed. The last step involves the last word of the word play, $\mathbf{w_{wp}}$, and the first two words of the remainder of the sentence, $\mathbf{w_{s1}}$ and $\mathbf{w_{s2}}$. If the sequence $<\mathbf{w_{wp}}\ \mathbf{w_{s1}}\ \mathbf{w_{s2}}>$ occurs in the trigram table, the punchline is valid, and the wordplay fits with the rest of the final sentences.

If the wordplay recognizer found several wordplays that "produced" a joke, the wordplay resulting in the highest trigram sequence probability was used.

Results and Analysis

A set of 65 jokes from the "111 Knock Knock Jokes" website[3] and one joke taken from "The Original 365 Jokes, Puns & Riddles Calendar" (Kostick, et al., 1998) was used as a training set. The Similarity Table, discussed in the Section on *Generation of Wordplay Sequences*, was modified with new entries until correct wordplay sequences could be generated for all 66 jokes. The training texts inserted into the bigram and trigram tables were chosen based on the punchlines of jokes from the set of training jokes.

The program was run against a test set of 130 KK jokes, and a set of 65 non-jokes that have a similar structure to the KK jokes.

The test jokes were taken from "3650 Jokes, Puns & Riddles" (Kostick, et al. 1998). These jokes had the punchlines corresponding to any of the three KK joke structures discussed earlier.

To test if the program finds the expected wordplay, each joke had an additional line, Line_6, added after Line_5. Line_6 is not a part of any joke. It only existed so that the wordplay found by the joke recognizer could be compared against the expected wordplay. Line_6 consists of the punchline with the expected wordplay instead of the punchline with Line_3.

The jokes in the test set were previously "unseen" by the computer. This means that if the book contained a joke, identical to the joke in the set of training jokes, this joke was not included in the test set.

Some jokes, however, were very similar to the jokes in the training set, but not identical. These jokes were included in the test set, as they were not the same. As it turned out, some jokes to a human may look very similar to jokes in the training set, but treated as completely different jokes by the computer.

Out of 130 previously unseen jokes the program was not expected to recognize eight jokes. These jokes were not expected to be recognized because the program is not expected to recognize their structure.

The program was able to find wordplay in 85 jokes, but recognized only seventeen jokes as such out of 122 that it could potentially recognize. Twelve of these jokes have the punchlines that matched the expected punchlines. Two jokes have meaningful punchlines that were not expected. Three jokes were identified as jokes by the computer, but their punchlines do not make sense to the investigator.

Some of the jokes with found wordplay were not recognized as jokes because the database did not contain the needed sequences. When a wordplay was found, but the needed sequences were not in the database, the program did not recognize the jokes as jokes.

In many cases, the found wordplay matched the intended wordplay. This suggests that the rate of successful joke recognition would be much higher if the database contained all the needed word sequences.

The program was also run with 65 non-jokes. The only difference between jokes and non-jokes was the punchline. The punchlines of non-jokes were intended to make sense with Line_3, but not with the wordplay of Line_3. The non-jokes were generated from the training joke set. The punchline in each joke was substituted with a meaningful sentence that starts with Line_3. If the keyword was a name, the rest of the sentence was taken from the texts in the training set. For example, Joke_6 became Text_1 by replacing "time for dinner" with "awoke in the middle of the night."

Joke_6: Knock, Knock
Who's there?
Justin
Justin who?
Justin time for dinner.

Text_1: Knock, Knock
Who's there?
Justin
Justin who?
Justin awoke in the middle if the night.

A segment "awoke in the middle of the night" was taken from one of the training texts that was inserted into the bigram and trigram tables.

The program successfully recognized 62 non-jokes.

Possible Extensions

The results suggest that most jokes were not recognized either because the texts entered did not contain the necessary information for the jokes to work; or because N-grams are not suitable for true "understanding" of text. One of the simpler experiments may be to test to see if more jokes are recognized if the databases contain more sequences. This would require inserting a much larger text into the trigram table. A larger text may contain more word sequences, which would mean more data for N-grams to recognize some jokes.

It is possible that no matter how large the inserted texts are, the simple N-grams will not be able to "understand" jokes. The simple N-grams were used to understand or to analyze the punchline. Most jokes were not recognized due to failures in sentence understanding. A more sophisticated tool for analyzing a sentence may be needed to improve the

[3] http://www.azkidsnet.com/JSknockjoke.htm

joke recognizer. Some of the options for the sentence analyzer are an N-gram with stemming or a sentence parser.

A simple parser that can recognize, for example, nouns and verbs; and analyze the sentence based on parts of speech, rather than exact spelling, may significantly improve the results. On the other hand, giving N-grams the stemming ability would make them treat, for example, "color" and "colors" as one entity, which may significantly help too.

The wordplay generator produced the desired wordplay in most jokes, but not all. After the steps are taken to improve the sentence understander, the next improvement should be a more sophisticated wordplay generator. The existing wordplay generator is unable to find wordplay that is based word longer than six characters, and requires more that three substitutions. A better answer to letter substitution is phoneme comparison and substitution. Using phonemes, the wordplay generator will be able to find matches that are more accurate.

The joke recognizer may be able to recognize jokes other than KK jokes, if the new jokes are based on wordplay, and their structure can be defined. However, it is unclear if recognizing jokes with other structures will be successful with N-grams.

Summary and Conclusion

Computational work in natural language has a long history. Areas of interest have included: translation, understanding, database queries, summarization, indexing, and retrieval. There has been very limited success in achieving true computational understanding.

A focused area within natural language is verbally expressed humor. Some work has been achieved in computational generation of humor. Little has been accomplished in understanding. There are many linguistic descriptive tools such as formal grammars. But, so far, there are not robust understanding tools and methodologies.

The KK joke recognizer is the first step towards computational recognition of jokes. It is intended to recognize KK jokes that are based on wordplay. The recognizer's theoretical foundation is based on Raskin's Script-based Semantic Theory of Verbal Humor that states that each joke is compatible with two scripts that oppose each other. The $Line_3$ and the wordplay of $Line_3$ are the two scripts. The scripts overlap in pronunciation, but differ in meaning.

The joke recognition process can be summarized as:

$Step_1$: joke format validation
$Step_2$: generation of wordplay sequences
$Step_3$: wordplay sequence validation
$Step_4$: last sentence validation

The result of KK joke recognizer heavily depends on the choice of appropriate letter-pairs for the Similarity Table and on the selection of training texts.

The KK joke recognizer "learns" from the previously recognized wordplays when it considers the next joke. Unfortunately, unless the needed (keyword, wordplay) pair is an exact match with one of the found (keyword, wordplay) pairs, the previously found wordplays will not be used for the joke. Moreover, if one of the previously recognized jokes contains (keyword, wordplay) pair that is needed for the new joke, but the two words that follow or precede the keyword in the punchline differ, the new joke may not be recognized regardless of how close the new joke and the previously recognized jokes are.

The joke recognizer was trained on 66 KK jokes; and tested on 130 KK jokes and 66 non-jokes with a structure similar to KK jokes.

The program successfully found and recognized wordplay in most of the jokes. It also successfully recognized texts that are not jokes, but have the format of a KK joke. It was not successful in recognizing most punchlines in jokes. The failure to recognize punchline is due to the limited size of texts used to build the trigram table of the N-gram database.

While the program checks the format of the first four lines of a joke, it assumes that all jokes that are entered have a grammatically correct punchline, or at least that the punchline is meaningful. It is unable to discard jokes with a poorly formed punchline. It may recognize a joke with a poorly formed punchline as a meaningful joke because it only checks two words in the punchline that follow $Line_3$.

In conclusion, the method was reasonably successful in recognizing wordplay. However, it was less successful in recognizing when an utterance might be valid.

References

Attardo, S. (1994) *Linguistic Theories of Humor*. Berlin: Mouton de Gruyter

Binsted, K. (1996) *Machine Humour: An Implemented Model Of Puns*. Doctoral dissertation, University of Edinburgh

Frisch, S. (1996) *Similarity And Frequency In Phonology*. Doctoral dissertation, Northwestern University

Hetzron, R. (1991) On The Structure Of Punchlines. *HUMOR: International Journal of Humor Research,* 4:1

Jurafsky, D., & Martin, J. (2000) *Speech and Language Processing*, New Jersey: Prentice-Hall

Kostick, A., Foxgrover, C., & Pellowski, M. (1998) *3650 Jokes, Puns & Riddles*. New York: Black Dog & Leventhal Publishers

Latta, R. (1999) *The Basic Humor Process*. Berlin: Mouton de Gruyter

Lessard, G., & Levison, M. (1992) Computational Modelling Of Linguistic Humour: Tom Swifties. *ALLC/ACH Joint Annual Conference*, Oxford

McDonough, C. (2001) *Mnemonic String Generator: Software To Aid Memory Of Random Passwords*. CERIAS Technical report, West Lafayette, IN

McKay, J. (2002) Generation Of Idiom-based Witticisms To Aid Second Language Learning. *Proceedings of Twente Workshop on Language Technology 20*, University of Twente

Raskin, V. (1985) *The Semantic Mechanisms Of Humour*, Dordrecht: Reidel

Ritchie, G. (1999) Developing The Incongruity-Resolution Theory. *Proceedings of AISB 99 Symposium on Creative Language: Humour and Stories*, Edinburgh

Ritchie, G. (2000) Describing Verbally Expressed Humour. *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, Birmingham

Stock, O., & Strapparava, C. (2002) Humorous Agent For Humorous Acronyms: The HAHAcronym Project. *Proceedings of Twente Workshop on Language Technology 20*, University of Twente

Suls, J. (1972) A Two-Stage Model For The Appreciation Of Jokes And Cartoons: An Information-Processing Analysis. In J. H. Goldstein and P. E. McGhee (Eds.) *The Psychology Of Humor* NY: Academic Press

Takizawa, O., Yanagida, M., Ito, A., & Isahara, H. (1996) On Computational Processing Of Rhetorical Expressions - Puns, Ironies And Tautologies. *Proceedings of Twente Workshop on Language Technology 12*, University of Twente

Yokogawa, T. (2002) Japanese Pun Analyzer Using Articulation Similarities. *Proceedings of FUZZ-IEEE*, Honolulu

On the Usefulness and Limitations of Diagrams in Statistical Training

Atsushi Terao (atsushi@edu.hokudai.ac.jp)
Graduate School of Education, Hokkaido University
Kita 11 Nishi 7, Sapporo 060-0811 Japan

Abstract

The purpose of this study was to examine the usefulness and limitations of vector diagrams, consisting of lines with arrows representing variables, in statistical training. Nineteen undergraduates learned advanced level statistics either with vector diagrams or in the conventional way and solved three problems. Vector diagrams sometimes helped the students understand descriptions in the text which were difficult in conventional explanations, but caused other difficulties. Vector diagrams were useful for solving one of the three problems, but not the other two. It is concluded that a property of diagrams or formulae can be a double-edged sword.

Students who are majoring in psychology or other relevant disciplines have to study statistics. Despite substantial effort by teachers, understanding statistics is often difficult for many students. This paper reports the results of a practical experiment in which the students learned to employ either "vector diagrams" or a conventional formula-based approach to the basics of regression analysis. The students were then asked to solve three problems using the given technique they learned.

Unlike many previous studies on using diagrams in educational settings, which focus only on the usefulness of diagrams, this study also investigates limitations of diagrams. Research on diagrammatic reasoning has found many "good" properties of diagrams (e.g., Barwise & Etchmendy, 1996; Cheng & Simon, 1995; Larkin & Simon, 1987). The researchers seem to consider these properties as if they are always support (at least do not impair) understanding and problem solving. The results of this study suggest that the same property, which definitely makes the solution of a problem easy, sometimes makes another problem difficult. Similarly, the results suggest that formulae do not necessarily have "bad" properties.

The vector diagrams used in this study consist of several vectors drawn as lines with arrows, each of which corresponds to a variable. For example, the correlation coefficient is defined as $\cos\theta$ where θ is the angle between two vectors, $\vec{x} = \{x_1 - \bar{x}, \cdots, x_n - \bar{x}\}$ and $\vec{y} = \{y_1 - \bar{y}, \cdots, y_n - \bar{y}\}$. The regression analysis is described as the projection of the dependent variable (actually the vector of the dependent variable like $\vec{y}$ in Figure 1) on the linear space of independent variables ($\vec{x_1}$ and $\vec{x_2}$ in Figure 1).

Tasks and Prediction

These are two basic assumptions in this study about the nature of diagrammatic and algebraic representations and operators.

One assumption is that the diagrammatic representation of a problem affords a far smaller number of operators than the algebraic representations of that problem. This assumption is two-fold. First, the problem space (the set of all possible problem states) is smaller when a problem is represented by a diagram. In the chapters on vectors in mathematics textbooks, the only diagrammatic operators one can commonly find are: extension (or reduction), rotation, projection, and (de)composition. Algebraic representations, by contrast, allow many kinds of manipulations such as the four basic operations of arithmetic, expansion or factorization, fraction operations, root operations (e.g., $\sqrt{a} * \sqrt{b} = \sqrt{ab}$), summation operations (e.g., $\sum(a + b) = \sum a + \sum b$), substitution, and so on. Second, I assume that the search space (the set of problem states a student actually considers), is also smaller when a problem is represented by a diagram. Whether a problem is represented by a diagram or a formula, students do not consider all possible operators because in each case some operators are difficult to use.

The other assumption is that a formula and its transformation become more concrete when they are connected to a diagram. This assumption seems to have no room for doubt because connecting a formula to a diagram increases the number of attributes the formula has. This assumption is related to the first assumption. Because of the limited number of diagrammatic operators, "diagrammatic inference" often requires using algebraic representations, although the diagrams play a crucial role in the inference. This means that students often have to do "heterogeneous inference," inference that use multiple forms of representation (Barwise & Etchemendy, 1996).

This study claims that any property of diagrams or formulae can be either a help or a hindrance in problem solving. For example, considering only a

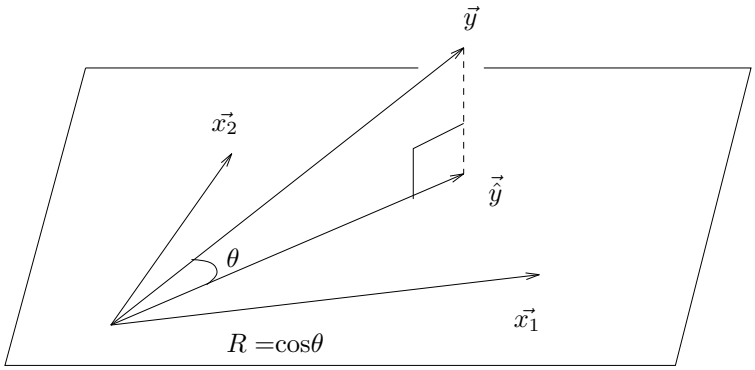


Figure 1: The definition of the multiple correlation coefficient R using a vector diagram

Table 1: Three Problems used in This Study.

Problem 1: The multiple correlation coefficient R indicates the goodness of fit of the model in multiple regression analysis. Consider the multiple correlation coefficient "R" in the case of simple regression. Please explain the relationship between this R and r, the simple correlation coefficient for the two variables x and y.
Problem 2: In the case of simple regression, if the number of paired values (x_i, y_i) is two, we can describe the values of one variable by using the other variable without any error, indicating $r = +1$ or -1. Give an explanation for the reason of this perfect description.
Problem 3: Explain the relationship between the regression coefficient $\hat{a_1}$ ($= S_{xy}/S_{xx}$) and the correlation coefficient r in the case of simple regression by using only the two variances S_{xx} and S_{yy}. S_{xy} means the covariance for the two variables x and y.

small number of diagrammatic operators can serve either as a constraint on the search (This is expected to be the case in Problem 1 in Table 1, as described below), or as a limitation if the correct solution path is outside of the search space (This is expected to be the case in Problem 2). The abstractness of formulae also can be an advantage or a disadvantage (This contrast is expected to be shown between Problem 1 and Problem 3).

Problem 1 If students learn regression analysis in a conventional way, R is defined by the formula

$$R = \frac{\frac{1}{n}\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\frac{1}{n}\sum(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum(\hat{y}_i - \bar{\hat{y}})^2}}.$$

This formula means that R is defined as the correlation coefficient between the expected value $\hat{y}$ and the observed value y. In the case of simple regression, we can say $\hat{y}_i = a_0 + a_1 x_{i1}$ and $\bar{\hat{y}} = a_0 + a_1 \bar{x}_1$. If these relations are used in the formula for R then the conclusion $R = |r|$ is obtained after a long series of algebraic manipulations.

Using a vector diagram, R is defined as $\cos\theta$ for two vectors $\vec{\hat{y}} = \{\hat{y}_1 - \bar{\hat{y}}, \cdots, \hat{y}_n - \bar{\hat{y}}\}$ and $\vec{y} = \{y_1 - \bar{y}, \cdots, y_n - \bar{y}\}$ as shown in Figure 1. Vector $\vec{\hat{y}}$ is the orthogonal projection of $\vec{y}$ on the plane spanned by $\vec{x}_1$ and $\vec{x}_2$. If we delete $\vec{x}_2$ from Figure 1 and redraw the orthogonal projection, we will obtain the answer as shown in Figure 2.

According to the basic assumption mentioned above, the problem/search space of the diagrammatic version of this problem is assumed to be smaller than the problem/search space of the formula version. The solution with vector diagram, consequently, should require less computation than the conventional solution. Note that in both solutions we started with the definition of R. In general, diagrammatic approaches often require less computation than conventional approaches (Cheng, 1992).

Other than the small problem/search space of the diagrammatic solution, the concreteness of diagrammatic operators also can contribute to finding the answer to this problem. In contrast, the formula version of the definition of R and its transformation are more abstract with no diagrammatic meaning, and the solution is a pure algebraic solution. Cheng and Simon (1995) pointed out that conventional mathematical approaches are often more complex than diagrammatic approaches because the bulk of the reasoning must center around abstract equations.

We therefore predict that a group of students which uses a vector diagram to solve this problem will show better performance than another group of students which tries to solve it in a conventional way.

In this study, after trying to solve the problems, the students read the correct solution and evaluated

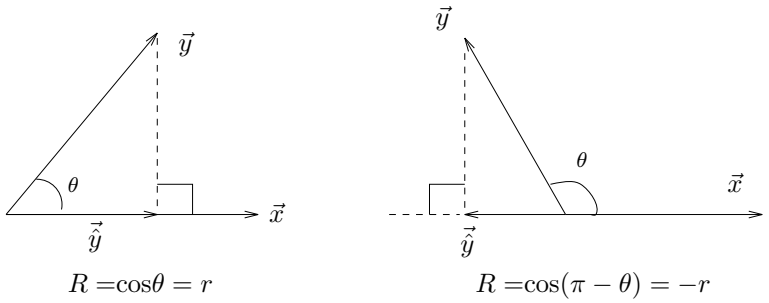


Figure 2: The answer to Problem 1 with vector diagrams

the degree of their understanding (to what degree they could understand the solution) and the degree of their conviction (to what degree they could accept it) on a 5-point scale. In Problem 1, we can expect some differences in these scores in accordance with the difference in difficulty of problem solving. However, there may be no difference in these evaluations between the two groups of the students. Even if it is difficult to make a long sequence of appropriate operators by their own efforts, just following the correct sequence may not be a tough task as long as the students are familiar with these operators.

Problem 2 This problem was chosen in this study to show that the limited number of diagrammatic operators, which is the property of vector diagrams considered to make Problem 1 easy, can also be a hindrance in problem solving. Among diagrammatic operators one can find in mathematics textbooks, I assume that the decomposition of a vector is relatively difficult to use for students because the pair of vectors that would be generated does not exist in the current problem state. If a diagrammatic solution of a problem requires students to use the decomposition operator, the correct solution path is likely to be outside of the search space, although this path is in the problem space. A crucial difference between the vector solutions of Problem 1 and 2 is in whether the correct solution path is within the search space or not, although some other differences remain uncontrolled. This experiment puts the external validity above the internal validity, and it is difficult in this kind of practical study to strictly control all factors.

In many conventional textbooks, the correlation coefficient is explained with a scatter diagram. In the case of Problem 2, two points will be plotted on the scatter diagram. The regression straight line is uniquely specified because two points define a unique line. For this problem, the comparison is not diagram vs. algebra but vector diagram vs. conventional way.

A vector diagram which can be used to solve this problem is shown in Figure 3. The vector $\vec{x}$ lies at right angles to the vector $(\bar{x}, \bar{x})$; the vector $\vec{y}$ lies at right angles to the vector $(\bar{y}, \bar{y})$. Students could find these spatial relations by drawing a diagram of concrete data chosen arbitrarily or by calculating the inner product. The fact that the vector $\vec{x}$ is parallel to the vector $\vec{y}$ means that the vector $\vec{y}$ is described as $\alpha\vec{x}$ where α is a scalar. Note that this solution needs only one kind of diagrammatic operator: Participants need to decompose each of (x_1, x_2) and (y_1, y_2) into two vectors as shown in Figure 3.

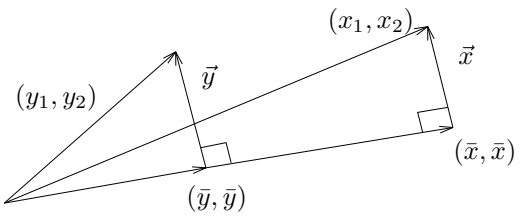


Figure 3: A vector diagram used to solve Problem 2

Our prediction is that the difficulty of using the decomposition operator will impair performance of the students. It is also expected that these students will have trouble in understanding and accepting the correct solution because the decomposition would be outside of the search space. This is contrary to the case of algebraic solution of Problem 1 because all problem states in this solution are expected to be included in the search space.

Problem 3 This problem was chosen to use in this experiment to show that the abstractness of algebraic solutions can sometimes help problem solving. Recall that it is thought that this property of formulae would make difficult the algebraic solution of Problem 1.

A conventional solution to this problem consists of a sequence of simple transformations of the equation defining the regression coefficient:

$$\hat{a}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r_{xy} \times \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}.$$

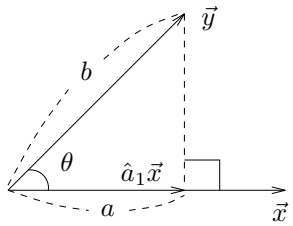


Figure 4: A vector diagram used to solve Problem 3

The transformations used in this solution look very formal and it is difficult to find any concrete meaning in them. For example, $\sqrt{S_{yy}}$ is forced to be put into the formula but this looks like a manipulation without any concrete meaning.

This problem represents a class of problems which could not be solved by purely diagrammatic thinking; rather, it requires heterogeneous inference, recruiting both diagrammatic and algebraic representations. Figure 4 shows a vector diagram which can be used to solve this problem. If two relations $a = \hat{a}_1\sqrt{\sum(x_i - \overline{x})^2}$ and $b = \sqrt{\sum(y_i - \overline{y})^2}$ are put into the equation $r = \cos\theta = \frac{a}{b}$, we will obtain the correct answer. The diagram gives some concrete meaning to this solution (the second basic assumption in this study mentioned above).

The prediction is that the students who use a purely algebraic approach will show better performance than the students who try to use a vector diagram. As mentioned in Problem 1, conventional algebraic approaches are often more complex than diagrammatic approaches because the bulk of the reasoning must center around abstract equations (Cheng & Simon, 1995). This study, however, claims that the abstractness of algebraic manipulation is not a "bad" property of formulae by nature. Heterogeneous inference requires students to use multiple representations simultaneously and it can burden students with a cognitive load. A pure diagrammatic solution, if any, is thought to be easier if this solution is as simple as the algebraic part of the heterogeneous inference.

We can expect some differences in the score of understanding and acceptance in accordance with the difference in difficulty of problem solving. However, similar to the case of Problem 1, there may be no difference in these evaluations between the two groups of students because just following the solution steps might not be very difficult.

Method

Participants

Participants were 19 undergraduate students majoring in psychology at Wakayama University, Japan. They had all taken or were taking a first introductory statistics course for psychology students, but did not know about the regression analysis taught in this experiment. In Japan, most of the undergraduate students learn algebra and vectors in high schools. This means they are ready for learning statistics either by conventional method or by an alternative, diagrammatic method.

Design

There were two experimental groups. In *formula* group, participants studied the basics of regression analysis in the conventional way in which formulae were mainly used. In *vector* group, the basics of regression analysis were taught with vector diagrams. The participants were assigned to one of these groups.

Before this experiment, participants received a simple pretest, the purpose of which was to evaluate their basic knowledge of statistics. This pretest consisted of five items, which required students to write formulae of mean, variance, SD, covariance, and correlation coefficient. Students got one point for each correct answer giving a maximum score of 5 points.

I tried to make sure that the two groups were of roughly comparable ability. Based on the results of the pretest, students were divided into nine pairs with one left over. Paired students' scores differed by a maximum of one point. Within a pair, the students were randomly assigned to one of the two groups. One remaining participant was assigned to the vector group. Thus, the vector group had 10 and the formula group had 9 participants. The two groups had roughly comparable spread of ability.

Materials

The text material was written by the author because no appropriate material was found. Two types of textual material was used corresponding to the two groups. To make these two textual materials have the same difficulty as much as possible, I first wrote the material to be used in the formula group and then "translated" it into the text used in the vector group.

Three problems shown in Table 1 were used in this experiment. All of the statistical concepts that were needed to solve these problems were explained in the textual material for both groups. Because of space limitation, I omit the detailed description of the content of these textual materials.

Procedure

There were two sessions in this experiment: understanding the text material, and problem solving. The experiment was conducted in groups of 3 to 8 participants.

In the first session, participants tried to understand the text material. If they found a description that was difficult to understand, they were required to underline that part in the textbook and to note the reason for the difficulty in the margin. The participants took about one hour to finish this session although there was no time limit.

After reading the text material, each participant reported to the experimenter (i.e. me) their difficulties in understanding the text. The experimenter gave each participant additional explanations about the difficult points in the text. All of the questions were resolved before the participants proceeded to the second session.

In the second session, the three problems shown in Table 1 were presented one by one. Fifteen minutes were allocated to solving each problem. Participants were allowed to look at the text material at any time. All participants were given a paper-and-pencil version of the test.

The participants received the correct answer after they had finished trying to solve each problem. They were required to evaluate to what degree they could understand

each solution and to what degree they could accept it on a 5-point scale ranging from 1: "very difficult to understand (or accept)" to 5: "very easy to understand (or accept)."

Results

Understanding text

It turned out that the two textual materials were similar in the sense that they had almost the same difficulty. Column 4 in Table 2 presents the number of descriptions reported as being difficult to understand in the text for each participant. There was very little difference in the number of reported difficulties during the learning session between the two groups in this experiment. The number was 8 in the formula group and 9 in the vector group.

A closer look at the reported difficulties revealed that vector diagrams often helped the students in understanding several points in which the students in the formula group had a difficulty, while other obstacles arose with vector diagrams. In the formula group, 5 of 9 participants (N, P, Q, R and S) said that it was difficult to understand the proof which showed that the range of correlation coefficient is from -1 to $+1$. Two participants (O and S) found difficulty in understanding the reason why dividing covariance by two standard deviations was the most proper way to capture the relation between two variables. One participant (R) said that she had trouble in understanding where a_0 and a_1 in the formula $\hat{y}_i = a_0 + a_1 x_i$ came from. In the vector group, all these points were not problematic for the students. No students in this group reported any difficulties in understanding the corresponding points in their textual material. Instead, they had trouble in understanding other points. Four of the 10 participants (D, G, H, and J) said that they did not know the concept of orthogonal projection (see Figure 1). Two participants (G and H) said understanding inner product—which was used in this group to define correlation coefficient—was difficult. Two participants (H and I) had difficulty in imagining n-dimensional vectors. One participant said the equation which describes the relation between the variance and a vector was difficult.

Problem solving

Columns 5, 8 and 11 in Table 2 present the performance of each participant and success S (%) in problem solving for each group; F means failure in problem solving. For each problem, the two columns to the right of the column indicating success or fail in problem solving show participants' self-evaluation of the degree of understanding and acceptance of the correct solution presented after their attempt at problem solving.

All in all, the results supported our prediction.

Problem 1 Vector representations facilitated solution of Problem 1. In the vector group, one participant (Participant F in Table 2) reached the conclusion $R = |r|$ and 5 participants found the answer $R = r$ in the case of $r \geq 0$. All of these students used vector diagrams. In the formula group, no participant got the answer $R = |r|$ or $R = r$ to this problem. The difference in success S (%) between the two groups was significant (Fisher's exact test, $p = .011$).

No significant difference was found in the self-evaluation scores for understanding and acceptance of the given correct solution.

Problem 2 and 3 In contrast to the good performance on Problem 1, no participant in the vector group succeeded in solving Problem 2 and Problem 3. The participants in the formula group showed relatively good performance. The difference in success S (%) between the two groups was significant in Fisher's exact test, $p = .003$ for Problem 2 and $p = .011$ for Problem 3.

The scores for understanding and acceptance of the correct solution to Problem 2 in the vector group were lower than the scores in the formula group. The differences in means between the two groups were significant, $t(9.0) = 5.28, p = .001$, for understanding; $t(17) = 3.15, p = .006$, for acceptance.

There was no significant difference in the scores of the two groups for understanding and acceptance in Problem 3.

Discussion

The limited number of diagrammatic operators can make the problem space smaller, and raise the probability of reaching the correct answer. We predicted that this property would improve performance on Problem 1 and the results supported this prediction. Formulae allow students to do many kinds of manipulation. For example, in Problem 1, participant R tried to get $R \times \frac{1}{r}$ and participant M considered $\frac{\{1/n \sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{R}$. Note that it is next to impossible to do these manipulations on a vector diagram. If a diagram rules out these messy manipulations, it must be a big help for students.

Interestingly, the same property, namely, affording a small number of operators, could prevent students from finding the solution and understanding an explanation. This is the case in Problem 2. No participant in the vector group succeeded in solving this problem. The participants also had trouble in understanding and accepting the correct answer to this problem. After the experiment, participant A told me that understanding the decomposition of vectors was difficult, especially, $(\bar{x}, \bar{x})$ and $(\bar{y}, \bar{y})$ looked strange. This feedback suggests that the students were likely to rule out the decomposition operator necessary

Table 2: Summary of the Data from Experiment 2.

				Problem 1			Problem 2			Problem 3		
Groups	Participants	Pre	Difficulties	S/F	Un	Ac	S/F	Un	Ac	S/F	Un	Ac
Vector	A	3	0	F	5	2	F	2	1	F	4	4
	B	3	1	S	4	4	F	2	3	F	2	2
	C	3	0	S	5	5	F	4	3	F	5	5
	D	1	1	F	1	1	F	4	2	F	4	2
	E	1	0	S	5	5	F	2	3	F	5	5
	F	1	0	S	4	4	F	4	4	F	4	4
	G	1	2	F	3	3	F	1	1	F	5	5
	H	1	3	F	4	3	F	1	1	F	1	1
	I	1	1	S	4	4	F	4	4	F	4	4
	J	1	1	S	4	5	F	4	4	F	4	4
	Mean/%correct	1.60	0.90	60.0%	3.90	3.60	0.0%	2.80	2.60	0.0%	3.80	3.60
	SD	0.92	0.94		1.14	1.28		1.25	1.20		1.25	1.36
Formula	K	3	0	F	4	3	F	5	3	F	5	4
	L	3	0	F	2	4	S	5	5	S	5	5
	M	2	0	F	4	3	S	5	5	S	4	3
	N	2	1	F	4	4	S	5	5	F	5	5
	O	1	1	F	5	5	F	5	4	S	5	5
	P	1	1	F	4	4	S	5	5	F	4	2
	Q	1	1	F	5	5	F	5	2	S	3	3
	R	1	2	F	4	4	S	5	5	S	4	4
	S	0	2	F	4	2	S	5	5	F	4	4
	Mean/%correct	1.56	0.89	0.0%	4.00	3.78	66.7%	5.00	4.33	55.6%	4.33	3.89
	SD	0.96	0.74		0.82	0.92		0.00	1.05		0.67	0.99

Notes. Pre: the score of pretest (1–5)
Difficulties: the number of descriptions in the text which were difficult to understand
S/F: success (S) or failure (F) in problem solving
Un: the score of evaluating the degree of understanding the correct solution (1–5)
Ac: the score of evaluating the degree of acceptance of the correct solution (1–5)

to solve this problem. The low ratings for understanding and acceptance of the correct answer refute the possibility that the inability to solve this problem means that the participants carelessly failed to apply a familiar operator.

Similar to the case of properties of diagrams, a property of formulae can be either an advantage or a disadvantage. Abstractness is an example of such properties. This property was predicted to work against solving Problem 1 but to be an aid in solving Problem 3 in the formula group. The results of the experiment were consistent with this prediction. A formula and its transformation become more concrete when they are connected to a diagram. This is the case in heterogeneous inference, inference that use both diagrammatic and algebraic representations. A pure diagrammatic solution is easier if this solution is as simple as the algebraic part of the heterogeneous inference.

Conclusion

Previous research on diagrammatic reasoning has pointed out many "good" properties of diagrams and has claimed advantage for diagrammatic approaches over conventional (usually algebraic) approaches. From the results presented here, it seems that the story is not so simple. The results of this experiment indicate that the vector diagram is not a panacea for students struggling with statistics. The same property of certain diagrams or formulae can be either an advantage or a disadvantage. Teachers should keep this in mind and ponder how properties of diagrams or formulae can work in a particular situation.

Acknowledgement

I wish to thank Sciencedit (http://www.sciencedit.com/) and two of my friends, Ryan Baker and Erik Lindsley, for editing the manuscript of this paper.

References

Barwise, J., & Etchmendy, J. (1996). Visual information and valid reasoning. In G. Allwein & J. Barwise (Eds.), *Logical Reasoning with Diagrams*. New York: Oxford University Press.

Cheng, P. C. -H. (1992). Diagrammatic reasoning in scientific discovery: Modeling Galileo's kinematic diagrams. *Reasoning with Diagrammatic Representations: Papers from the 1992 Spring Symposium*. (pp. 33–38) Menlo Park, CA: AAAI Press

Cheng, P. C. -H., & Simon, H.A. (1995). Scientific Discovery and Creative Reasoning with Diagrams. In S. M. Smith, T. B. Ward & R. A. Finke (Eds.), *The Creative Cognition Approach*. Cambridge, MA: MIT Press.

Larkin, J., & Simon, H.A (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99.

An fMRI study of the Interplay of Symbolic and Visuo-spatial Systems in Mathematical Reasoning

Atsushi Terao (atsushi@edu.hokudai.ac.jp)
Graduate School of Education, Hokkaido University, Kita 11 Nishi 7
Sapporo, 060-0811 JAPAN

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3890 USA

Myeong-Ho Sohn (mhsohn@andrew.cmu.edu) Yulin Qin (yulinQ@andrew.cmu.edu)
John R. Anderson (ja+@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave. Pittsburgh, PA 15213-3890 USA

Cameron S. Carter (CarterCS@msx.upmc.edu)
Imaging Research Center, University of California at Davis
1 Shields Ave. Davis, CA 95616 USA

Abstract

The purpose of this fMRI study was to provide evidence for the mathematician's belief that mathematical thinking emerges from the interplay between symbolic and visuo-spatial systems. Twelve participants were given algebra word problems and depicted the quantitative relations on a mental number line or made parts of an equation. The regions activated in depicting the picture were also recruited to make an equation.

Mathematics is a language. Many scientists say that mathematics is a language to describe the nature of phenomena they are looking at. It is well known that Nicolas Burubaki, a group of mathematicians, stressed the crucial role of formal symbol systems in mathematics.

On the other hand, many mathematicians and physicists emphasize the role of visuo-spatial reasoning in mathematics, which recruits qualitative, language-independent representations. For example, Albert Einstein stated "Words and language, whether written or spoken, do not seem to play any part in my thought process."

As Dehaene, Spelke, Pinel, Stanescu, and Tsivkn (1999) suggested, mathematical thinking may emerge from the interplay between symbolic and visuo-spatial systems. In this fMRI study, we approach this problem and provide evidence for this kind of mathematical thinking.

Psychological studies have revealed that if a problem apparently looks like a pure symbolic task, it can require students to have some visuo-spatial representations. For example, Griffin, Case and Siegler (1994) showed that the mental "number line", a qualitative representation of the number system, is crucial readiness for early arithmetic. Lewis used a number-line-like diagram to train undergraduate students having difficulty to solve "compare" word problems (problems containing more-than or less-than relations), and succeeded in improving their performance. Paige and Simon (1966) proposed that solving word problems is not a simple translation of problem sentences into equations, as Bobrow's (1966) STUDENT did, but needs "physical cues," a visuo-spatial representation. The 6th grade students who used our "Picture Algebra" strategy (Koedinger & Terao, 2002) to solve the compare word problem showed relatively high performance. We expect that using this strategy may better prepare students to learn formal algebra.

Functional magnetic resonance imaging (fMRI) gives us a new source of information about the mental representations used in mathematics. Dehaene et al. (1999) showed two different mental representations are used for different tasks. Exact calculation (e.g., 4+5=9) elicited left-lateralized activation in the left inferior frontal lobe, together with left angular gyrus and left anterior cingulate. This pattern was interpreted as suggesting that the participants recruited their symbolic systems and did language dependent encoding. Approximation (e.g., 4+5 is closer to 8 than 6), on the contrary, elicited bilateral parietal lobes activation. This pattern was interpreted as suggesting that the participants recruited visuo-spatial systems and did language independent encoding. Dehaene, Piazza, Pinel and Cohen (2003) reviewed neuro-imaging and neuropsychological evidence concerning various numerical tasks and proposed a hypothesis that three parietal circuits are related to number processing. The horizontal segment of the intraparietal sulcus (HIPS) appears to be a core quantity system, analogous to a mental number line. This area seems to be supplemented by two other circuits. One is the bilateral posterior superior parietal lobule (PSPL), which is considered to be involved in attention orientation on the mental number line. The other is the left angular gyrus, which is likely to support manipulations of numbers in a symbolic form (e.g., exact calculation).

Dehaene et al. (1999, 2003) suspected that mathematical thinking may emerge from the interplay between symbolic and visuo-spatial systems but did not provide direct evidence for this idea. For example, exact calculation and approximation mainly depend on the symbolic system and the visuo-spatial system, respectively, not necessarily a collaboration between these two systems.

In this study, we try to provide direct evidence for such a collaboration. We decided to use algebra word problems for three reasons. First, previous psychological research suggests that visuo-spatial reasoning plays a crucial role in solving these problems while they explicitly require students to use symbols (i.e., equations). This kind of problem is expected to show the interplay between symbolic and visuo-spatial systems. Second, algebra word problems are widely used in school mathematics curriculum, so that we can say the observed interplay is a prevailing form of reasoning, not a special form isolated to a very specific task. Third, there are plenty of studies using algebra word problem, the accumulated findings help us in valid reasoning from our results.

If algebra word problems recruit the visuo-spatial system as well as the symbolic system, we should see activation of some visuo-spatial areas when students try to make a correct equation for a problem. To find visuo-spatial areas, we asked our participants to make a pictorial representation of the problem in one condition. This task should activate visuo-spatial areas and most of these areas should also be activated when we ask the participants to make an equation of the same problem in another condition. We especially expect that the two hypothesized parietal visuo-spatial systems, HIPS and PSPL, show activation in both conditions.

Method

Participants

Participants were 12 right-handed, native English speakers. They were recruited by advertisement posted on an electric bulletin board in Carnegie Mellon University. Participants were provided written informed consent in accordance with the Institutional Review Boards at the University of Pittsburgh and at Carnegie Mellon University.

Tasks and Design

There was one *representation* factor and four *problem* factors. They were all manipulated within subjects. The representation factor was the mental representation the participants made from the problems. In the *picture* condition, the participants draw a mental image describing the critical relations in the problem; in the *equation* condition, the participants were required to construct an equation to the problem.

Table 1 shows two example problems. Each problem consisted of three problem sentences and used three letters as unknown quantities. The first sentence was an assignment sentence. In the equation condition, the participants memorize what letter the "x" is assigned to. In the picture condition, the participants imagined a number line and picked locations for the letters. The second and the third sentences described a qualitative relation between two letters. Hereafter we call them R1 (meaning "Relation 1") and R2, respectively. Participants could not find a numerical solution to these problems because they did not have a sentence referring to the total amount of the three unknown quantities. For example, if the first problem in Table 1 has the sentence "The total of the three quantities is 44," we can make an equation to find the three quantities like x+(x+6)+(x+8)=44. In the equation condition, the participants were told that the total sentence would be omitted and were required to make the left side of the correct equation like x+(x+6)+(x+8). For the first problem in Table 1, the participant can make the parts (x+6) and (x+8) from R1 and R2, respectively. In the picture condition, the participants imagine a picture describing the critical relations in the problem as shown in Figure 1. For the first problem in Table 1, the participants can imagine B to the right of the location A on the mental number line and add the distance, 6, to this picture. When they read and represent R2, the whole picture can be obtained.

Table 1: Sample of Problems.

Consistent/more-than/intransitive problem	
Assignment	x=A
R1	B is 6 more than A.
R2	C is 8 more than A.
Inconsistent/less-than/transitive_n problem	
Assignment	x=A
R1	A is 6 less than B.
R2	B is 2 less than C.

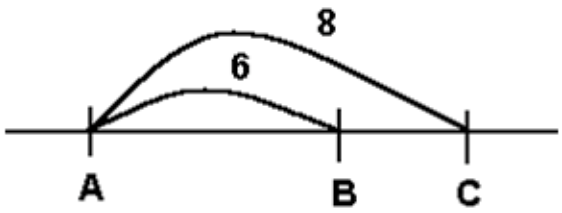


Figure 1: An example of pictures made from problems.

A first problem factor was the relation, whether R1 and R2 use *more-than* or *less-than* relations.

A second problem factor was the consistency, whether the relations used in R1 and R2 were consistent with the correct equation. A more-than problem was labeled either as a *consistent* problem if the correct equation uses "+"or labeled as an *inconsistent* problem if the correct equation uses "-". We use a similar labeling for the less-than problems.

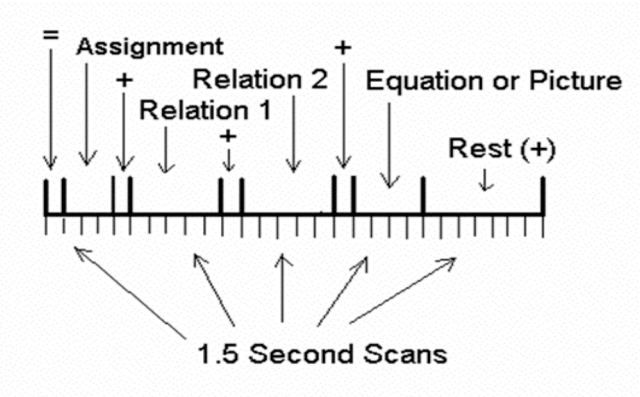


Figure 2: The 42-second structure of an fMRI trial.

A third condition was defined in accordance with the target stimuli presented at the end of each problem (trial). The problem was labeled either as a *correct* problem if the target is the correct target or as an *incorrect* problem if the target is incorrect.

A forth factor was transitivity. This factor was defined by the two relational sentences, R1 and R2. Considering the picture the participants were required to make seems to be an easy way to explain this factor. For the *intransitive* problem, an arc will be drawn over another arc as shown in Figure 1. This will be the case if either the former letter or the latter letter is common in R1 and R2. For example, R1 and R2 of the first problem in Table 1 use the same latter letter (A), and this is an intransitive problem. For the *transitive_n* problem (n stands for Normal), one arc should be drawn at the next position to another arc in the correct picture. A third level of this factor is represented by another problem. This type of problem, called the *transitive_d* problem hereafter (d stands for Delay), was made by changing the R1 and R2 of the transitive_n problem. In the transitive_d problem, participants in the equation condition are not able to make a part of the equation until they read R2. They need to remember R1 and make the two parts of the equation when R2 is presented. The purpose of making this unusual problem was to find brain regions which play a role in making equations going beyond simple encoding of problem sentences. Comparing the transitive_n problem with the transitive_d problem in the time period of presenting R1 might reveal differences between memory and processing areas but we will not say much about this comparison in this paper. There may be no difference between the transitive_n and transitive_d problems in the picture condition because the participants can describe the relation when they see R1. For example, for the transitive_d problem made from the second problem in Table 1 (R1: "B is 2 less than C."), the participants can imagine the spatial relation between B and C by just ignoring A used in the assignment sentence.

Combinations of the four problem factors (2x2x2x3) yielded 24 types of problems. The participants went through all of these 24 types in both the picture condition and the equation condition in the MRI scanner, so that each participant encountered 48 problems. The 48 problems were divided into four blocks: two blocks in the picture condition and the other two blocks in the equation condition.

Procedure

Pre-scan Practice Participants took about 20 minutes of pre-scan practice just before the scan. They went through one block of 12 trials (problems) in the picture condition and another block of 12 trials in the equation condition. Half of the participants started with the picture condition and the other half of the participants started with the equation condition. The time course in each trial was the same as the one in the scanner.

Event-related fMRI scan Event-related fMRI data were collected by using a singe-shot EPI acquisition on a Siemens 3T scanner, 1500 TR, 30 ms TE, 60^0 flip angle, 210 mm FOV, 26 axial slices/scan with 3.2 mm thick, 64x64 matrix, and with AC-PC on the 6^{th} slice from the bottom. There were 28 scans (42 seconds) for each trial, 12 trials for a block and 4 blocks for each participant. Two of these 4 blocks were for the picture condition and the other two blocks were for the equation condition. Half of the participants started with the first block in the picture condition and proceeded to the second block in the equation condition, the third block in the picture condition, and the last block in the equation condition. The other half of the participants started with the first block in the equation condition, then went through picture, equation, and picture conditions in this order.

The protocol of each trial of scan is illustrated in Figure 2. The three problem sentences appeared on the screen one by one. The assignment sentence was on the screen for 3500 ms; R1 and R2 were on the screen for 7500 ms. A target equation or picture was presented after the disappearance of R2. Participants responded to this target by pressing the button. If they thought the target was correct they pressed a button with the index finger of the right hand; if they thought the target is incorrect, they pressed the other button with the middle finger of the right hand.

fMRI data analysis Data processing was conducted using SPM99 software (http://www.fil.ion.ucl.ac.uk/spm/). Slice timing was corrected first and images were realigned. Realigned images were normalized to Talairac coordinates. Normalized images were smoothed with an 8 mm FWHM isotropic Gaussian Kernel. Analysis was carried out using the general linear model with a box-car waveform convolved with a hemodynamic response function. Only correct trials were used for analysis.

To find brain regions of interest (ROI), a random effects model was used. At the first level, mean images for each participant were created, depicting the subtraction of BOLD response during assignment sentence from BOLD response during R1 in each condition (picture and equation). Data from transitive_d problems were excluded to do this subtraction because of the unique nature of these problems. At the second level, these mean images were combined in one-sample *t*-test. We used a height threshold of $P < 0.0005$

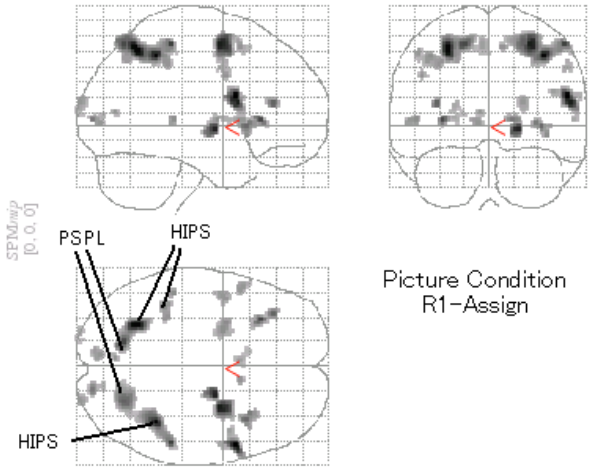


Figure 3: Regions of the brain that show activation in depicting a relation between two quantities in the picture condition as compared to the encoding of the assignment.

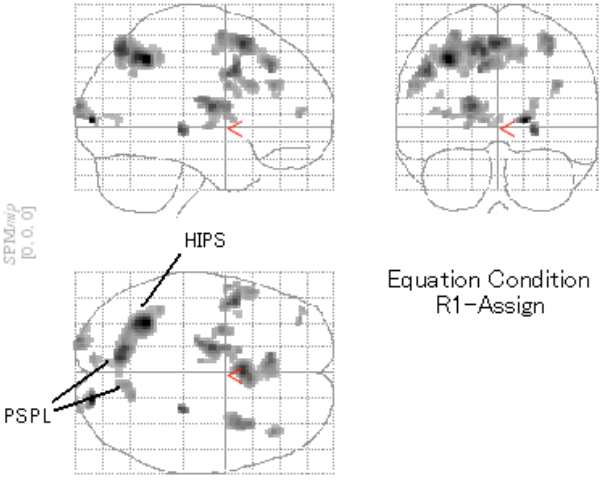


Figure 4: Regions of the brain that show activation in constructing parts of an equation in the equation condition as compared to the encoding of the assignment.

uncorrected (t > 4.44), with an extent threshold of eight contiguous voxels in an ROI. This analysis should show us the regions activated when constructing a number line to describe a quantitative relation or constructing parts of an equation from this relation. The subtraction between picture and equation conditions might be interesting but we did not conduct it, because this subtraction might hide visuo-spatial areas that are recruited not only in picture condition but also in equation condition.

Results

Figure 3 shows the brain regions which show activation during the period of R1 as compared to the assignment sentence in picture condition. We can infer that these regions may be related to using a mental number line to depict a relation between two quantities.

Figure 4 shows the brain regions in the equation condition obtained by the same subtraction (R1 - assignment). These regions should be related to constructing parts of an equation from sentences.

Comparing these two figures, we can see an overlap of activation especially in the parietal lobe (HIPS and PSPL). This means that constructing an equation, which apparently is a symbolic task, recruits the visuo-spatial system.

Areas for constructing a mental number line

A pattern of bilateral activation was obtained for drawing a mental number line to represent a quantitative relation. As we had expected, the active areas in parietal lobes occupied HIPS and PSPL (Tarairach coordinates of main peaks: -40, -40, 48, Z=3.75; -28, -60, 54, Z=4.63; 38, -46, 48, Z=4.62; 24, -68, 56, Z=4.12). Activation was also found during constructing a number line from R1 in the bilateral premotor cortices (-30, 2, 52, Z=3.79; -46, 2, 34, Z=3.72; -48, 0, 50, Z=3.68; 28, -2, 56, Z=4.52; 54, -2, 40, Z=3.51), bilateral supplementary motor areas (-4, 12, 56, Z=3.59; 10, 12, 54, Z=3.73), left Broca area (-52, 12, 4, Z=3.69), right inferior frontal sulcus (54, 8, 20, Z=4.75), left and right insula (-32, 26, 4, Z=4.27; 36, 16, 0, Z=4.05), left and right corpus striatum (20, -6, -2, Z=4.49; -24, -2, 6, Z=3.51), and right DLPFC, dorsolateral prefrontal cortex (-36, 34, 16, Z=4.07). Activation found in left and right visual cortex should reflect longer exposure to the visual stimulus: The period of R1 was longer than the period of assignment sentence.

Areas for constructing an equation

A pattern of left-lateralized activation was obtained for constructing parts of an equation from a sentence referring to a relation between two quantities. The bilateral PSPL and left HIPS activation were also found as in the picture condition (-34, -54, 46, Z=4.93; 16, -62, 50, Z=3.85). This means that the hypothesized parietal visuo-spatial system (Dehaene et al., 2003) was recruited when constructing parts of an equation from a problem sentence. Activation was also found during construction of parts of the equation from R1 in the left premotor cortex (-58, 2, 28, Z=3.87), bilateral supplementary motor areas and right Brodmann area 8 (-28, -2, 62, Z=4.00; -14, 8, 58, Z=3.79; 0, 12, 56, Z=4.53; 34, 8, 58, Z=4.06), bilateral inferior frontal sulci (-38, 8, 22, Z=3.44; -48, 6, 38, Z=4.43), left basal ganglia including thalamus and globus pallidus (-16 -10 14, Z=4.43), right parahippocampal gyrus (24, -30, -2, Z=4.38), and left and right DLPFC (40, 32 28, Z=3.95; -36, 50, 8, Z=3.77).

To confirm that several brain areas activated in picture condition also showed activation in equation condition, we plotted percent signal change along the time course. The base line for calculating the signal change was set by the average of first two scans. Remember that these areas were found in picture condition and no data from equation condition were used. Data from intransitive problems and transitive_n problems were combined in each condition.

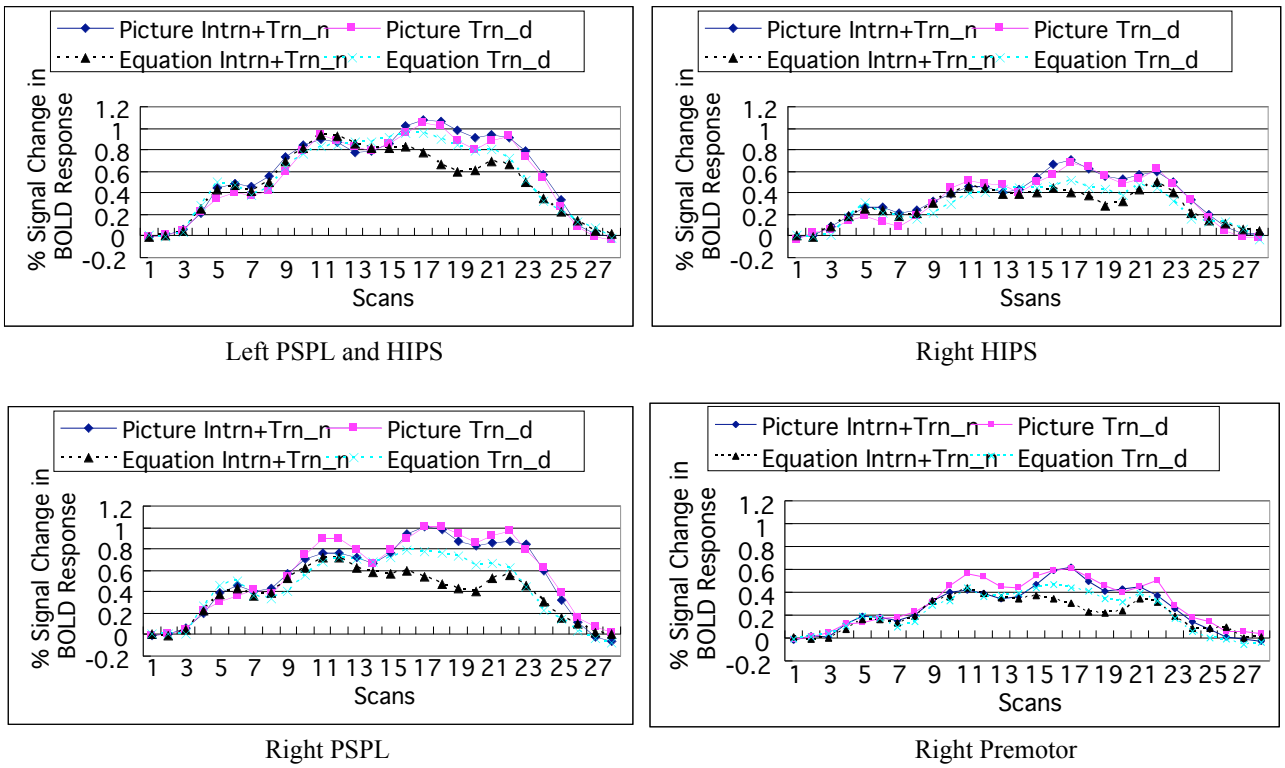


Figure 5: Percent signal change in BOLD response in four ROIs found in picture condition.

After finding ROIs, the signal change was calculated for each of the four problem types (picture--transitive_d, picture--others, equation--transitive_d, and equation--others) and for each participant, and then averaging across the 12 participants. Among many ROIs found in picture condition, because of the limited space of this paper, we only show the percent signal change at left PSPL and HIPS (-28, -60, 54; The number of voxels is 333), right HIPS (38, -46, 48; The number of voxels is 303), right PSPL (24, -68 56; The number of voxels is 241), and right premotor cortex (28, -2, 56; The number of voxels is 220.). These are the four biggest clusters of active voxels.

Figure 5 shows the percent signal change in each of these four ROIs. We can see that these areas found in picture condition also played a role, more or less, in equation condition. Other areas found in picture condition also showed a similar pattern of activation.

Discussion

The results of this experiment indicate that mathematical thinking emerges from the interplay between symbolic and visuo-spatial systems. The hypothesized two parietal visuo-spatial regions (Dehaene et al., 2003) showed activation not only when participants imagined a picture from a problem sentence but also when participants constructed parts of an equation from the sentence. We cannot deny a possibility that these parietal regions are involved in non-visuospatial mathematical reasoning. But it is reasonable we consider them as picture regions until some evidence is fond for this possibility in brain imaging research.

We might expect language areas to show greater activation on these symbolic tasks than in the more visuo-spatial picture task. However, while we found activation in language areas, particularly Broca's and the inferior frontal sulcus, we did not find clear differences in those areas between the two conditions. Perhaps the language areas are doing different kinds of computations in the two conditions, but we found no evidence to support this claim.

Our results seem to be consistent with the recent version of the ACT-R theory (Anderson et al., submitted). The ACT-R theory hypothesizes several buffers and their locations in the brain. For example, the goal buffer is supposed to be in DLPFC and the imaginal buffer in the parietal lobe. The theory also hypothesizes that production rules are stored in the corpus striatum. We found activation in these regions in this experiment. The recent version of ACT-R theory can predict the percent signal change in BOLD response based on the task analysis. We did a task analysis before conducting this experiment. Constructing an ACT-R model could provide us an explanation about the signal change in this experiment.

Readers might suspect that the parietal activation in PSPL and HIPS only reflects longer exposure to the relational sentence R1 (7500 ms) than the assignment sentence (4500 ms). This effect might be the case but it cannot explain the different pattern of activation between the picture and

equation conditions (see Figures 3 and 4) because it should have the same effect in both conditions.

Readers might also suspect that activation of visuo-spatial areas in the equation condition was an effect of use of the mental number line in the picture blocks spilling over to the equation blocks. Because we used a within-subject design, we cannot deny this possibility. Even if it is true, however, it is the effect that we expect in educational settings. Those participants who used visuo-spatial systems on equations may have enhanced their performance. Using a framework of production systems, we can write production rules that represent a purely symbolic processing of a problem sentence. For example, we can think of the following production rule to process R1:

IF x is bind to $A
  and R1 says "$B is $N more than $A
THEN represent $B as x+$N,

where the letter with $ means a variable. Visuo-spatial systems are not necessary if using this kind of production rule but it appears participants still used visuo-spatial reasoning to construct an equation. The following alternative set of production rules illustrates how use of the visuo-spatial systems may make this task easy:

IFR1 says "$B is $N more than $A
THEN
  $B is to the right of $A on the number line

IF x is bind to $A
  and $B is to the right of $A on the number line
THEN represent $B using a plus sign, as x+$N,

If the first production is already exists prior to algebra instruction, the second rule is easier to learn, perhaps, then the one above.

It has been shown that using a pictorial representation helps students solve algebra word problems (e.g., Koedinger & Terao, 2002; Lewis, 1989). These results can be interpreted that the students who learned to use visuo-spatial systems improved their ability to solve algebra word problems. Even if the results of this study only show that students can use visuo-spatial systems to solve algebra word problems only after trained with a pictorial representation, this encourages the educational practice of using a pictorial representation as a scaffold of learning.

There is an interesting episode in the pre-scan, practice session. A few participants showed bad performance in the equation condition. We asked them what they did to make equations. They revealed that they used a "direct translation" strategy which was similar to the strategy Bobrow's (1966) STUDENT used. For example, if R1 says "B is 6 more than A," they translated this sentence into the form of "B=6+A" before substituting "x" for a quantity. This strategy does not seem to need any visuo-spatial reasoning. The fact that the poor performer first used this kind of strategy suggests that students having difficulty with word problems may not learn to make use of visuo-spatial systems.

This study is still in progress and we only have scratched the surface in dada analysis. Further data analysis (e.g., statistical comparisons of the two conditions) should be done and it will provide insight into mathematical thinking.

Conclusion

Mathematical thinking emerges from the interplay between symbolic and visuo-spatial systems. Algebra word problems, which are widely used in current school curriculum, are not a pure language processing task. They appear to depend on the use of visuo-spatial systems.

Conclusion

Mathematical thinking emerges from the interplay between symbolic and visuo-spatial systems. Algebra word problems, which are widely used in current school curriculum, are not a pure language processing task. They appear to depend on the use of visuo-spatial systems.

Acknowledgments

This work was supported by NSF ROLE grant.

References

Anderson, J. R., Bothell, D., Byrne, M. D., & Lebiere, C. (submitted). An integrated theory of mind. http://act-r.psy.cmu.edu/publications/

Bobrow, D. G. (1966). Natural language input for a computer problem solving system. *MAC-TR-1, Project Mac.* MIT

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*, 487-506.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and Brain-imaging evidence. *Science*, *284*, 970-974.

Koedinger, K. R., & Terao, A. (2002). A cognitive task analysis of using pictures to support pre-algebra reasoning. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 542-547). Mahwah, NJ: Lawrence Erlbaum Associates.

Lewis, A. B. (1989). Training students to represent arithmetic word problems. *Journal of Educational Psychology*, *81*, 521-531.

Paige, J. M., & Simon, H. A. (1966). Cognitive processes in solving algebra word problem. In B. Kleinmuntz (Ed.), *Problem solving: Research, method and theory*. New York: John Wiley & Sons.

Griffin, S. A., Case, R., R Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for social failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice.* Cambridge, MA: MIT Press.

Shared Knowledge in Collaborative Problem Solving: Acquisition and Effects

Susanne Thalemann (thaleman@cognition.uni-freiburg.de)
Gerhard Strube (strube@cognition.uni-freiburg.de)
Institut für Informatik und Gesellschaft, Friedrichstr. 50
79098 Freiburg, Germany

Abstract

Whenever heterogeneous experts work together, shared knowledge comes into play. In recent years, two important research questions have emerged that we will address in the present paper. First, we analyzed if collaborative problem solving leads to the construction of shared knowledge. The second goal of this study was to demonstrate the assumed positive effect of shared knowledge on collaborative problem solving. Following Newell and Simon's (1972) classic view of problem solving, we distinguish shared knowledge about the initial situation, the goals, and the operators. The kind and amount of prior shared knowledge was varied as an independent variable in four experimental conditions. We showed that participants acquired shared knowledge during the cooperation and that most of this information was correct. Further results indicate that if collaborating partners have knowledge in common, their overall problem solutions are better than if they lack any kind of shared information. However, this effect seems to be mostly due to shared knowledge about initial situation and goals, as this leads to better solutions than shared knowledge about operators.

Introduction

Complex tasks often require the collaboration of experts with heterogeneous background knowledge. If we refer to this kind of collaborative problem solving, the construct of shared knowledge plays a very important role. Two lines of research dealing with the role of shared knowledge can be distinguished. On the one hand, there are studies that address the *acquisition* of shared knowledge in collaborative activities, especially in collaborative learning. On the other hand, many researchers focus on the *effects* of shared knowledge as a prerequisite for effective collaborative problem solving. In the present paper, we will address both issues in a net-based scenario.

Concerning the *acquisition*, it is mostly assumed that collaborative learning leads to the construction of shared knowledge (e.g. Pfister, Wessner, Holmer & Steinmetz, 1999; Roschelle & Teasley, 1995) through interaction and communication (Clark & Brennan, 1991). Although there is evidence for this assumption, this issue is seldom addressed quantitatively. In addition, the possibility of the emergence of false shared knowledge is often neglected. Only a few studies try to quantify the amount of shared knowledge (e.g. Fischer & Mandl, 2000; Jeong & Chi, 2000) and to analyze the way of its emergence (either through individual but similar experience with the learning material or by co-construction). There are several possibilities of co-constructing shared knowledge. One is that two interacting partners mutually build a piece of shared knowledge that contains novel information, which none of them possessed in full before (e.g. Roschelle & Teasly, 1995). Another possibility, analyzed here, is that one person communicates some of his specific information to his partner, so that the latter learns and understands it, which constitutes shared knowledge between the two. One goal of this article is to analyze and quantify if collaborative problem solving leads to the acquisition of shared knowledge and to determine if it is correct or not. As collaborative problem solving also requires participants to communicate and interact in order to find a solution and, furthermore, in this experiment, to represent it in a common whiteboard, we hypothesize that participants will acquire some sort of shared knowledge.

Concerning its' *effects*, shared knowledge is supposed to be an important variable determining the functioning of groups consisting of members with heterogeneous background knowledge (e.g., expert groups; Hinsz, Tindale & Vollrath, 1997, Smith, 1994). Shared knowledge about the distribution of information within the group is regarded as a major constituent of the group's transactive memory system, which guides information encoding and retrieval on the group level (Wegner, 1987). This shared meta-knowledge augments the group's memory capacity and enables effective retrieval of information held by the group members when needed for problem solving (Moreland, Argote & Krishnan, 1996). Shared mental models comprising shared knowledge about task, team, equipment and situation have been found crucial for effective expert team decision making, problem solving and co-ordination in complex dynamic environments (Cannon-Bowers, Salas & Converse, 1993; Rouse, Cannon-Bowers & Salas, 1992). Furthermore, effective communication requires common ground and shared vocabulary (Clark & Brennan, 1991; Waern, 1992). As shown by this brief summary, the importance of shared knowledge for collaborative activities is widely acknowledged throughout the literature, although the term 'shared knowledge' refers to rather different things. Empirical studies on problem solving that demonstrate the effectiveness of shared knowledge are still rare. In the present paper, we therefore want to develop a taxonomy of shared knowledge that is applicable to problem solving activities in general. Furthermore, we will analyze it's facilitating effects in collaborative problem solving and find out whether some components of shared knowledge are more effective than others.

A Taxonomy of Shared Knowledge for Collaborative Problem Solving

In line with Stasser and Titus (1985, 1987) we define *shared* knowledge as the kind of information two or more persons have in common, i.e., possess and understand in essentially the same way. Information available to only one person is called *individual* or *distributed* knowledge. Based on this definition our taxonomy of shared knowledge follows Newell and Simon's (1972) classic view of problem solving. According to Newell and Simon, knowledge about the initial situation, goals and operators is needed to solve a problem. Expanding this view to collaborative problem solving, shared knowledge can be conceptualized as *shared knowledge about the initial situation, the goals and the operators*. This taxonomy provides a more detailed description of shared knowledge and allows for the testing of differential effects of its components. Moreover, it is applicable to any problem without undue commitments concerning specific initial situations, goals and operators.

Analyzing the Effectiveness of Shared Knowledge

Design tasks are a widely used but scarcely investigated example of problem solving (Smith & Browne, 1993; Strube, Garg & Wittstruck, 2002). For our study, we chose the domain of web design with its heavy use of net-based communication between multiple experts as our experimental task: Dyads of subjects adopting the role of an information technology advisor (IT expert) and a representative of a fictitious company (company expert) have to design an online-shop that meets the company's needs. From the perspective of problem solving, the company expert's knowledge is characterized by information about the company's initial situation and goals. The expertise of the IT expert comprises knowledge about operators, i.e., rules to transform the company's goals and constraints into technical solutions.

As hypothesized above, shared knowledge is needed and facilitates collaborative problem solving whenever heterogeneous experts work together. Following this hypothesis and taking into account that individual problem solving requires individual knowledge about initial situation, goals and operators (Newell & Simon, 1972), we assume that collaborative problem solving requires shared knowledge about the initial situation, goals and operators. Therefore, we state the following hypotheses:

1. Shared knowledge, in general, leads to better problem solving performance than a total lack of shared knowledge.
2. Sharing knowledge about all components, the initial situation, goals and operators, is more effective than just sharing knowledge about the initial situation and goals.
3. Shared knowledge about the initial situation, goals and operators is more effective than sharing knowledge about operators alone.
4. Finally, we are interested in the particular difference of shared knowledge about the initial situation and goals as compared to shared knowledge about operators. Information about the initial situation and goals frame the problem and help to decide when a solution is reached. In contrast operators are means to transform one state into another, so we expect that sharing knowledge about the initial situation and goals has different effects than sharing knowledge about operators.

Method

Tasks and Materials

Design Task. Information needed for the design task was provided in two introductory texts, one for the role of the company expert and one for the role of the IT advisor.

After having acquired their task-relevant information (see next section), two participants worked on the design problem, one as the company expert and the other as the IT advisor. They were instructed to discuss solution features of the online-shop in preparation, via chat. Features which were chosen as part of the final solution were to be noted on a common whiteboard.

The introductory text for the company expert comprised 36 knowledge elements, 27 about the company's initial situation (IS) and goals (G), and 9 containing either shared knowledge or irrelevant filler information (depending on the experimental condition). Knowledge about IS and G was considered as a whole because a) this resembles the natural expertise of a company expert, e.g. a manager and b) knowledge elements about IS and G were of the same form, with the only difference that the goals were formulated as a request, e.g. "The company wants their customers to find the desired products quickly and easily". The introductory text for the IT advisor consisted of 27 knowledge elements about operators (O) and 9 shared elements or fillers. The operators had the form of if-then clauses with the 'then' part specifying a technical feature of the solution, and the 'if' part describing constraints to be fulfilled by the company. If the company's initial situation and goals match the 'if' part, the solution feature provided in the 'then' part should be marked as a desirable feature (DF) of the online-shop. If not, the solution feature should be marked as not being part of the online-shop's functionality (No DF). By integrating the 27 knowledge elements about initial situation and goals with the 27 operators, a total of 24 solution features (15 to be part of the shop and 9 to be rejected) were unambiguously determined. The following example should illustrate the task and a filler item (FI) for the company expert:

IS: Products cost between 7 and 180 Euro.

O1: If medium payments (5-1000 Euro) are expected, then the online-shop should provide the opportunity to pay by check or credit card.

O2: If micro payments (0.1-5 Euro) are expected, the online-shop should use systems like e-Cash.

DF: The online-shop should provide the possibility to pay via check and credit card.
No DF: E-Cash should be used for payments.
FI: The well known designer Philippe Starck designed the layout of the stores.

As participants were not true experts in the domain assigned to them, materials had to be simplified as compared to reality. Despite this simplification, the knowledge elements and solution features provided should be correct. Therefore, three experts on matters of the internet rated the correctness of the knowledge elements and the resulting solutions. As a result, several items were changed or eliminated.

Memory Test. To make sure that information provided in the introductory texts was acquired sufficiently, participants performed a memory test, similar to a recognition test. (The difference was that test items were not identical verbatim, but instead in meaning, to items presented during the learning phase.) The following example shows target (T) and distractor (D) for a knowledge element concerning the company's initial situation (IS). Items were presented in random order and should be classified as correct or false.

IS: Online sales should be integrated into the current business practices of the company, as it would be too costly to add additional personnel to process the online shop's orders by hand.
T: Due to personnel and administrative considerations online sales should be integrated into the business practices.
D: Due to personnel and administrative considerations online sales should be separated from the current business practices.

If the criterion of 95 percent correct responses was not reached, text and the subsequent test were presented again. This procedure was repeated until both subjects reached the criterion, or a time limit of 80 minutes was exceeded.

Further Tasks. In addition to the design task, participants had to complete two additional tasks. Prior to the experiment, they answered a questionnaire about relevant aspects of the IT advisor's knowledge. In order to make the experimental manipulation effective, only naive participants were accepted. As a fictitious company was used, there was no need to control for prior knowledge concerning the company. The final task was an unexpected repetition of the memory test already applied after reading the introductory texts. As opposed to the first time, the test now comprised *all* information relevant to the task for both subjects so that the amount of acquired shared knowledge can be determined.

Experimental Design

The kind and amount of prior shared knowledge (SK) was manipulated as an independent variable in four experimental conditions. In the first condition, no shared knowledge is available prior to collaboration (No SK). The company expert's introductory text only contains the 27 knowledge elements about the initial situation and goals (IS+G), as well as 9 fillers. Accordingly, the IT-expert is provided with 27 operators (O) and 9 fillers. In the second condition, SK (IS+G), knowledge about IS+G is shared. Prior knowledge of the company expert remains unchanged, whereas the 9 fillers of the advisor's prior knowledge are replaced by 9 knowledge elements about initial situation and goals. These 9 elements form the dyad's shared knowledge. Similarly, condition 3 with shared operators is realized, SK (O): Fillers in the company expert's introductory text are replaced by 9 operators, whereas the advisor's prior knowledge stays the same as in the first condition. Condition 4, SK (IS+G+O), provides shared knowledge about both components, starting situation and goals, as well as operators. Like in conditions SK (IS+G) and SK (O), the company expert receives 9 knowledge elements about operators, and the advisor 9 elements about initial situation and goals, in addition to the 27 knowledge elements of their own expertise.

By sharing either 9 IS+G elements or 9 operators in conditions SK (IS+G) and SK (O), 9 solution features can be determined by the advisor (condition IS+G) or by the company expert (condition O) alone, since the shared knowledge elements provide all the information they need in addition to their own expertise. Although each participant is provided with some shared information in condition SK (IS+G+O), the number of solution features that do not require any collaboration remains 9 because the shared knowledge is distributed redundantly. Both partners are able to find the same 9 solution features on their own in condition SK (IS+G+O). In sum, the maximum possible correct solution features is 24 in all conditions.

The variation of shared knowledge as described above focuses on task-relevant shared information such as operators or goals. However, as many of the knowledge elements contained technical terms, providing some of the same knowledge elements to both participants might also have encouraged participants to develop a shared vocabulary around the terms both were exposed to. Furthermore, shared meta-knowledge is another aspect of what is meant by shared knowledge in the present experiment: In conditions SK (IS+G), SK (O) and SK (IS+G+O) participants were told that there is some information they both have in common.

Dependent Variables

Acquisition of Shared Knowledge. The amount of shared knowledge that was acquired during the collaboration was measured by the relative number of items that were recognized in the same way by both cooperation partners in the final recognition test. Initially shared items in the conditions with shared knowledge were excluded as well as all items that had not been discussed during the cooperation

task[1]. This measure also allowed to differentiate between correct (items that were recognized correctly by both partners like a 'hit') and false shared knowledge (items recognized in the same way by both partners, but falsely like a 'miss').

Effects of Shared Knowledge. As an indicator of the effectiveness of shared knowledge, we chose solution quality as our dependent variable. It was measured by the amount of correctly noted solution features, ranging from 0 to 24. All features in the whiteboard were analyzed, only features mentioned more then once were excluded. Solution features were scored as correct if they corresponded to the ideal solution, as determined through the integration of the company expert's and the IT advisor's knowledge. Conversely, solution features contradicting the ideal solution were scored as false. Features that were neither correct nor false, were coded as irrelevant, e.g., intrusions or underspecified features.

In order to detect eventual differences in the difficulty of learning the knowledge supplied about IS+G and O, we measured the time required to read the introductory texts and to acquire the information, until the learning criterion was reached.

Subjects and Procedure

64 participants (47 females, 17 males) were randomly assigned to the 4 experimental conditions, as well as to the role of company expert or IT advisor. None of them had task-relevant prior knowledge as controlled by the first questionnaire. After having acquired the information presented in the introductory text according to the learning criterion, participants performed the design task for 50 minutes. This limit seemed to be appropriate as subjects finished their task quite easily within this timeframe. Afterwards, they completed the net-based questionnaire. The only task not expected by the participants was the final memory test. The experiment took between 2.5 and 3 hours all together.

Results

Acquisition of Shared Knowledge

As expected the hypothesis stating that collaborative problem solving leads to the acquisition of shared knowledge could be accepted for correct ($t_{(31)} = 23.25$, $p < .05$) as well as for false shared knowledge elements ($t_{(31)} = 4.71$, $p < .05$).

As table 1 shows, about one third of the discussed knowledge elements enter the pool of subject's shared knowledge. The amount of shared knowledge did not differ significantly between experimental conditions ($F_{(3,56)} < 1$), but participants acquired significantly more correct than false shared items ($F_{(1,56)} = 447.86$, $p < .05$).

[1] Note that information in the partner's domain can only be acquired by communication within the dyad.

Table 1: Mean relative number and standard deviations of correct and false shared knowledge (SK).

Condition	Correct SK	False SK
No SK	.37	.01
	(.07)	(.02)
SK (IS+G)	.36	.02
	(.06)	(.06)
SK (O)	.38	.09
	(.12)	(1.0)
SK (IS+G+O)	.33	.03
	(.09)	(.04)

Effects of Shared Knowledge

As the 24 solution features were unambiguously determined by combining the company expert's and the IT advisor's knowledge, solution features noted on the whiteboards had to be compared to this ideal solution. Because the features that subjects noted were often incorrectly formulated and thus ambiguous, criteria were defined to determine what variations could be accepted as correct. For example, if a solution feature was called 'eMoney' instead of 'eCash' this was counted in the same way. To control for objectivity, a second rater, blind to the hypotheses of the experiment, assessed all solution features again. 94 percent of all features were rated identically.

Table 2: Mean number and standard deviations of correct, false and irrelevant features.

Condition	Correct features	False features	Irrelevant features
No SK	9.00	1.50	4.50
	(3.46)	(1.41)	(3.46)
SK (IS+G)	14.50	.88	6.38
	(3.16)	(.35)	(3.85)
SK (O)	11.25	2.00	4.88
	(2.71)	(1.20)	(2.85)
SK (IS+G+O)	14.50	1.88	3.75
	(2.39)	(1.25)	(3.24)

ANOVA was carried out to compare the quality of the solution in the four experimental conditions. Table 2 shows the means in the 4 experimental conditions (main effect of shared knowledge: $F_{(3,28)} = 6.59$, $p < .05$). Linear contrasts were carried out to assess overall effectivity of shared knowledge (hypothesis 1), as well as the difference between shared knowledge about initial situation and goals, compared to shared knowledge about operators (hypotheses 4). Hypotheses 2 and 3 were tested by post hoc tests (Scheffé).

The first hypothesis was tested via a linear contrast of condition (No SK), in contrast to the average of conditions SK (IS+G), SK (O) and SK (IS+G+O). As expected, solution quality was poorer when shared knowledge was missing, compared to conditions SK (IS+G), SK (O) and SK (IS+G+O), which had different components of shared

knowledge ($t_{(28)} = 3.65$, $p < .05$). Concerning the relative effectiveness of shared knowledge about IS+G (hypotheses 4), compared to shared knowledge about O, the former lead to significantly better solutions than the latter ($t_{(28)} = 2.20$, $p < .05$). Hypothesis 2, stating better results with shared knowledge about IS+G+O, compared to shared knowledge about IS+G alone, was refuted. The same is true for hypothesis 3, stating better results with shared knowledge about IS+G+O, compared to shared knowledge about O alone, although means pointed towards the expected direction. All these analyses were computed for the correct answers.

Only a few false solution features were noted on the whiteboards, without significant differences between experimental conditions ($F_{(3,28)} = 1.60$, $p > .05$). The number of irrelevant features was also unaffected by the amount of prior shared knowledge ($F_{(3,28)} < 1$).

Subjects needed significantly more time to read and learn information about operators ($M = 2420$ m/sec, $SD = 618$), than those about the initial situation and goals ($M = 2025$ m/sec, $SD = 632$; $F_{(1,56)} = 6.61$, $p < .05$). There were no differences between the experimental conditions ($F_{(3,56)} < 1$).

Discussion

Concerning knowledge acquisition we assumed that the interaction during the design task would lead to the construction of shared knowledge. As results showed, participants in fact acquired shared knowledge even if the amount was smaller than it could have been. This result can be conceived as a first indicator for the importance of shared knowledge as it was acquired even though participants were not instructed to do so and testing was unexpected. Furthermore, most of this shared knowledge elements were correct, so that they can really help participants to construct a common representation of the task and the solution. In contrast to other studies (Fischer & Mandl, 2000; Jeong & Chi, 2000) we did not have problems in determining how this shared knowledge came about, as participants had no common learning or working material, so that the only way this could have happened is through co-construction. We also could be sure, that this shared knowledge did not exist prior to collaboration as the company expert's information was fictitious and knowledge in the domain of online-shops was controlled for.

Concerning the effectiveness, shared knowledge prior to collaboration leads to better solutions in net-based collaboration than working together without shared knowledge. Overall, solution quality was higher if there was shared knowledge (as in conditions IS+G, O and IS+G+O) than if shared knowledge was lacking, as in condition No SK. Although this general benefit of shared knowledge could be demonstrated, certain kinds of shared knowledge were more helpful in our task than others. Shared knowledge about initial situation and goals seemed to be more effective than shared operators, since the solution quality did not improve when shared operators were added in condition SK (IS+G+O). This result contradicted the hypothesis that sharing information about both components of shared knowledge, initial situation and goals on the one hand, and operators on the other, is most effective. In the present study, sharing initial situation and goals alone, was as effective as additionally sharing operators. Apparently, sharing operators did not seem to have a facilitating effect on collaborative problem solving. Although the comparison between conditions SK (IS+G+O) and SK (O) lacked statistical significance, means were in line with the assumption that sharing operators is not very effective in a task like our experimental one. This interpretation is strengthened by the results of directly comparing the effectiveness of shared knowledge about the initial situation and goals, with that about the operators. Shared knowledge about IS+G clearly led to better performance than sharing operators.

But how can this effect be explained? First, the structure of the operators is more complex than that of the initial situation and goal elements. While the latter are simple sentences, the operators always combine an 'if' part with a 'then' part. This means that the operators contain technical solution features (in the 'then' part), in addition to information about possible situations and goals of the company (in the 'if' part). Besides their more complex structure, operators contained more information, framed in technical terms, that was new to the participants. Finally, information about the company provides company experts with a more coherent view of the company, whereas operators, in comparison, form a rather loose collection of rules for possible situations. We may assume, therefore, that operators are more difficult to process than knowledge elements about IS+G. This assumption is supported by the higher reading and learning time for operators than for IS+G. As a consequence, understanding and applying operators should be more difficult for the company expert than understanding information about IS+G is for the IT-advisor. From this, it follows quite naturally that sharing knowledge about IS+G should be more effective than sharing knowledge about operators.

Secondly, the structure of the operators has an impact on the ease of integrating information about IS+G into the 'if' part of the operators. As an operator's 'if' part already contains information about possible states of the company's initial situation and goals, it should be easy for the advisor to integrate the shared IS+G information into his or her operators, since it is just an instantiation of what is already provided in the operators' 'if' part. Thus IS+G information, either shared, or newly received during collaboration, does not provide information completely new to the IT advisor. In contrast, if the company expert receives operators, the solution provided in their 'then' part is completely novel information to him, and not just a specification of what he already knows. To deduce the resulting solution feature might therefore be more difficult. In addition, the advisor is able to realize (when he or she receives the shared knowledge elements) that the task-relevant information from the company expert is just a specification of the knowledge already provided by his or her operators. Asking

the right question to gather the information needed should therefore be easier for the advisor than for the company expert. Having shared knowledge about operators does not allow the company expert to infer other possible solution features, since knowledge about the initial situation and goals does not contain underspecified information about possible solution features. So in contrast to the IT advisor, it should be more difficult for the company expert to develop a coherent view of the partner's expertise and thus ask for the appropriate information.

In conclusion, shared knowledge is an effective variable facilitating collaborative problem solving. In the present paper, we show that sharing knowledge about operators is less effective than sharing the initial situation and goals, as the latter is easier to understand and integrate, and additionally, allows making assumptions about the partners' expertise. But shared knowledge is not only a facilitator for collaborative problem solving that can be provided externally or already exists but is also co-constructed during the problem solving process. Of course, it should also be noted that on the basis of the single experiment reported here, it is an open question whether the results hold for other problems and materials as well. Furthermore, the problem reported here is a well-defined one with a clear solution. Concerning ill-defined problems it can be assumed that the positive effect of shared knowledge about IS+G is reduced. In fact, operators will always be more complex, because per definition they consist of an 'if' and a 'then' part. What is said about the ease of integrating the IS+G information compared to integrating operators and inferring the partner's knowledge will also hold true for ill-defined problems. But it is questionable whether these factors will be as helpful *in finding a solution* because in ill-defined problems several solutions are possible. So, having all information to integrate operators and IS+G information does not unambiguously allow to deduce all solution features. Therefore, we assume that sharing knowledge about IS+G is still effective, given the structural differences explained above, but that this effect is less strong.

Acknowledgements

We thank the Deutsche Forschungsgemeinschaft for funding the work of the first author (DFG, Virtual Graduate Program). We also thank Dietmar Janetzko for his helpful comments on an earlier draft of this paper and Kristen Drake for her language assistance.

References

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In Castellan, J. Jr. (Ed.), *Individual and group decision making. Current issues.* Hillsdale, NJ: Lawrence Erlbaum.

Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. B. Resnick, J. M. Levin, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition.* Washington: American Psychological Ass.

Fischer, F., & Mandl. H. (2000). The construction of shared knowledge in face-to-face and computer-mediated cooperation. *Paper presented at the 81th Annual Meeting of the American Educational Research Association* (AERA).

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin, 121*(1), 43-64.

Jeong, H. & Chi, M. T. H. (2000). Does collaborative learning lead to the construction of common knowledge? In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Moreland, R. L., Argote, L., & Krishnan, R. (1996). Socially shared cognition at work. Transactive memory and group performance. In Nye, J. L., & Brower, A. M. (Eds.), *What's social about social cognition?* Thousand Oaks: SAGE.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewoods Cliffs, NJ: Prentice-Hall.

Pfister, H.-R., Wessner, M., Holmer, T., & Steinmetz, R. (1999). Negotiating about shared knowledge in a cooperative learning envirnonment. In *Proceedings of the Third Confernce on Computer Support for Collaborative Learning*. California: Stanford University.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer Supported Collaborative Learning* (pp. 69-97). Berlin: Springer.

Rouse, W. B., Cannon-Bowers, J. A., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE Transactions On Systems, Man, And Cybernetics, 22*(6), 1296-1308.

Smith, J. B. (1994). Collective intelligence in computer-based collaboration. Hillsdale, NJ: Lawrence Erlbaum.

Smith, G. F., & Browne, G. J. (1993). Conceptual foundations of design problem solving. *IEEE Transactions On Systems, Man, and Cybernetics, 33*(5), 1209-1219.

Stasser, G. & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48,* 1467-1478.

Stasser, G., Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology, 53,* 81-93.

Strube, G., Garg, K., & Wittstruck, B. (2002). Interfaces to expert knowledge: Modelling communication of knowledge in multi-expert teams. *International Conference on Cognitive Modeling (ICCM-2003),* 303-304.

Waern, Y. (1992). Modelling group problem solving. *Zeitschrift für Psychologie, 200,* 157-174.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior.* New York: Springer.

The Impact of Prior Task Experience on Bias in Predictions of Duration

Kevin E. Thomas (k.e.thomas@plymouth.ac.uk)
School of Psychology, University of Plymouth, Drake Circus, Plymouth. Devon. PL4 8AA. UK.

Stephen E. Newstead (s.newstead@plymouth.ac.uk)
School of Psychology, University of Plymouth, Drake Circus, Plymouth. Devon. PL4 8AA. UK.

Simon J. Handley (s.handley@plymouth.ac.uk)
School of Psychology, University of Plymouth, Drake Circus, Plymouth. Devon. PL4 8AA. UK.

Abstract

The effect of prior experience on bias in time predictions on two different types of laboratory task was examined in two studies. Experiment 1 revealed that prior experience of performing a substantial part of the same task led to greater time prediction accuracy. However, contrary to the weight of previous research, there was little evidence of the temporal underestimation indicative of the planning fallacy. In fact, temporal underestimation only occurred on a longer duration task when it was preceded by a much shorter task, which was either related (Experiments 1 and 2) or unrelated to it (Experiment 2). In contrast, temporal overestimation prevailed on tasks ranging from about 30 seconds' to four minutes' duration. Contrary to the theory of the planning fallacy, these studies indicate that people do take account of their performance on previous tasks and use such distributional information when predicting task duration. The potential role of the anchoring and adjustment cognitive heuristics in determining temporal misestimation is discussed.

Introduction

The process of predicting task duration has been the subject of considerable research (e.g., Buehler, Griffin & MacDonald, 1997; Koole & Van't Spijker, 2000). In general, such research has produced evidence of temporal underestimation on various laboratory (e.g., Josephs & Hahn, 1995) and real world tasks (e.g., Buehler, Griffin & Ross, 1994). Such research supports the cognitive judgment phenomenon known as the planning fallacy, which is the tendency to underestimate task duration despite being aware that previous similar activities took longer than anticipated (Buehler, Griffin & Ross, 2002).

The planning fallacy was identified by Kahneman and Tversky (1979), who suggest that distinct two types of data are available to people when predicting task duration. Namely, singular information, which is data about the task at hand; and distributional information, which concerns data about previous tasks. An aspect of singular information is the amount of work involved in completing a current task, whereas personal performance on previous similar tasks is an aspect of distributional information.

Kahneman and Tversky (1979) propose that the planning fallacy is a consequence of heuristic information processing whereby singular information becomes the focus of attention at the expense of distributional information, which is overlooked. Hence, temporal underestimation occurs because the current task is treated as a unique event, which is dissociated from previous similar activities.

Given that the neglect of distributional information has been suggested as a possible cause of the planning fallacy (Kahneman & Tversky, 1979), it is notable that the issue of prior task experience has received little empirical treatment in relation to time estimation. One exception is the work of Thomas, Newstead and Handley (2003; see also Thomas, Handley & Newstead, 2004), which revealed that prior experience of performing (or mentally planning how to complete) certain laboratory tasks led to a reduction in temporal misestimation. However, Thomas et al. (2003) found little evidence of the temporal underestimation indicative of the planning fallacy on short duration (i.e., up to four minutes' duration) laboratory tasks such as the Tower of Hanoi.

In fact, there was evidence of general temporal overestimation on such tasks, with underestimation only occurring on longer tasks when they were preceded by a shorter version of the same task. For example, temporal underestimation prevailed on the five-disk Tower of Hanoi task only when the three-disk version of this task was performed beforehand. The findings of Thomas et al. (2003) indicate that there are certain tasks on which the temporal underestimation indicative of the planning fallacy does not occur and is reversed.

Thomas et al. (2003) suggest that the temporal underestimation they observed may have been a consequence of participants using the anchoring and adjustment cognitive judgmental heuristics (Tversky & Kahneman, 1982). That is, information such as the perceived duration of the first task served as an anchor for time predictions on the second task, which were insufficiently adjusted according to the greater demands of the upcoming task. Such a judgment strategy would be expected to result in temporal underestimation if the perceived duration of a just-completed shorter task served as a basis for time predictions on a current task.

A principal aim of the present research was to further address the issue of prior experience by employing laboratory tasks that are not only less artificial than those employed by Thomas et al. (2003), but are more akin to the ones used in previous research supporting the planning fallacy (e.g., Byram, 1997). That is, tasks that have well-

defined components and must be completed sequentially by following a set of instructions.

The present studies also sought to determine the direction in which time predictions were biased (i.e., under or overestimation) on a laboratory task that takes longer to complete than those used in our earlier work, but is of similar duration to some of the laboratory tasks employed in previous research (e.g., Buehler et al., 1997). Given that the tasks employed by Thomas et al. (2003; 2004) were of shorter duration than the laboratory tasks used in research supporting the planning fallacy (e.g., Josephs & Hahn, 1995), it could be that temporal underestimation is only evident on tasks that take longer than four or five minutes to complete. Consistent with this suggestion, temporal underestimation has been observed on laboratory tasks ranging in duration from about 10 minutes (e.g., Francis-Smythe & Robertson, 1999) to over one hour (Byram, 1997). The issue of task duration was addressed in the present studies by employing tasks that were of similar duration to those used in our previous research alongside one that took longer to complete.

Experiment 1

The issue of task duration was addressed in this study by using three different versions of the same miniature construction kit (i.e., toy castle) manufactured by Playmobil®. One of these tasks (long duration task) took about 11 minutes to complete whilst the others took either four minutes (medium duration task) or 30 seconds to finish (short duration task). The medium and short tasks were sub-component versions of the long duration task, and involved constructing different parts of the same miniature castle. The issue of prior task experience was addressed by varying the order in which the long duration task was performed. That is, whether time prediction bias differed when the long task was performed after, or was preceded by, one of the two shorter tasks.

Method

Participants. Eighty (64 female and 16 male) students at the University of Plymouth participated voluntarily in partial fulfillment of a psychology course requirement. No biographical data other than gender was recorded.

Design, Materials and Procedure. A 2 (time: predicted vs. actual duration) x 4 (task experience: long then short task vs. short then long task vs. medium then long task vs. long then medium task) mixed factorial design was used. The time factor was a repeated measure, with participants producing a predicted and actual task completion time. Task experience was manipulated between groups, with participants being randomly assigned to one of the four equal-sized conditions.

Prior to judging task duration, the amount of time that participants were given to view the task components and instruction booklet differed according to the type of task that was about to be performed. Pilot testing revealed that 80 seconds were needed to preview the instruction booklet and the plastic components of the long task. Pilot testing revealed that the instruction booklet and the plastic components of the short and medium tasks could be previewed in 20 and 40 seconds, respectively.

The long duration task involved building a multi-turreted castle with surrounding jetty and battlements by assembling a series of molded plastic components in a pre-specified order. The medium duration sub-component task involved building the castle without the surrounding jetty and battlements. The short duration sub-component task involved building one wall of the castle. A digital stopwatch was used to measure task duration.

Results

Means (and standard deviations) of predicted and actual completion time on the second task are presented in Table 1.

Table 1: Predicted and Actual Duration of the Second Task Per Task Experience Condition (In Seconds).

Time	Task Experience Condition			
	Short-Long Task (n = 20)	Long-Short Task (n = 20)	Medium-Long Task (n = 20)	Long-Medium Task (n = 20)
Predicted	435.00 (181.70)	28.15 (20.12)	550.50 (129.59)	254.25 (109.20)
Actual	556.25 (147.18)	18.55 (18.00)	497.85 (124.14)	178.95 (35.68)

There was considerable variability within the predicted and actual task completion time data, and frequency distributions of these data from each task experience condition were positively skewed. Hence, these data were subjected to a logarithmic transformation before being statistically analyzed.

A 2 (time) x 4 (experience) split-plot analysis of variance (ANOVA) produced a main effect of task experience, $F(3,76) = 637.66$, $MSE = .14$, $p < .001$, with overall time being longest in the medium then long task condition. Pairwise comparisons (Scheffé) revealed significant differences between the means of all conditions ($ps < .05$) except those of the medium then long task and the short then long task conditions ($p > .10$). The main effect on the time factor was not significant ($F < 3$, $p > .10$).

The ANOVA also produced an interaction, $F(3,76) = 6.12$, $MSE = .11$, $p < .01$ (see Figure 1 below). This revealed that temporal overestimation was evident on the medium and short duration tasks, whereas the direction in which time predictions were biased on the long task differed according to the relative duration of the previous task. Specifically, temporal overestimation was evident when the medium task had just been completed, whereas underestimation occurred when the short task was performed initially. Tests for simple effects revealed that predicted and actual time differed significantly on the

medium and short duration tasks (ps < .05). On the long duration task, the difference between predicted and actual time was significant when the short task (p < . 05), but not the medium task was performed beforehand (p > .10).

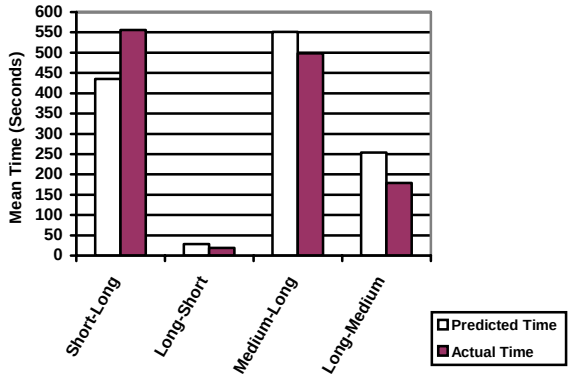


Figure 1: Predicted and actual completion time on the second task per task experience condition

Discussion

Temporal overestimation was evident on the medium and short duration sub-component tasks when they were performed after the long task. Consistent with the work of Thomas et al. (2003; 2004), this finding indicates that the temporal underestimation indicative of the planning fallacy (Kahneman & Tversky, 1979) is not evident (and is reversed) on another type of laboratory task with a duration of less than five minutes.

There was some evidence of temporal underestimation on the long duration task, but only when the short sub-component task had just been completed. The latter finding concurs with the notion that individuals might use the anchoring and adjustment cognitive heuristics when judging the duration of the second of two consecutive tasks (Thomas et al., 2003). That is, information about the previous task serves as an anchor for time predictions, which are insufficiently adjusted according to the relative demands of the current task.

In contrast, temporal overestimation (rather than underestimation) prevailed on the long task when the shorter duration medium sub-component task was performed beforehand. Thus, it seems that an anchoring and adjustment strategy was not used when judging the duration of the second task when a sizeable portion of this task had just been completed.

A possible explanation for these findings is that, due to differences in the extent of prior experience, different kinds of task-related information were used when predicting the duration of the long task. As completion of the medium task involved assembling half of the long task, participants in this experimental condition possessed considerable information about the nature of the long task when predicting its duration.

Given the extent of these individuals' prior task experience, they might have engaged in more thorough information processing when making a second time prediction. For example, they may have calculated the number of large plastic components required to complete the medium task, and appropriately scaled up this figure as a function of the greater number of major components involved in finishing the long duration task.

Using such a judgment strategy could result in temporal overestimation if it involved thinking about factors that delayed the completion of the previous task (e.g., fitting some plastic components together incorrectly). Thus, these participants may have erred on the side of caution because they took account of their previous task performance. In fact, time predictions were more accurate when the long task was preceded by the medium rather than the short task, suggesting that greater prior task experience was used to good effect.

As participants who performed the short task initially constructed only one wall of the Playmobil® castle (i.e., one part of the long task), they possessed little information about how to complete the long task when predicting its duration. In the absence of substantial prior task experience, these individuals may have used heuristic information processing when making a second time prediction. For example, time predictions may have been anchored on the perceived duration of the first task with insufficient upward adjustment for the longer duration of the second task. Thus, due to insufficient prior task experience, these participants may have relied on the anchoring and adjustment cognitive heuristics when judging the duration of the long task.

Whilst this study suggests that time predictions on the longer of two successive tasks might be based on information about the first task, the nature of this task-related information is not known. Given that the present tasks differed in duration, it could be that an anchoring and adjustment judgment strategy involving the perceived duration of the previous task is responsible for temporal underestimation when a longer task follows a shorter one.

In contrast, as the present tasks share the same structure (i.e., they are different versions of the same task), it could be that information about the nature of the first task formed the basis of time predictions on the second task. For example, the number of major plastic components involved in completing the previous task could serve as an anchor for time predictions on the current task. Temporal misestimation would be expected to occur as a consequence of using this kind of judgment strategy if the number of major plastic components differed between the first and second tasks. That is, if time predictions were not sufficiently adjusted from an anchor value to take account of the number of major plastic components needed to complete the second task.

Having found evidence of temporal underestimation on the long duration task when a much shorter version of it had just been completed, Experiment 2 sought to determine the type of information about a just-completed shorter task that

formed the basis of time predictions on the long duration task.

Experiment 2

The issue of the relevance of prior task experience was addressed in this study, where a related or an unrelated shorter duration task was performed before the long Playmobil® task. The related task was the short duration task from Experiment 1, whereas the unrelated task was the three-disk version of the Tower of Hanoi task. Pilot testing revealed that the three-disk task and the short sub-component task were of similar duration (Ms = 28.59 and 25.37 seconds, respectively).

Performing the short Playmobil® task initially would provide participants with some information about the nature of the long duration task, whereas no information about the long duration task would be acquired whilst completing the three-disk task.

If time predictions were based on information about the nature of the previous task, then they should be more accurate on the long duration task when the related task was performed beforehand. Conversely, if time predictions were based on information such as the perceived duration of the previous task, then the extent of judgment bias on the long task should not differ according to the relevance of prior experience.

Method

Participants. Fifty-six (42 female and 14 male) students at the University of Plymouth participated voluntarily. Forty-three individuals participated in partial fulfillment of psychology course requirement whilst the remainder were paid £2.50 each. No biographical information other than gender was recorded.

Design, Materials and Procedure. The long duration Playmobil® task, and a wooden Tower of Hanoi task apparatus containing three different-sized disks were used. A digital stopwatch was used to measure task duration.

There were two equal-sized groups of participants who performed either the three-disk task or the short sub-component task before the long duration Playmobil® task. The amount of time that participants were given to preview the plastic Playmobil® task components and instruction booklet differed according to the type of task that was about to be completed.

On the three-disk and short duration sub-component tasks, participants were given 20 seconds to preview the instructions and task apparatus or plastic components. Participants previewed the plastic components and instruction booklet of the long duration task for 80 seconds.

Results

Means (and standard deviations) of predicted and actual completion time on the long task are presented in Table 2.

Table 2: Predicted and Actual Duration of the Long Task Per Prior Experience Condition (In Seconds).

Time	Prior Experience Condition	
	3-disk Task First (n = 28)	Short Task First (n = 28)
Predicted	412.50 (137.99)	432.86 (177.66)
Actual	614.04 (125.32)	599.07 (140.91)

For the same reasons that were specified in Experiment 1, the predicted and actual completion time data were subjected to a logarithmic transformation before being statistically analyzed. A 2 (time) x 2 (task experience) split-plot ANOVA produced a main effect of time, $F(1,54) = 61.81$, $MSE = .07$, $p < .001$, with completion times exceeding predictions (Ms = 606.56 and 422.68 seconds, respectively). This finding indicates that temporal underestimation was evident on the long duration task. The main effect of prior experience and the interaction were not significant ($Fs < 1$, $ps > .10$). The absence of a significant interaction suggests that the extent of temporal underestimation on the long duration task did not differ according to the type of shorter task that was performed beforehand.

Discussion

Consistent with the results of Experiment 1, temporal underestimation was evident on the long duration Playmobil® task when the short sub-component task was performed beforehand. Temporal underestimation also prevailed on the long task when it was preceded by an unrelated task that was of much shorter duration (i.e., the three-disk Tower of Hanoi task).

The presence of temporal underestimation on the long task is consistent with Thomas et al.'s (2003) suggestion that the anchoring and adjustment cognitive heuristics are used when judging the duration of the second of two successive tasks. However, as the extent of underestimation did not differ significantly according to the relevance of prior experience, it seems that information about the nature of the previous task was not used as a basis for time predictions on the long duration task. Instead, some other kind of task-related information presumably served as a basis for time predictions on this task.

A possible candidate source of such information is the perceived duration of the previous task. That is, individuals judged how long the first task took to complete, and used this figure as a basis for their second time prediction. Indeed, at the end of the first experimental trial, several participants commented that the just-completed task had taken them less time to finish than they predicted. Whilst

such evidence is purely anecdotal, it indicates an awareness of temporal misestimation on both of the short tasks, and suggests that several individuals estimated the duration of the first task retrospectively.

An anchoring and adjustment judgment strategy involving the perceived duration of the previous task should result in time prediction bias when successive tasks differ in duration. That is, temporal misestimation is a consequence of failing to increase or decrease the current prediction according to the longer or shorter duration of the upcoming task (Thomas et al., 2003). Such insufficient adjustment from an anchor value (i.e., the perceived duration of the previous task) would lead to temporal underestimation if the current task took longer to complete than the previous one.

General Discussion

The present studies provide further insight into the role of prior experience in the process of predicting task duration. In Experiment 1, we found that, relative to building just one wall of the Playmobil® castle initially, constructing half of the castle on the first trial resulted in greater time prediction accuracy on the long duration task. This finding is consistent with previous research, which has found that prior experience attenuates bias in temporal (e.g., Josephs & Hahn, 1995) and non-temporal judgments (e.g., Smith & Kida, 1991) of task performance.

More importantly, this finding suggests that performance on previous similar activities is not only considered when judging task duration, but can also be used to good effect (i.e., to improve time prediction accuracy). Given that distributional information seems to be a key component of the planning fallacy, the role of prior task experience in mediating temporal misestimation is in need of further study. That is, further insight into how such distributional information can be used effectively will enhance our understanding of the planning fallacy phenomenon.

Whilst it has been shown that possessing considerable prior task experience reduces temporal misestimation (Experiment 1), the present research also indicates that the use of such distributional information does not always improve judgment accuracy. In both studies, there was evidence of temporal underestimation on the long Playmobil® task when it was preceded by a much shorter duration sub-component task. However, the extent of this temporal underestimation was similar when either the short sub-component task or an unrelated short duration task was performed initially (Experiment 2).

Consistent with our previous work (Thomas et al., 2003; 2004), this finding indicates that information about a just-completed similar task is considered when predicting task duration, but can lead to judgment bias. If, as we propose, an anchoring and adjustment strategy involving the perceived duration of a previous shorter task forms the basis of time predictions on a longer task, then an alternative interpretation of the planning fallacy suggests itself. That is, temporal underestimation is a consequence of time predictions being based on the shorter duration of a previous task, but being insufficiently scaled up according to the greater demands of the current task.

Whilst it is for future research to determine whether the present findings generalize to more everyday kinds of task, the use of the anchoring and adjustment cognitive heuristics could explain the prevalence of the planning fallacy on many large scale projects. That is, individuals who undertake such projects will typically have experience of performing similar but less complex tasks (Kidd, 1970). Moreover, as large scale (e.g., construction) projects tend to be performed infrequently, judgments of their duration can only really be based on the shorter duration of previous less complex tasks. If time predictions are anchored on the duration of previous smaller scale tasks, then temporal underestimation would be expected to occur.

In both studies, there was some evidence of the temporal underestimation indicative of the planning fallacy on the long duration Playmobil® task. This finding suggests that temporal underestimation might only be evident on laboratory tasks that are of longer duration than those employed in our earlier research. However, temporal underestimation was not evident on the long duration task when the medium sub-component task was performed initially (Experiment 1).

It was suggested that temporal overestimation on the long duration task was due to participants taking account of factors that delayed the completion of the medium task (e.g., incorrectly fitting some plastic components together) and incorporating such information into their second time prediction. Although further research is required to test the validity of this claim, it has been shown that thinking about such information can reduce bias in non-temporal judgments of task performance (e.g., Koriat, Lichtenstein & Fischhoff, 1980).

Given the present findings, it could be that, when prior experience is substantial, people incorporate potential impediments to optimal task completion into their temporal judgments on subsequent tasks. This kind of judgment strategy might lead to temporal overestimation, and also to greater time prediction accuracy. Support for the latter suggestion comes from Experiment 2, where time predictions on the long duration task were less biased when participants possessed more extensive prior task experience. That is, when the medium rather than short duration sub-component task was performed beforehand.

The existence of temporal overestimation on the short and medium duration sub-component tasks (Experiment 1) highlights the directional nature of time prediction bias. A possible explanation for the presence of temporal overestimation on tasks with a duration of up to four minutes is that people tend to judge task duration in whole minutes rather than seconds, or by using longer temporal units such as 5 or 10 minutes (Fraisse, 1984).

Given the duration of the two shorter tasks used in Experiment 1, temporal overestimation should prevail if participants used temporal units such as five minutes when judging their completion times on the medium sub-

component task. Likewise, giving a time prediction of two or three minutes would be expected to result in temporal overestimation on the short sub-component task. Thus, the reversal of the temporal underestimation indicative of the planning fallacy on the two shorter Playmobil® tasks could be a consequence of the type of time unit used to judge task duration.

References

Buehler, R., Griffin, D. & MacDonald, H. (1997). The role of motivated reasoning in optimistic time predictions. *Personality and Social Psychology Bulletin, 23, 3,* 238-247.

Buehler, R., Griffin, D. & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology, 67, 3,* 366-381.

Buehler, R., Griffin, D. & Ross, M. (2002). Inside the planning fallacy: The causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.

Byram, S. J. (1997). Cognitive and motivational factors influencing time predictions. *Journal of Experimental Psychology: Applied, 3, 3,* 216-239.

Fraisse, P. (1984). Perception and estimation of time. *Annual Review of Psychology, 35, 1,* 1-36.

Francis-Smythe, J. A. & Robertson, I. T. (1999). On the relationship between time management and time estimation. *British Journal of Psychology, 90, 3,* 333-348.

Josephs, R. A. & Hahn, E. D. (1995). Bias and accuracy in estimates of task duration. *Organizational Behavior and Human Decision Processes, 61, 2,* 202-213.

Kahneman, D. & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in the Management Sciences, 12,* 313-327.

Kidd, J. B. (1970). The utilization of subjective probabilities in production planning. *Acta Psychologica, 34,* 338-347.

Koole, S. & Van't Spijker, M. (2000). Overcoming the planning fallacy through willpower: Effects of implementation intentions on actual and predicted task-completion times. *European Journal of Social Psychology, 30,* 873-888.

Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6, 2,* 107-118.

Smith, J. F. & Kida, T. (1991). Heuristics and biases: Expertise and task realism in auditing. *Psychological Bulletin, 109, 3,* 472-489.

Tversky, A. & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic & A. Tversky (Eds.). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Thomas, K. E., Handley, S. J. & Newstead, S. E. (2004). The effects of prior experience on estimating the duration of simple tasks. *Current Psychology of Cognition, 22, 1,* 83-100.

Thomas, K. E., Newstead, S. E. & Handley, S. J. (2003). Exploring the time prediction process: The effects of task experience and complexity on prediction accuracy. *Applied Cognitive Psychology, 17, 6,* 655-673.

Acknowledgments

This research was supported by a research grant from the Economic and Social Research Council (ESRC) of the United Kingdom (Award No. R42200034413) to the first author. Requests for re-prints of this article should be addressed to: Kevin Thomas, School of Psychology, University of Plymouth, Drake Circus, Plymouth. Devon. PL4 8AA. UK.

Visual Expertise Depends on How You Slice the Space

Brian A. Tran (b3tran@ucsd.edu)
UCSD Department of Computer Science and Engineering, 9500 Gilman Drive
La Jolla, CA 92093-0114 USA

Carrie A. Joyce (cjoyce@cs.ucsd.edu)
UCSD Department of Computer Science and Engineering, 9500 Gilman Drive
La Jolla, CA 92093-0114 USA

Garrison W. Cottrell (gary@cs.ucsd.edu)
UCSD Department of Computer Science and Engineering, 9500 Gilman Drive
La Jolla, CA 92093-0114 USA

Abstract

Previous studies using fMRI have found that the Fusiform Face Area (FFA) responds selectively to face stimuli. More recently however, studies have shown that FFA activation is not face-specific, but can also occur for other objects if the level of experience with the objects is controlled. Our neurocomputational models of visual expertise suggest that the FFA may perform fine-level discrimination by amplifying small differences in visually homogeneous categories. This is reflected in a large spread of the stimuli in the high-dimensional representational space. This view of the FFA as a general, fine-level discriminator has been disputed on a number of counts. It has been argued that the objects used in human and network expertise studies (e.g. cars, birds, Greebles) are too "face-like" to conclude that the FFA is a general-purpose processor. Further, in our previous models, novice networks had fewer output possibilities than expert networks, leaving open the possibility that learning more discriminations, rather than learning fine-level discriminations, may be responsible for the results. To challenge these criticisms, we trained networks to perform fine-level discrimination on fonts, an obviously non-face category, and showed that these font networks learn a new task faster than networks trained to identify letters. In addition, all networks had the same number of output options, illustrating that visual expertise does not rely on number of discriminations, but rather on how the representational space is partitioned.

Introduction

The Fusiform Face Area (FFA) in the ventral temporal lobe has recently received much attention. Initial work appeared to show that this area was selective for processing faces. Several fMRI studies showed high activation in the FFA only to face stimuli and not other objects (Kanwisher et al., 1997; Kanwisher, 2000). Further, studies involving patients with *associative prosopagnosia*, the inability to identify individual faces (Farah et al., 1995), and *visual object agnosia*, the inability to recognize non-face objects (Moscovitch et al., 1997), seemed to indicate a clear double dissociation between face and object processing. Prosopagnosic patients had lesions that encompassed either right hemisphere or bilateral FFA, while object agnosic patients' lesions did not (De Renzi et al., 1994).

Gauthier and colleagues have challenged the notion of the face specificity of the FFA by pointing out that the earlier studies failed to equate the level of experience subjects had with non-face objects, to the level of experience they had with faces (Gauthier et al., 1997; Gauthier et al., 1999a; Gauthier et al., 1999b). She showed that the FFA was activated when bird and dog experts were shown pictures of the animals in their area of expertise. Further, she illustrated that, if properly trained, individuals can develop expertise on novel, non-face objects (e.g. Greebles), and subsequently show increased FFA activation to them (Gauthier et al., 1999a). Expertise in these studies was operationally defined as the point in training when a subject's default response level (i.e. entry level) "shifts" from basic to the individual level. This is indexed by the subject's reaction time for verifying individual names becoming as fast as the time to verify category membership.

Neurocomputational models done first by Sugimoto and Cottrell (2001) and later extended by Joyce and Cottrell (2004) began to address the question of how and why the FFA gets recruited for these other tasks (Sugimoto & Cottrell, 2001, Joyce & Cottrell, 2004). Using four different types of stimulus classes (books, cans, cups and faces), Sugimoto and Cottrell found that the amount of expert-level experience on a previous task correlates with faster subordinate level learning relative to a system that processes the same stimuli, but not to a subordinate level. Thus, an area that is used for one expertise task will learn a second expertise task faster than an area used only for basic level discriminations. Joyce and Cottrell (2004) further found that an expert network's ability to separate individuals is reflected in highly variable responses at the representational layer (the hidden layer). This response variability extended to novel categories, permitting faster learning of these

categories. This suggests that the FFA is primed to win the competition for a new expertise task because of its ability to fine-tune its feature representations when given a novel fine-level discrimination task (Joyce & Cottrell, 2004).

While the human and computational studies of expertise are compelling, they are not undisputed. For example, proponents of the view that the FFA is face-specific claim that the objects used in human expertise studies, such as cars, dogs, birds, or Greebles, are "face-like", meaning they possess properties similar to faces. Thus any response of FFA to these stimuli is due to their featural similarity with faces, not because the FFA is a general-purpose, fine-level discriminator. While the network simulations, which illustrate expertise across a wide variety of non-face objects, may seem to argue against this criticism, a methodological issue makes these results less compelling. In previous simulations, non-expert networks were trained on a lesser number of discriminations (4 category labels) than expert networks (10 individual labels plus the 4 category labels). It has been argued (Mike Tarr, personal communication) that if an object recognition network simply had to make as many discriminations as the expert one, then it would also be able to learn Greebles faster.

The current simulations were designed to address the criticisms cited above. First, we train the networks to perform fine-level discrimination on an obviously non-face category: fonts. In this case, the basic level networks learn to identify letters presented in a variety of different fonts (a task any human can do with ease) while the subordinate level networks learn to distinguish the particular font in which a letter is written (a task few humans can do). To address the second criticism, we present both basic and subordinate level networks with the same stimulus set and have them perform an equal number of discriminations (e.g. 6 letter vs. 6 font discriminations). Thus, any advantage to learning to distinguish Greebles by the font network over the letter network cannot be due to the number of discriminations learned.

Experiments

We ran two sets of experiments. In the first, we investigated the ability of our basic visual object processing architecture (Dailey & Cottrell, 1999; Dailey et al. 2002; Joyce & Cottrell, 2004) to recognize letters and fonts. This allowed us to discover which fonts were difficult and which letters gave good generalization once their font had been learned (by training on other letters). We then used these results in the second set of experiments to perform a very controlled version of our previous "basic versus expertise network" experiments, and investigate generalization to Greeble expertise.

Experiment 1: Stimuli and Methods

The images used were 300x300 pixel images of letters. For this experiment, 15 different fonts were used, and for each of those 15 fonts, we had images of all 26 letters. The fonts were chosen to be somewhat difficult. Image preprocessing

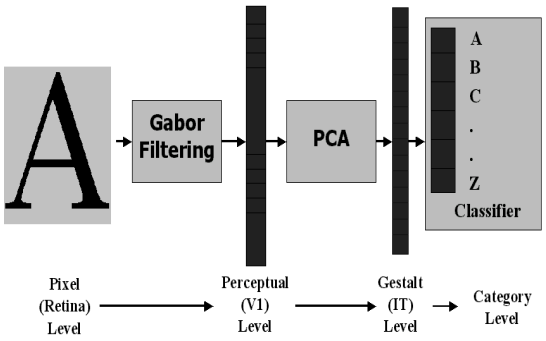


Figure 1: The expertise model.

of the different letters and fonts followed the procedures outlined in Dailey and Cottrell (1999). Each image was first processed using 2-D Gabor wavelet filters (5 spatial frequencies at 8 orientations each), a simple model of complex cell responses in visual cortex. The filters were applied at 64 points in an 8x8 grid, resulting in a vector of 2560 elements (Buhmann, Lades & von der Malsburg, 1990; Dailey & Cottrell, 1999). The vectors were then normalized via z-scoring (scaled and shifted so that they had zero mean and unit standard deviation) on a per-filter basis, a local operation. A principal components analysis (PCA) was then applied to the normalized vectors. The top 40 components were saved and renormalized. Projections of the stimuli onto these 40 dimensional vectors constituted the input to the networks. Figure 1 shows the expertise model, which includes the image preprocessing procedure.

A standard backpropagation network architecture was used for learning classifications. The network had 40 input units, each representing a principal component vector, a 30-unit hidden layer using the logistic sigmoid function, and 15 linear output units for the font network, and 26 linear outputs for the letter network. The learning rate and momentum were 0.01 and 0.5, respectively.

Letter training Letter networks were trained to identify letters across a subset of the 15 different fonts. Each network was given the letters from 13 different fonts as the training set and another font as the holdout set. It was then

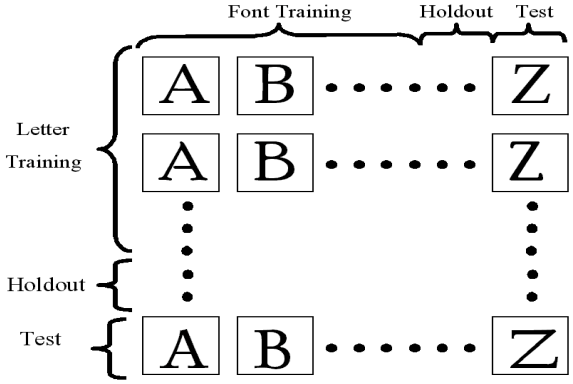


Figure 2: Training for Experiment 1. 26 letters and 15 fonts were used.

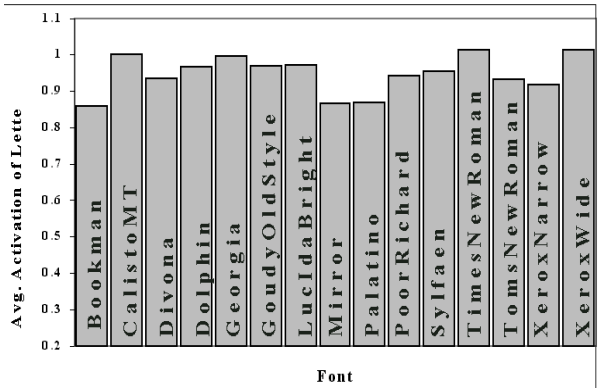


Figure 3: Average activation of letters for test fonts.

tested on the letters from the remaining font. Training was stopped at either an RMSE of 0.02 or when overtraining started to occur. The result was 15 letter networks, all trained and tested on different fonts. Figure 2 illustrates the training and test sets for Experiment 1.

Letter networks learned their task quickly. Figure 3 illustrates the average activation of letters for each font when that font was used as the test set. The amount of activation of an output unit can be thought of as the level of confidence that the letter unit activated corresponds to the correct letter. Although the letters in some fonts were harder to generalize to than others, the average activations were quite high across all fonts. Accuracy of the networks was also computed: if the activation of the unit corresponding to the correct letter is the highest among all other units, then the network was correct in naming the letter. As expected, all letter networks were able to name the correct letter with 100% accuracy.

Font training Training networks to be font experts (i.e. identify the font a letter is written in) for 15 different fonts proved to be quite difficult. Our networks never satisfactorily learned the problem. In order to determine which fonts were easy enough to learn, we performed multidimensional scaling on the distances between the fonts. Distances between fonts were defined as one minus the

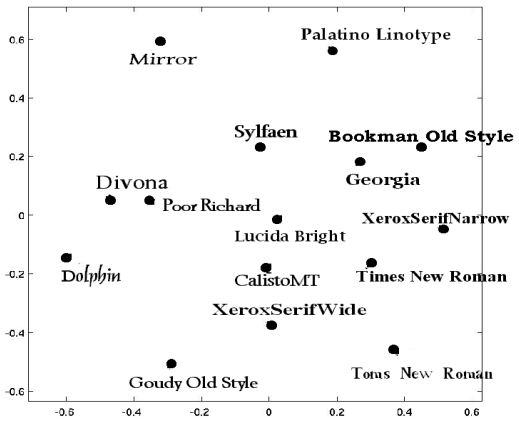


Figure 4: MDS of fonts.

	Avg. RMSE	Avg.Accuracy(%)
Easy Font Network	0.3419	86.19
Hard Font Network	0.4382	76.62

Table 1. Average RMSE and accuracy for font networks

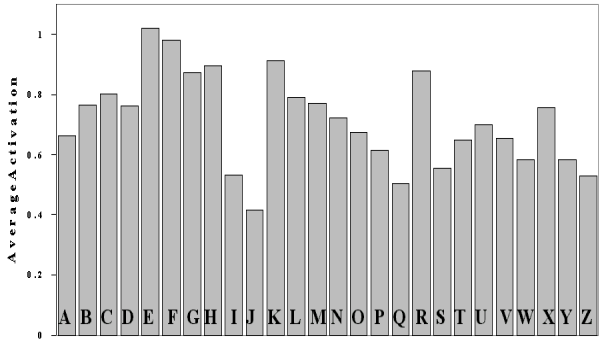


Figure 5: Average activation of fonts across a test letter.

average correlations between their corresponding letters, using their PCA representations. We then formed a 15 by 15 matrix of inter-font distances, and submitted this to a standard non-metric multidimensional scaling routine. The results for a two dimensional solution are shown in Figure 4. The plot shown had a stress of 1.9694.

We used this graph to find the three most separated and the three most correlated fonts. One group of networks was trained on the easier fonts (3 least correlated) and another group of networks was trained on the harder fonts (3 most correlated). Here, 24 letters from each font were used as the training set, 1 letter as the holdout set, and 1 letter as the test set. This was repeated so that each letter had a chance to be the test letter once. Training was stopped when overtraining started to occur. The result is 52 different networks, 26 from the easy font training and 26 from the hard font training.

With these reduced training sets, networks were successfully able to learn to discriminate fonts. Verifying our analysis, the hard font networks had a slightly harder time learning the task than the easy font networks (Table 1). Although the RMSE for both networks were high, they were still accurately able to name the correct font. In fact many of the networks had an accuracy of 100%.

We again computed average activations of output units, except this time for fonts across a test letter (Figure 5). The importance of this plot comes in the activation for particular letters. A high activation means that the network had an easier time generalizing to the font in that letter. This assisted us in choosing the highly generalizable letters as stimuli for Experiment 2.

Experiment 2: Stimuli and Methods

As discussed in the Introduction, Experiment 2 was carried out in order to provide a novel control for our computational model of the visual expertise hypothesis. We used the six

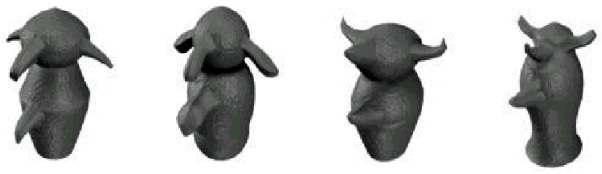


Figure 6: Examples of Greeble images.

most discriminable fonts, and the six letters that were the easiest to generalize to between fonts. Two sorts of networks, both with exactly the same training set, preprocessing, architecture, and number of outputs, were then trained to be either letter classifiers or font experts. While one might consider a letter network an "expert," for our purposes, we consider it a basic level classifier. The main characteristic of basic level categorization is that similar things are classified into the same category. The font expert, on the other hand, must take similar things (the same letter in different fonts) and differentiate between them. Our hypothesis is that such a network will learn the Greeble task faster than a letter network. Thus, training in Experiment 2 was divided into two separate phases. Phase 1 involved training the letter and font networks in a manner similar to that of Experiment 1. In Phase 2, the letter and font networks trained in Phase 1 were then trained to classify Greebles. Examples of Greebles are shown in Figure 6.

Using the results from Experiment 1, the 6 most generalizable letters and the 6 most discriminable fonts were chosen as the stimuli. In addition to these 36 stimuli, 5 different images of 10 unique Greebles were introduced in phase 2. The five different images were produced by jittering the image of a specific Greeble a few pixels on the x, y or x and y axes. Preprocessing of the images was as described in Experiment 1. Greeble images were also preprocessed using Gabor filters and PCA, however they were not included in the generation of the PCA eigenvectors. Rather, the eigenvectors produced via the PCA on the letter/font stimuli were applied to the Greebles. Thus, the PCA representations given to the networks contained no *a priori* information about how Greebles fit into the representational space.

As in Experiment 1, the networks consisted of 40 input units. The hidden unit layer was increased to 40 units due to the increased difficulty of having to solve two tasks. Finally, there were 16 output units, where 6 represented the category (fonts or letters), and 10 the Greebles. Learning rate and momentum remained the same.

Training procedures in Phase 1 were similar to that of Experiment 1, except that only 6 letters and 6 fonts were used. Here, 10 letter networks were trained such that for each network, the letters for a randomly selected font were used as the test set, the letters from another font were the holdout set, and the remaining 4 were used for training. For font networks, each of the 10 networks was tested on the fonts for a randomly selected letter, another randomly selected letter was used as holdout, and the rest were for training. All networks were trained to 2560 epochs. At each log base 2 epoch of training in Phase 1, the weights of the

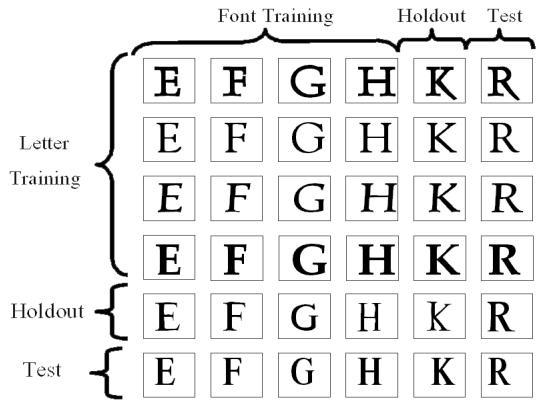


Figure 7: Phase 1 training for Experiment 2. 6 letters and 6 fonts were used.

letter and font network were saved. These weights were used as the starting points for networks in Phase 2 in order to show how varying levels of experience with a preliminary task affected learning of a secondary subordinate level task. For this phase, both font and letter networks ignored the 10 Greeble output units. This training procedure is shown in Figure 7.

In Phase 2 training, the networks trained in Phase 1 were trained to perform subordinate level classification on 10 Greebles. Training for this phase stopped when an RMSE of 0.05 was reached.

Experiment 2: Results

Phase 1 Training Based on the results from Experiment 1, we trained letter and font networks on stimuli that seemed the easiest to generalize to. Both networks were able to learn the task with extremely low error. As expected, the letter networks initially had an easier time learning the letters than the font networks did learning fonts. More importantly, accuracy on the fine-level discrimination task (classifying fonts) became just as good as basic level discrimination (classifying letters).

Phase 2 Training In the second phase of Experiment 2, the letter and font networks were trained to perform fine-level discrimination on Greebles. Again the results were as

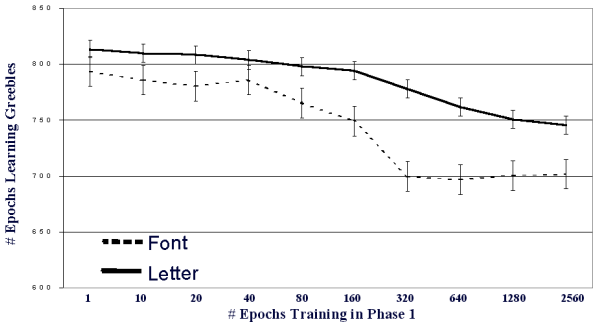


Figure 8: Number of epochs to learn the new task. Error bars denote standard error.

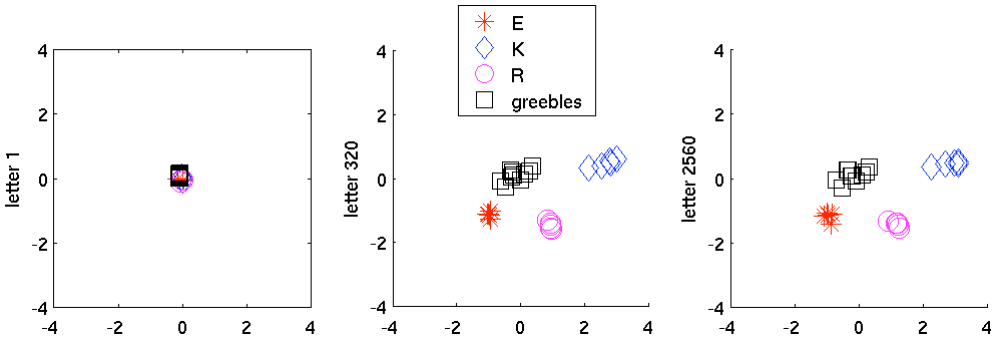


Figure 9: PCA of hidden units of letter network. Grouped by letter.

expected. Figure 8 shows the time in epochs needed for both the letter and font networks to learn the Greeble task as training on the initial task (either classifying letters or fonts) increased. All font networks, regardless of amount of training, learned the Greeble task faster than the letter networks. In addition, more experience with the font task resulted in improvement on learning the Greeble task while more experience with the letter task yielded little improvement (although there is some indication that the letter networks may catch up eventually). Further work will be necessary to evaluate this trend. However, the point remains that expertise in fonts is better than expertise in letters for Greeble training.

To further understand the behavior of the networks, PCA was done on the hidden unit representations prior to Greeble training. Figures 9 and 10 illustrate the spread of the stimuli in representational space based on the 2nd and 3rd principal components (the first PC just codes the overall magnitude change in the weights). In Figure 9 the six points in each symbol represent a given letter in 6 different fonts for a letter network, with one additional symbol representing how Greebles are represented prior to any training on Greebles. In Figure 10, each symbol represents a given font for a font network, and each individual point in that symbol a different letter.

Notice that for the letter network (Figure 9), the letters are grouped together by letter identity regardless of font. Similar inputs (the letters) are made *more similar* by this mapping. In the font network (Figure 10), over training, the fonts spread farther apart over time. Hence, in order to classify the font of each letter, the network must *amplify* small differences between similar items -- all the stimuli representing the same letter must be classified differently. This generalizes to the Greebles; in the font network, the Greebles are more spread out, making it easier for the font network to learn the distinctions between them. Figure 10 also shows that the fonts appear less spread out than the Greebles. This is because the network has learned to see all of the letters in the same font as "the same," whereas it has not learned anything about Greebles yet. It should be noted that each Greeble point is a different Greeble, so the network is already individuating them to some extent. These results are similar to those gathered in our previous network simulations using faces, cups, cans, books, and Greebles (Sugimoto & Cottrell, 2001; Joyce & Cottrell, 2004) illustrating that expertise in the font networks is due to the same mechanism as expertise in face and non-face object networks.

Conclusion

The current studies illustrate that: 1) expertise can be obtained with decidedly non-face-like stimuli and that font expertise exhibits similar properties to that of face and non-face objects seen in previous simulations, and 2) the expertise in previous simulations cannot be explained by a

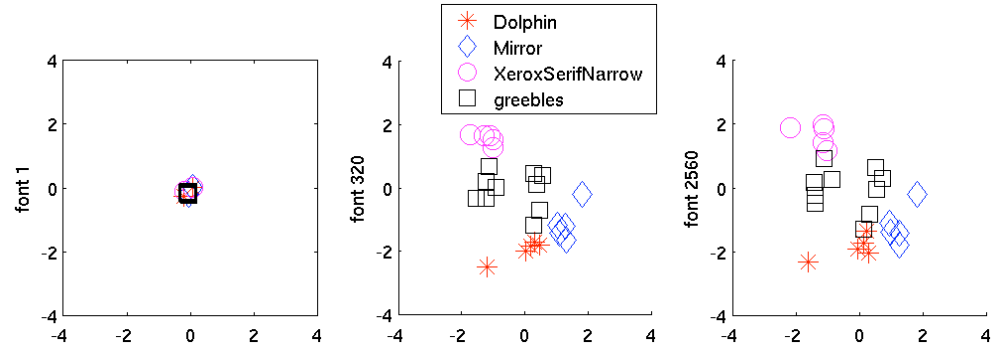


Figure 10: PCA of hidden units on font network. Points grouped by font.

greater number of subordinate level discriminations than basic level discriminations: in the current work these were equated and the results were qualitatively similar to those we have obtained previously.

Our first experiment gave us useful preliminary data for training font experts; it showed that the task of classifying fonts was possible, and revealed which letters and fonts were the easiest to generalize to and train on. The behavior of the networks in the second experiment was similar to previous studies, although our stimuli were different fonts, not "face-like" objects. When training the networks on a new task, the font expert networks learned Greeble classification faster than the letter networks, suggesting that previous visual expertise, whether it be on object or non-object, leads to relatively faster learning in a novel discrimination task. In addition, an equal number of discriminations were required of both letter and font networks. Thus, the expertise advantage could not be due to the sheer *number* of partitions the representational space was divided into, but instead is due to *how* the space was divided. We conclude that visual expertise does not depend on the type of stimuli, nor on the number of stimuli used for training, but on how you slice the space.

Future Work

We plan to train face networks to become font experts, thus generalizing the Greeble expertise work. We expect face expert networks will learn font expertise faster than basic level categorizers. We then plan to train human subjects to become font experts, using fMRI to image both prior to and after training to ascertain if font expertise training engages the FFA. We expect that the letter areas found in the left hemisphere will not become more highly activated by font training. Although it should be obvious from the way that letters are grouped together by the letter network, a future simulation should show that letter networks are difficult to train in font expertise.

Acknowledgements

We would like to thank Gary's Unbelievable Research Unit, the Perceptual Expertise Network, and the anonymous reviewers for comments. This work was supported by NIMH grant MH57075 to GWC and McDonnell Foundation grant #15573-S6 to the Perceptual Expertise Network, Isabel Gauthier, PI.

References

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. *Proceedings of the IJCCN San Diego*, pp. II-411-416.

Dailey, M.N. and Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, 12(7-8):1053-1074.

Dailey, M.N., Cottrell, G. W., Padgett, Curtis, and Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *J. Cog. Neuro.* 14(8):1158-1173.

De Renzi, E., Perani, D., Carlesimo, G., Slveri, M., and Fazio, F. (1994). Prosopagnosia can be associated with damage confined to the right hemisphere – An MRI and PET study and a review of the literature. *Psychologia,* 32(8):893-902.

Farah, M. J., Levinson, K. L., and Klein, K. L. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia,* 33(6):661-674.

Gauthier, I., Anderson, A. W., Tarr, M. J., Skudlarski, P., and Gore, J. C. (1997). Levels of categorization in visual recognition studied with functional MRI. *Current Biology,* 7:645-651.

Gauthier, I., Behrmann, M., Tarr, M. J., (1999a). Can face recognition really be dissociated from recognition? *Journal of Cognitive Neuroscience,* 11:349-370.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999b). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience,* 2(6):568-573.

Joyce, C. A., Cottrell, G. W. (2004). Solving the visual expertise mystery. To appear in *Proceedings of the Neural Computation and Psychology Workshop 8*.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience,* 3(8):759-762

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17:4302-4311.

Moscovitch, M., Winocur, G., and Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with vusual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience,* 9(5):555-604.

Sugimoto, M., and Cottrell, G. W., (2001) Visual Expertise is a General Skill. *Proceedings of the 23rd Annual Cognitive Science Conference*, pp. 994-999.

A Stochastic Comparison-Grouping Model of Multialternative Choice: Explaining Decoy Effects

Takashi Tsuzuki (tsuzuki@rikkyo.ac.jp)
College of Social Relations, Rikkyo University, Nishi-Ikebukuro, Toshima-ku, Tokyo 171-8501, Japan

Frank Y. Guo (fyguo@ucla.edu)
UCLA, Department of Psychology, 405 Hilgard Ave.
Los Angeles, CA 90095-1563, USA

Abstract

Based on Guo and Holyoak's (2002a, 2002b) work, we propose a stochastic comparison-grouping theory of multialternative decision making to explain three context-induced violations of rational choice. The attraction effect and the similarity effect are explained by stochastic comparison grouping, according to which similar alternatives are compared more frequently than dissimilar alternatives are. The compromise effect is explained by the assumption that attribute values are perceived according to a basic psychophysical function, in addition to the comparison grouping mechanism. Furthermore, this model explains individual differences in choice by assuming interpersonal differences in pre-existing attitude toward products.

Introduction

Rational theories of decision making suggest that choice is intrinsically determined by the utilities of individual alternatives, thereby unaffected by relationship among alternatives, which is a part of choice context. However, violations of this tenet have been found in many studies (e. g., Huber, Payne, & Puto, 1982; Simonson, 1989). Three much-studied findings of the so-called context-dependent choice warrants specific attention, as they constitute violations of axioms that are believed to be fundamental to rational choice. They are addressed together in this paper as they share important commonalities and can be explained by a unified framework. These findings include the attraction effect, the similarity effect, and the compromise effect (Huber, Payne, & Puto, 1982; Tversky, 1972; Simonson, 1989; Simonson & Tversky, 1992).

These effects all occur with the addition of a third alternative, called the decoy, to a two-alternative choice set and are all called *decoy effects*. Like in most research of the same line (e. g., Guo & Holyoak, 2002b; Roe, Busemeyer, & Townsend, 2001), they are examined in the present paper in a two-attribute form, which is schematized in Figure 1. The alternatives that constitute the core set are commonly referred to as the target and the competitor (also called the core alternatives in this paper), and the addition the decoy. The target and the competitor form a trade-off, that is, one is better than the other on one attribute but worse than the other on the other attribute. Depending on the position of the decoy relative to that of the target, three phenomena could occur. Two of them happen when the decoy is more similar to the target than to the competitor. If it is inferior to the target on all attributes, the choice probability of the target would increase relative to that of the competitor. This is called the attraction effect (Huber, Payne, & Puto, 1982). On the other hand, if a trade-off exists between the decoy and the target, the choice probability of the target would decrease relative to that of the competitor. This is called the similarity effect (Tversky, 1972). The third phenomenon occurs when the decoy sits between the target and the competitor, in which case the decoy, now constituting a compromise of the core alternatives, would be chosen most often. This is called the compromise effect. All three phenomena would potentially lead to violations of axioms of rational choice (will explain in detail later).

A number of explanations have been advanced for each of the three findings (e. g., Simonson & Tversky, 1992; Tversky, 1972; Tversky & Simonson, 1993), however, Roe et al. (2001) were the first to explain all three (in addition to other findings) with a single framework, implemented in a connectionist model derived from a previous stochastic mathematical theory (Busemeyer & Townsend, 1993). Their model accounts for these findings by variable lateral inhibition determined by similarity relations among alternatives and momentary shifting of attention from attribute to attribute. Subsequently, Guo and Holyoak (2002b) proposed a connectionist model accounting for the attraction effect and the similarity effect that is also based on inter-alternative similarity. They conceived the decision process as divided into two stages: the two more similar alternatives (i. e., the target and the decoy) are compared first, and joined by the competitor later. The first stage has an impact on the second stage and finally leads to these phenomena (will explain in detail later). The two-stage model derives its idea from perceptual grouping, according to which similar shapes are visually perceived as forming a unit. Analogously, similar alternatives are processed together at the early stage of decision process. In analogy to perceptual grouping, this mechanism is called *comparison grouping* in the present paper. Compared to Roe et al.'s model, the two-stage assumption is more consistent with some empirical studies that investigate decision processes of multialternative choice (Russo & Rosen, 1975; Satomura, Nakamura, & Sato, 1997). To explain the compromise

effect, Guo and Holyoak (2002a) used another feature of the same model in addition to the two-stage assumption.

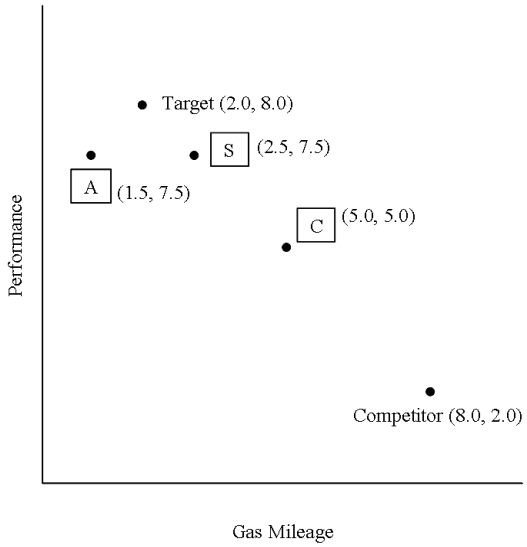


Figure 1: A summary of the phenomena simulated. The letters S, A, and C stand for the decoys for the similarity effect, the attraction effect, and the compromise effect respectively. The numbers in parentheses are attribute ratings.

Despite its explanatory simplicity and consistency with certain experimental data, the two-stage model seems oversimplified for describing human behavior – it is unlikely that people completely limit evaluation to just one pair of alternatives for a long period of time. Studies have shown that in multialternative choice tasks that resemble those giving rise to the three effects, people 1) momentarily shift attention across pairwise comparisons, and 2) similar pairs were compared *more frequently* than dissimilar pairs were (Russo & Rosen, 1975; Satomura et al., 1997). In addition, the second stage in that model was proposed to consist of triple-wise comparisons, whereas these studies suggested that choice predominantly consists of pairwise comparisons. Based on data from these studies, we propose a stochastic comparison-grouping model, in which all possible types of comparisons are performed momentarily with differential frequencies (Russo & Rosen, 1975; Satomura, Nakamura, & Sato, 1997). In addition, whereas Guo and Holyoak's model estimates choice probabilities from results of just one simulation by a mathematical conversion (Luce, 1959), the present model runs a large number of simulations to reflect decisions across individuals, thereby directly estimating choice probabilities. The psychophysical assumption (Guo & Holyoak, 2002a), proposed in conjunction with comparison grouping to explain the compromise effect, remains unchanged in the current model.

The Model

Decision Scenario and Model Architecture

The decision scenario used for simulation is adapted from that used by Roe et al. (2001). The decision maker has to choose one car from a set of two or three alternatives by evaluating two attributes; gas mileage and performance, which are measured on a 1 – 10 scale (see Figure 1). Accordingly, a connectionist model is constructed (see Figure 2). Each attribute or alternative is represented by one node (circle) in the network, with their relations represented by connections (lines with arrowheads). Each node has a certain degree of activation. For an alternative node, the activation stands for the valuation of the corresponding alternative; for an attribute node, it stands for the evaluative importance of that attribute. Node activations are within the range of 0.0 - 1.0.

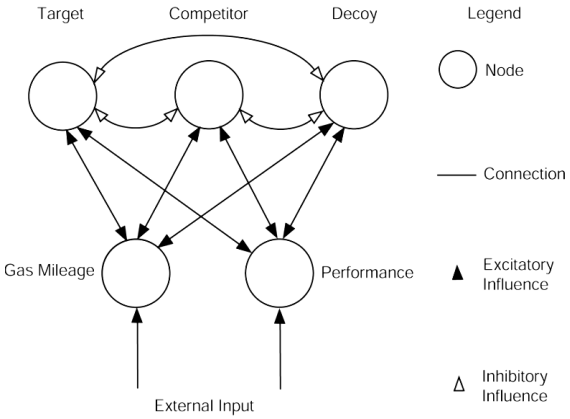


Figure 2: The architecture of the model. External Input represents the motivational and attentional sources that drive the decision process.

The connection between an attribute node and an alternative node, called the *attribute-alternative connection*, has an excitatory weight (i. e., when one node is more active, the other would be more active as well). The connection between each pair of alternative nodes has an inhibitory weight (i. e., when one node is more active, the other would be less active), also known as *lateral inhibition* in the literature. The lateral inhibition reflects the competitive relationship among alternatives. Via this mechanism, which would commonly result in one node achieving higher activation than the rest, the model "chooses" the winning alternative. All connections are bi-directional, reflecting the idea that influences can go either way between factors involved in decisions. The external inputs to the attribute nodes represent the motivational and attentional sources that drive the decision making.

This network representation was similar to Guo & Holyoak's (2002a, 2002b) model, and is consistent with common connectionist architecture used in decision modeling (e. g., Holyoak & Simon, 1999; Tsuzuki, Kawahara, & Kusumi, 2002).

Connection Weights and Initial Activations

The attribute-alternative weights reflect the perceived goodness of alternatives according to their attribute ratings, and were initially set to the corresponding attribute-alternative ratings. For example, the performance-target and gas mileage-target weights were first set to 8.0 and 2.0 respectively.

Recall that in Guo and Holyoak's model (2002a, 2002b) the perception of goodness follows a basic psychophysical function. In particular, this function reflects the idea that perceived goodness increases with negative acceleration with actual attribute value. Consistent with this idea, each attribute-alternative weight was further transformed by a logarithmic function:

$$w_{ij} = (\log_e (w_{ij} + \alpha) + \beta) / \gamma . \qquad (1)$$

Here, w_{ij} is the weight of the connection from node j to node i. α, β, and γ are set to 31.00, –3.35, and 0.905 respectively. Equation 1 reflects a weakly convex function. In addition, it serves as a normalization function that compresses these weights to values within a small range, which is comparable in magnitude to node activations (whereas the attribute values range from 0.0 to 10.0, the w_{ij}s range from 0.090 to 0.400.)

The lateral inhibitions are all set to -0.60. The initial activations of all nodes are conveniently set to 0.5, the middle point of the activation range, with the following qualification.

In reality, people usually have different pre-existing preferences regarding products. Accordingly, randomness was introduced to the initial activations of the alternative nodes, which were in the range of 0.5±0.25. The values follow a uniform distribution. As will be seen later, this randomness provides an explanation for individual differences in choice.

Running the Model

Connectionist models commonly run in an iterative fashion. In each iteration the activation of each node is updated according to the total influence it receives from the rest of the model – the activation increases if the influence is positive and decreases if otherwise. This influence can be understood as the overall reason for liking or disliking an alternative or attribute. A common activation function is used to specify this process (c. f., McClelland & Rumelhart, 1988).

$$a_i(t+1) = a_i(t) + \Delta a_i(t), \text{ where} \qquad (2)$$
$$\text{if } netinput_i > 0,$$
$$\Delta a_i = netinput_i (MAX - a_i) - decay . a_i$$
$$\text{otherwise}$$
$$\Delta a_i = netinput_i (a_i - MIN) - decay . a_i$$

$a_i(t+1)$ is the activation of node i at iteration (or time) $t + 1$; it is a function of $a_i(t)$, the activation of the same node at the previous iteration (or moment). $\Delta a_i(t)$ is the amount of activation change. The decay parameter reflects how much a neural signal decays over time (connectionist models draw analogies to neural processing), and is set to 0.04. The decay, however, does not play an important role in explaining the effects. *MAX* and *MIN* are the upper (1.0) and lower (0.0) limits of node activations. This equation specifies that node activation asymptotically approaches the upper or lower activation limit as a consequence of the total influence it receives from other components of the network. The total influence, $netinput_i$, is determined by the following equation.

$$netinput_i = istr . intinput_i + estr . extinput_i, \text{ where} \qquad (3)$$
$$intinput_i = \Sigma_j w_{ij} a_j(t)$$
$$extinput_i = 1$$

Intinput is the internal input that comes from all the attribute and alternative nodes, and depends on both the activation of the node feeding input to i and the connection strength that links the two nodes. *Extinput*, standing for the external input, should be a stable source of attention and motivation and is set to a constant. *Istr* and *estr*, set to 0.12 and 0.05 respectively, are constants that scale down activation changes so that the changes are not abrupt. Since the internal input is the source of these effects, *istr* is set to be larger than *estr*.

The model runs iteratively – in each iteration, the activation of each node in the model is updated according to Equation 2 and 3. The process reflects the evolution of valuation over time. This iterative process continues until an externally determined period of deliberation time, arbitrarily set to 100 iterations, is met.[1] The final winning choice is the alternative with the highest activation.

Stochastic Comparison Grouping

In a series of eye fixation studies, Russo & Rosen (1975) found that pairwise comparisons between similar options happen earlier and more frequently than other types of comparisons in multialternative choice. Consistent with that, Satomura et al. (1997) found that in decision tasks leading to the attraction effect, for the participants who chose the target (i. e., exhibited the attraction effect), 74% retrospectively reported that they compared the target to the decoy, while only 19% of them reported that they compared the target to the competitor. These studies give rise to the following modeling assumptions – All four kinds of possible comparisons (target-decoy, competitor-decoy, target-competitor, and target-decoy-competitor) are performed momentarily with different frequencies. To instantiate this in the model, for each iteration, a specific type of comparison is randomly chosen according to specified probabilities, and only the involved alternative nodes are updated.

[1] Other stopping criteria might be used. For example, the model can stop when the amount of activation difference across nodes is very large, or the amount of activation change becomes very small. However, according to our analysis, the type of criterion does not affect the qualitative pattern of the simulation results.

In simulating the attraction effect, the frequencies of these types of comparisons, in the order mentioned above, were set to the percentages of 74.0%, 10.4%, 10.4%, and 5.2% respectively. For example, for a random 74.0% of the iterations, only the activations of the target and the decoy were updated. These percentages were arbitrarily determined to roughly reflect previous experimental data (Russo & Rosen, 1975; Satomura et al, 1997), which suggested pairwise comparisons constitute the majority of the deliberation process, and the two similar alternatives were compared most often. To be consistent, the same decision process was employed for the similarity effect.

Simulations and Results

A total of 10,000 independent simulations, each standing for the deliberation of one individual, have been performed. For each simulation, the alternative with the highest final activation is the one chosen. Choice probability was obtained across all the simulations. The simulation results are presented both as choice probability (Table 1) and average node activation (Table 2). Note that choice probability is the criterion by which the modeling is judged.

Modeling Individual Differences

Note that node activation evolves as a continuous function of time. This means a node with high initial value tends to stay strong. For instance, if the initial value of the target is higher than that of the competitor, the target would tend to maintain relative advantage over the competitor in deliberation. Recall that initial activation values randomly vary across simulations. This randomness therefore leads to choice differences across simulations, and explains why sometimes the unlikely alternative was chosen. For example, the decoy was chosen with a slim chance in the attraction effect scenario – when the initial value of the decoy was very high relative to the target and the competitor, such initial advantage was too strong to be offset by the comparison-grouping mechanism. The modeling is consistent with the intuition that people have different pre-existing beliefs that randomly favor one alternative over another and tend to bias later decisions.

Binary Choice

The target and the competitor are set to equal additive attribute ratings: the target is rated 2.0 on gas mileage and 8.0 on performance, whereas the competitor is rated 8.0 on gas mileage and 2.0 on performance. The two attributes are assumed to be equally important. Consistent with the trivial prediction that their choice probabilities should be the same, the simulations yielded probabilities of 0.504 and 0.496 for them respectively. The slight inequality was due to the randomness in initial activations of the alternative nodes.

Attraction Effect

When the attraction effect occurs, the target benefits from the addition of the decoy more than does the competitor. Under some circumstances, this tendency leads to a higher choice probability for the target in the trinary set than in the core set. This constitutes a violation of the regularity principle of rational choice, according to which adding alternatives to a given choice set should not increase the probability of any alternative (Huber et al., 1982). In the simulation, the decoy was chosen to have attribute values of 1.5 and 7.5 for gas mileage and performance, respectively.

Comparison grouping, in conjunction with the competitive relationship among the alternatives, is able to explain this effect. Any time when the target is compared with the obviously inferior decoy (in modeling terms, this means the target node receives more input via the attribute-alternative connections than does the competitor), the activation of the target node increases whereas the activation the decoy node decreases. This differentiation is an intrinsic property of this type of connectionist model (c. f., McClelland & Rumelhart, 1988). Given that the deliberation process primarily consists of target-decoy comparisons, the target node would finally acquire higher activation than does the competitor.

The above analysis suggests that if the initial node activations were identical across the alternatives, the target would have been chosen for all simulations. However, with some randomness, it is possible that the competitor has a higher initial activation than does the target. If this initial difference, which has an impact on later comparisons, is large enough, the competitor would be chosen. This also suggests that with extreme initial values even the rather inferior decoy might be chosen. This seems consistent with the intuition that pre-existing beliefs regarding products carry a weight on later decisions.

The simulated choice probabilities of the target, the competitor, and the decoy were 58.7%, 36.6%, 4.8%. The probability of the target exceeds that in the binary choice scenario, thereby leading to a violation of the regularity principle.

Similarity Effect

In the decision situation that leads to the similarity effect, the target looks less attractive relative to the competitor once the decoy is introduced. Under certain situations, this would lead to a change of rank order of the target and the competitor. For example, in the simulated scenario, the core alternatives rank the same in the binary set, but the competitor would rank higher than the target if the similarity effect occurs. This constitutes a violation of the independence of irrelevant alternatives principle of rational choice, which states that adding a decoy to an original choice set should not alter the rank order of the alternatives (c. f., Tversky, 1972). Decoy was set to have attribute ratings of 2.5 and 7.5 for gas mileage and performance respectively. Note that its additive attribute rating is identical to that of the target.

Like in the case of the attraction effect, comparison grouping and inter-alternative competition are able to explain the similarity effect. Any time when the target is compared with the similarly attractive decoy, the activation

of both the target and the decoy nodes decrease due to their mutual inhibition of equal strength. This again is an intrinsic property of this type of connectionist model (c. f., McClelland & Rumelhart, 1988). Because the target-decoy comparison is the predominant type of comparison, compared to the competitor node, the target node hurts more from the comparison with the decoy, and would finally acquire lower activation than does the competitor.

The simulated choice probabilities of the target, the competitor, and the decoy were 27.8%, 39.7%, and 32.6%. Note that the tie between the target and the competitor was broken, indicating a violation of the independence of irrelevant alternatives principle.

In summary, the similarity effect and the attraction effect can be explained by frequency difference between the target-decoy and competitor-decoy comparisons. This also suggests that simulations of these effects should not depend on a particular specification of frequency ratio for the four types of comparisons – so long as this frequency difference is substantial, the two effects should be observed. In fact, other frequency ratios were used in our simulations and the same pattern was obtained (not reported here due to space limit).

Table 1: Simulation results as choice probability (estimated from 10,000 simulations).

Choice scenarios	Choice probability		
	Target	Competitor	Decoy
Binary choice	0.504	0.496	-----
Attraction effect	0.587	0.366	0.048
Similarity effect	0.278	0.397	0.326
Compromise effect	0.213	0.219	0.568

Table 2: Simulation results as average node activation and *SD*.

Choice scenarios	Average Node Activation (*SD*)		
	Target	Competitor	Decoy
Binary choice	0.293 (0.060)	0.294 (0.060)	-----
Attraction effect	0.320 (0.034)	0.305 (0.046)	0.233 (0.038)
Similarity effect	0.286 (0.033)	0.300 (0.046)	0.291 (0.033)
Compromise effect	0.275 (0.021)	0.276 (0.021)	0.291 (0.022)

Note. The results are computed from the simulations summarized in Table 1.

Compromise Effect

When the decoy for the similarity effect moves toward the competitor and finally reaches the middle point between the target and the competitor, the similarity effect turns into the compromise effect – The decoy changes from the least popular to the most popular alternative. In a decision scenario slightly different from the present one, this effect can also lead to a violation of the regularity principle (see Simonson, 1989 for more detail).

The comparison grouping assumption suggests that frequency of pairwise comparison increases with inter-alternative similarity. Accordingly, the percentages of the target-decoy, competitor-decoy, and target-competitor comparisons have the ratio of 2 : 2 : 1, inversely proportional to psychological distance[1]. The triple-wise comparison, being the least frequent, was arbitrarily set to one half as frequent as the least frequent pairwise comparison. Hence, the percentages of the four types of comparisons were set to 36.36%, 36.36%, 18.18%, and 9.10%.

The psychophysical assumption implemented in Equation 1 gives rise to this phenomenon. (This mechanism was still at work in the simulations of other two phenomena, but it did not play a causal role in producing them.) Take the target-decoy comparison for an example. The advantage of the decoy over the target (ratings of 5 versus 2 on gas mileage) looms larger than the advantage of the target over the decoy (ratings 8 versus 5 on performance) after the attribute ratings have been transformed into connection weights. (Calculated by Equation 1, the sum of the two attribute-alternative weights is 0.512 for the decoy, higher than the 0.505 for the core alternatives.) Hence the total input the decoy node receives via the attribute-alternative connections is the largest among the alternative nodes, making the decoy the winner.

The simulated choice probabilities of the target, the competitor, and the decoy were 21.3%, 21.9%, and 56.8%. Note that the specification of comparison percentages is not unique – so long as there is no frequency difference between the target-decoy and the competitor-decoy comparisons, neither the target nor the competitor would be bolstered relative to the other. The psychophysical mechanism would then guarantee choosing the decoy.

Comparison grouping provides a unified framework toward understanding the three phenomena. In particular, it explains why difference between the core alternatives exists in the similarity effect but disappears in the compromise effect, as comparison grouping can be modified by changing the similarity between the decoy and the core alternatives.

[1] This is just one way of specifying the inverse relationship between frequency ratio and similarity, which should be viewed as qualitative rather than quantitative. Note that in simulations of the other two effects, frequency ratios are not determined by the same function and just roughly reflect the inverse relationship.

Conclusion

We propose a stochastic comparison-grouping theory cast in a connectionist model to explain three important violations of rational choice. In addition, this model lends us understanding of the decision processes involved in these tasks.

A comparison is made between this model and previous accounts of the same findings. It extends Guo & Holyoak's (2002a, 2002b) model by incorporating insights from experimental data (Russo & Rosen, 1975; Satomura et al., 1997). In addition, it better accounts for individual differences in choice by introducing randomness in intial beliefs to the model. In comparison with Roe et al.'s (2001) model, both models use similarity relationship, but in different manners. Their model uses variable lateral inhibition that increases with inter-alternative similarity, whereas the current model proposes a similarity-based grouping mechanism. In addition, both models suggest momentarily shifted attention, again in different manners. In their model attention shifts from attribute to attribute, whereas in the present model attention shifts across different types of pairwise comparisons. The assumptions of this model seem more consistent with the aforementioned experimental data. Future studies are in order to further test the relative merits of the two models.

One apparent problem of the present model is that the modeling seems to depend on manually specified parameter values rather than psychological principles. Our justification is that these parameters specify linear transformations that do not alter the essence of the modeling assumptions. In addition, the same set of parameter values applies to all three phenomena.

Finally, this model is consistent with theoretical frameworks that relate cognition to perceptual processes (c. f., Medin, Goldstone, & Markman, 1995, Goldstone & Barsalou, 1998), and its proposed perceptual mechanisms might help us understand decision behavior at large.

Acknowledgments

We are grateful for comments from Keith J. Holyoak and John Hummel.

References

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*, 432-459.

Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition, 65*, 231-262.

Guo, F. Y., & Holyoak, K. J. (2002a, April). *Coincidence effect versus compromise effect: A neural network explanation.* Paper presented at meeting of the Western Psychological Association, Irvine, CA.

Guo, F. Y., & Holyoak, K. J. (2002b). Understanding similarity in choice behavior: A connectionist model. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 393-398). Mahwah, NJ: Lawrence Erlbaum Associates.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*, 3-31.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*, 90-98.

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review, 2*, 1-19.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge, MA: MIT Press.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, *108*, 370-392.

Russo, J. E., & Rosen, L. D. (1975). An eye fixation analysis of multialternative choice. *Memory & Cognition*, *3*, 267-276.

Satomura, T., Nakamura, H., & Sato, E. (1997). Consumers' attitude toward price (4): Experiments of reference price. *Distribution Information*, *5*, 18-24. (In Japanese)

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects, *Journal of Consumer Research*, *16*, 158-174.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*,*29*, 281-295.

Tsuzuki, T., Kawahara, T., & Kusumi, T. (2002). Connectionist modeling of higher-level cognitive processes. *Japanese Journal of Psychology, 72*, 541-555. (In Japanese with English summary)

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*, 281-299.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science, 39*, 1179-1189.

How expert dealers make profits and reduce the risk of loss in a foreign exchange market?

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Yusuke Uchida (uchi-fcs@poppy.ocn.ne.jp)
Department of General System Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku
Tokyo, 153-8902 JAPAN

Kiyoshi Izumi (kiyoshi@ni.aist.go.jp)
Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology, 2-41-6 Aomi, Koto-ku
Tokyo, 135-0064 JAPAN

Yusuke Ito (itoy@simplexinst.com)
Simplex Institute, Inc., Shuwa Kamiyacho Building, 6th Floor, 4-3-13, Toranomon, Minato-ku
Tokyo, 105-0001 JAPAN

Abstract

The paper clarified how actual expert dealers made profits and reduced the risk of loss in a virtual foreign exchange market, by comparing the way novice or general investors showed. As a result, we could find that the experts were risk-averse for losses and risk-seeking for profits while the novices showed the opposite behavioral tendency. This result was analyzed in terms of the prospect theory: It was found that the expert could be (partially) free from deposition effect and sunk-cost effect while the novices were not. It was also suggested that the experts' strategy of dumping losing currencies could be socially transmitted by the dealers' managers.

Introduction

How people make decision and learn in dispersed complex social systems has recently been considered important to be investigated. Here a dispersed complex social system is defined as a system without a centralized information resource and mechanism for deciding its behavior and, at the same time, with rapidly changing environments which provide their decision makers with information that continually changes. In this paper, we will focus on the (spot) foreign exchange market as a typical example of the complex social systems. This is because (1) the foreign exchange market have, as their constituents, a lot of market players (such as banks, investors, governments, and so on) who are different from one another in investment motivation, speculation, and stake etc., (2) it is not a centralized market, as the stock market, so that its market players have to fathom out market consensus based on limited information, and (3) it has rapidly changing environments which provide its decision makers with information on the order of seconds – the rate of change is typically faster than the decision makers can respond. So it is considered a good target of cognitive research to clarify how expert or skilled dealers make decisions, especially make profits and reduce the risk of loss, in the exchange market.

How can we investigate the process of expert dealers' decision making? For this purpose, experimental market (or economics) approach (for instance, Friedman & Sunder, 1994) seems useful: Experimental market is the field that attempts to understand the behaviors of the markets and their players by conducting experiments using real human subjects. From a cognitive perspective, the experimental market approach can be used to clarify the way of human biased decision making, by analyzing how market players actually make decisions in their conducting transactions on a virtual financial market that simulates the corresponding real market (Smith, 1996; Ueda, Taniguchi & Nakajima 2003). Namely, from the result of such a cognitive experiment on the markets, it is considered to understand how people or market players make decisions, in the real markets, under some biases such as representativeness (Kahneman & Tversky, 1972). This paper also adopts this **cognitive experimental market approach,** as its research methodology, in order to clarify what kind of biases influence expert dealers' decision making (profit making and risk hedge) in a virtual exchange market. We will pursue this research target, specifically by comparing the performance of expert or skilled dealers with that of novice or general investors.

There are some previous studies using the cognitive experimental market approach. For example, the U-Mart project aims to analyze the behaviors of the market, especially the conditions under which bumpy ride occurs, using a virtual future stock market (Kita, Sato, Mori & Ono 2003; Ueda, Taniguchi & Nakajima 2003). Because human players, as well as computer agents, participated in a series of experiments, it is possible to analyze the relation between the macroscopic behaviors, such as price fluctuation, of the market and the microscopic features of human players, which has not been realized yet.

Smith (1996 & 1997) analyzed, using three trained dealers as his subjects, how they reduced risks in dealing on a virtual spot currency market and proposed a feedback-control model of their risk management. Because he did not

compare the performance of his skilled subjects with that of novice or general investors, he did not clarify what kind of biases the skilled dealers had or were free from. In addition, it is a question that his subjects were actually expert dealers. In the experimental environment used in (Smith 1996; Smith 1997), each skilled subject was asked to make deals with a pre-installed computer dealer while, in that used in (Izumi, Nakamura & Ueda 2002), subjects were asked, in pairs, to make deals with one another through computer network, which virtually realized human-human dealing. In this sense, the latter environment can be said to be closer to the real dealing situation than the former. However, it is also a question that all of their subjects were actually expert dealers.

So, in this paper, we will ask real expert dealers to participate in our experiments and compare their performance with novice or general investors' one. Here, "expert" dealers mean those who had engaged in currency dealing or stock trading for more than five years, since this business field is a competitive jungle so that "engaging for five years" can be an index of being expert or skilled. On the other hand, we will ask graduate students majoring in economics to participate, as novice or general investors, in our experiments. Under this circumstance, we made two experiments in order to clarify how and under what kind of biases the expert dealers make profits and reduce the risk of loss in a virtual foreign exchange market. In this virtual market, only dollars and yen will be dealt.

This paper will be constructed as follows: In the second section, necessary terminology will be introduced. In the third and fourth section, the method of two experiments, which were made to clarify dealers' decision making, and their results will be explained. In the fifth section, the results obtained will be discussed from the perspective of the prospect theory (Kahneman & Tversky, 1979). In the final section, this paper will be concluded.

Terminology

Before we will explain our experiments, technical terms necessary for understanding dealers' trading behavior should be introduced.

Position: We will use this term from the point of the amount of dollars that each dealer has. "Long" means owning or holding dollars (i.e., the amount of dollars bought exceeds that of dollars sold). "Short" is the opposite of a long position. "Square" means the situation that the amount of dollars bought is equal to that of dollars sold. In our experiments, all the subjects were asked to start from the square position and to go back to the square at the end of the experiments.

Unrealized profits and losses (UPL): An increase/decrease in the value of dollars that is not "real" or "unrealized" because the dollars have not been sold. Once dollars are sold by a dealer, the profits/losses are "realized" by the dealer. If a dealer started from the square and bought one dollar at the rate of $1=Y100 and, after that, the rate has changed to $1=Y110, the dealer has Y10 as UPL, which will be realized when he/she sells the dollar to get back to the square. In our experiments, the performance of each dealer will be estimated in terms of UPL.

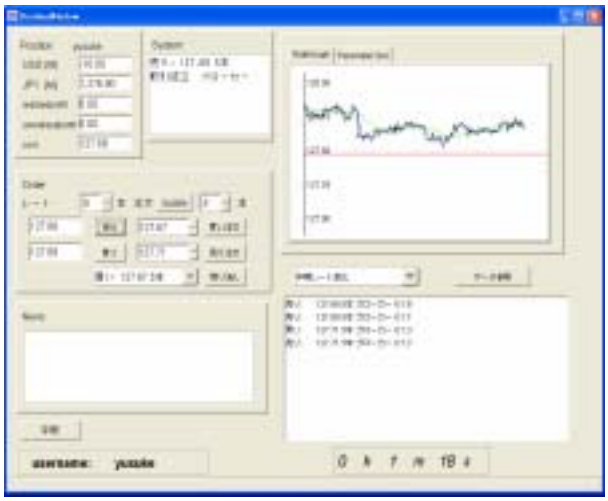


Figure 1: The user display of VDS (stand-alone type).

Lengthening vs. liquidation, profit-taking vs. loss cut: In this paper, such a dealing that the absolute amount of position increases is called "lengthening" while such a dealing that the absolute amount of position decreases is called "liquidation". Because it is related to dealers' risk management, the latter will be analyzed in detail. Moreover, liquidation can be divided into two sub-categories: profit-taking and loss cut.

Experimental Environment

The virtual dealing system (for abbrev., VDS; see Figure 1)[1], which we originally developed using Java language, was used in our experiments. The VDS was constructed so that users (subjects) could make dealings with one another through computer network; it simulates the functions and display of the actual dealing systems, such as Reuter 2000, so that users can get various fundamental information, such as interest rates and balance of trade, news and trends to buy and sell dollar/yen with other users or a broker[2].

The VDS is available both as a server-client system, in which multiple users make dealing with one another, all at once, through computer network, as is actual dealings, and as a stand-alone system, in which only one user makes dealing with the system's broker, as was in the experiment by Smith (1996 & 1997). In Experiment 1, it was used as the server-client system while, in Experiment 2, it was as the stand-alone system.

The VDS is designed so that we can get various users' dealing logs: For example, logs about what type and amount of dealing a user made at which time, and those about what type of information, news or trends a user referred to in his/her dealing. Therefore, by using the VDS, we can

[1] Because it was so built that the Japanese dealers would use, this VDS has Japanese signage, as denoted in Figure 1.

[2] Only one broker is assumed to exist in this VDS.

analyze what type of decision a user made referring to what kind of information.

Experiment 1

Purpose

The purpose was to make a hypothesis about how expert dealers made profits and reduced the risk of loss in a virtual foreign exchange market, by comparing the performance of the expert dealers with that of novice or general investors. The target was a dollar-yen exchange market, as already explained. Because this experiment aimed to explore a hypothesis about the way experts made decision, the server-client system was used so that it could provide the subjects with a dealing environment similar to the actual one.

Subjects

Eleven dealers, who had engaged in actual dealing or trading for more than five years, participated as "expert" dealers (we call this group of subjects **expert group**) while ten graduate students, who majored in economics, participated as "novice" or "general" investors (we call this group of subjects **novice group**).

Procedure

All the subjects of each group were gathered together, at a time, into one meeting room; the experiment of the expert group and that of the novice one were made independently.

The news and fundamental information given to the two groups were the same actual data during August, 1997; in the first half of this period, the rate was gradually moving up while, in the latter half, it rapidly and sharply declined, which was caused both by a decline in U.S. stock prices and by Japan's current account surplus. The rate change during the experiment was not calculated endogenously, i.e. as a result of dealings in the VDS, but given exogenously, i.e. the same as the actual change during this period: This was because the rate would have fluctuated quickly if the rate change had been endogenously calculated in such a dealing environment only with small number of market players. All the subjects were, in fact, given such an instruction that exchange rates were calculated by the orders of participating players, since the dealing environment of the experiment needed to be as close to the actual one as possible.

The subjects of each group were first explained about how they could use the VDS interface and make dealings through the VDS for 30 minutes. They were then given the information about the economic situation and fundamentals just before August, 1997[3]. After that, they were asked to make dealings through the VDS for 20 minutes[4].

[3] Because all the names of the currencies dealt, the target nations, and the proper names that came on were renamed, all the subjects did not notice that the rate and data given were the actual ones in the past. Of course, during the experiment, news was, from time to time, given to the subjects while the data of economic fundamentals and the trends of rate were always available.

[4] The events occurred in August, 1997 was compressed in the time frame of 20 minutes.

The data collected were the logs of each subject's positions, orders, referred economic information and messages of chatting, through the VDS, with other subjects.

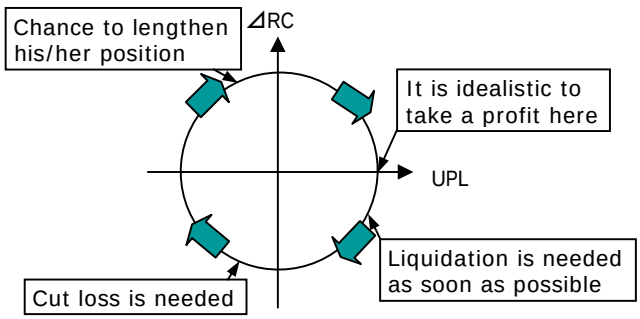


Figure 2: The risk space.

Method of analysis

The main purpose of the experiment was to make a hypothesis about how the respective groups made profits and reduced the risk of loss. So, for each subject, we plotted every log of dealing on the risk space (Smith, 1996): The risk space[5], which has an x-axis that denotes the value of UPL[6] for each dealing and has a y-axis that denotes the rate-of-change (RC) of exchange rates whose sign shows the increase in the absolute value of UPL[7], visualizes how each subject managed the risks caused by taking his/her positions (see Figure 2). By comparing the risk spaces of the expert group with those of the novice one, we can analyze how the expert dealers made decisions, especially reduced the risk of loss.

Result

Typical examples of the risk spaces plotted for the expert and the novice group are shown respectively in Figure 3 and 4; in these figures, a circle denotes buying dollars, a upside-down triangle does selling dollars, blue color does lengthening and red color does liquidation. From the nature of the figures, the lines of the graphs will extend towards high positive values of the x-axis if a subject waits to lock in profits after enough profits are made; in the same way, the lines will extend towards high negative value of the x-axis if he/she is so slow in cutting losses. Therefore, the higher the average profit is, the slower a subject tends to be in taking profits; the higher the average loss is, the slower he/she tends to be in cutting losses.

From these figures, we can find a tendency that the novices cut losses later than the experts. So, to statistically confirm the above, we calculated the average loss and profit respectively for the two groups in order to compare the average loss (or profit) of the expert group with that of the novice one. As a result, as for the average loss, we could

[5] The risk space used in this study was a little different from that used in (Smith 1996).

[6] The value of UPL was normalized, being divided by the amount of the maximum position of the subject.

[7] When a subject's position is long, $\Delta RC>0$ if dollar appreciation occurs while $\Delta RC<0$ if yen appreciation does.

find significant difference between the two groups (expert = -0.02854 (SD = 0.00035), novice = - 0.12315 (SD = 0.01124); p = 0.041<0.05, one-sided) while, as for the average profit, we could find no significant difference between the two groups (expert = 0.46511 (SD = 0.03293), novice = 0.37609 (SD = 0.02344); p = 0.190, one-sided).

From the above result, it is possible that experts can cut losses so fast that they may prevent the losses from increasing and, at the same time, can take profits at the right time while novices cannot: This is, however, considered to be little better than a hypothesis. It is because, as for the average profit, the number of dealings was so scarce that we could find no significant difference and because this result may be specific to a set of the exchange rate and news given to the subjects (hereafter, we call this set "scenario"). We therefore need to make an additional experiment to confirm the hypothesis obtained, with augmenting the number of dealings: The reason the number was small is considered to be attributed to the experimental environment of face-to-face dealing because almost all the expert subjects were acquainted with each other so that they, for a while, hesitated to make dealings. So we will make an experiment, as Experiment 2, by using a stand-alone type of VDS.

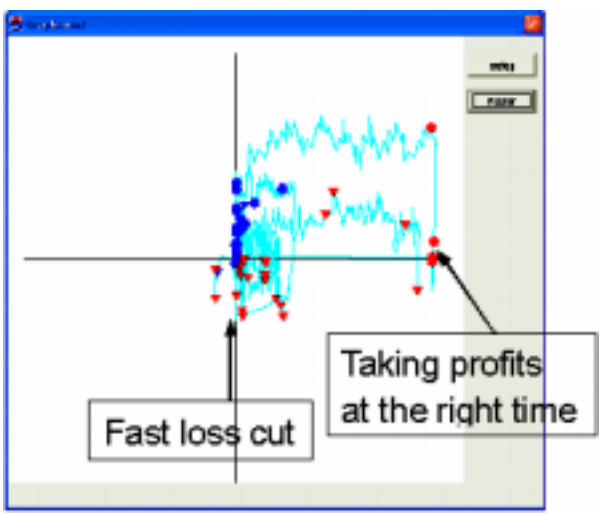


Figure 3: An example of the risk space of expert.

Figure 4: An example of the risk space of novice.

Experiment 2

Purpose

The purpose was to confirm the hypothesis derived from the result of Experiment 1: The hypothesis was that expert dealers could cut losses so fast that they might prevent the losses from increasing and, at the same time, could take profits at the right time while novices could not. Because of the reasons explained above, a stand-alone type of VDS was used in this experiment.

Subjects

Ten dealers, who had experienced more than five years' dealing, participated as "expert" dealers (**expert group**) while ten graduate students, who majored in economics, participated as "novice" investors (**novice group**).

Procedure

In this experiment, a stand-alone type of VDS was used so that all the subjects needed not to be gathered together at a time; each subject participated in this experiment independently. They were asked to make dealings with a computer dealer (computer-driven trading program) of the VDS.

The scenario (news and fundamental information) given to the two groups were the same actual data during September, 2000; through this period, the exchange rate showed a box or range rate, i.e. the rate fluctuated within the expected range. The rate change during the experiment was given exogenously, as was the same in Experiment 1[8].

All the subjects were first explained about how they could use the VDS interface and make dealings through the VDS for 40 minutes. They were then given the information about the economic situation and fundamentals just before September, 2000. After that, they were asked to make dealings through the VDS for 50 minutes.

The procedures other than the above were the same as those in Experiment 1.

Method of analysis

The same as that in Experiment 1.

Result

We calculated the average loss and profit respectively for the two groups in order to compare the average loss (or profit) of the expert group with that of the novice one. As a result, we could find significant difference between the two groups, both as for the average loss (expert = -0.01849 (SD = 0.00010), novice = - 0.031 (SD = 0.00005); p = 0.029 < 0.05, one-sided) and as for the average profit (expert = 0.09915 (SD = 0.00086), novice = 0.03619 (SD = 0.00013); p = 0.003 < 0.01, one-sided).

So our hypothesis was confirmed: We can say that the experts could cut losses so fast that they might prevent the losses from increasing and, at the same time, could take

[8] All the subjects were, in advance, informed of it.

profits (or lock in profits) after enough profits were made, whereas the novices could not.

Discussions

Analysis in terms of the prospect theory

From the results of the two experiments, we can say as follows: Expert dealers can cut losses so fast that they may prevent the losses from increasing and, at the same time, can take profits (or lock in profits) after enough profits are made. On the other hand, novices seem to be reluctant to sell currencies that lose value while they are inclined to lock in profits no sooner than the currencies they hold are profitable. Why is there a sharp contrast between the way experts make decision and that the novice do?

To answer this research question, we will introduce the prospect theory by Kahneman & Tversky (1979): The theory shows, based on the results of a laboratory experiment, that people's attitudes toward risks concerning profits may be quite different from their attitudes toward risks concerning losses. Namely the theory claims that people are, in general, risk-averse for profits and risk-seeking for losses.

By using the theory, the tendency that people are reluctant to lock in profits even when the losses increase while they are willing to do no sooner than their holdings become profitable can be explained by "deposition effect" (Shefrin & Statman 1985; Weber & Camerer 1998) and "sunk-cost effect" (Kahneman & Tversky 1979): The former means that people prefer certainty to uncertainty over a reference point while they show the opposite preference under a reference point. And the latter means that people tend to overestimate the cost needed for making a position; decision-makers are unduly influenced by resources that have already been spent and are therefore more likely to continue pursuing a previously chosen course of action.

This theoretical explanation seems to be true of the dealing behaviors of our novice subjects, because they sold promising currencies or winners too early and rode unpromising currencies or losers too long. That is, our novice subjects were considered not to be free from the biases of deposition effect and sunk-cost effect. On the other hand, our expert dealers seemed to avoid being distorted by these biases, because they could cut losses early and wait to lock in profits after enough profits were made.

Then the next question arises about whether expert dealers can be completely free from the biases. To clarify this point, we analyzed the dealing data in more detail. As a result, we could find that some expert dealers sometimes showed stepwise profit taking (see Figure 5). This stepwise way of profit-taking can be interpreted in two ways: One is that this indicates partial irrationality, which is subject to the biases, of their dealing behaviors and the other is that they were forced to take profits stepwise and to hedge risks caused by rapid price fluctuation because they could not fully predict the future exchange rate. Anyway, it can be said that even the expert dealers were not wholly free from the biases.

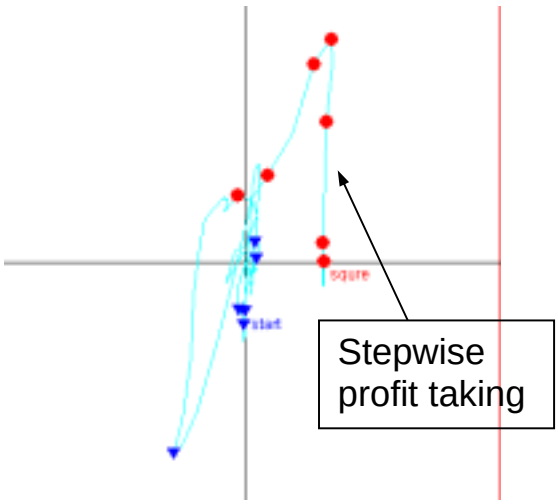


Figure 5: An example of the stepwise profit taking that some of the experts showed.

We also asked two investment managers why expert dealers can be free from the biases. Both of the managers said that dealers were taught, by their managers, to sell losers as early as possible because riding losers too long would only compound their losses, which might develop into a major management issue. On the other hand, the managers did not explicitly teach how to make profits. If this holds of dealers in general, it is possible that expert dealers have socially learned to be free from the biases. The investment managers added that a lot of dealers in the making had to drop out because it was quite difficult to learn to be free from the biases, which seemed easy on the surface.

To sum up, expert dealers can be (partially) free from the biases, in their decision making, that a lot of people are considered to have (Kahneman & Tversky 1979). This clarification, by using actual expert dealers as subjects, is the main contribution of this paper.

Conclusion

How people make decision and learn in dispersed complex social systems has recently been one of the important research issues. Especially the way of decision making needs to be investigated in real situations or settings. So, in this paper, we focused on the (spot) foreign exchange market as a typical example of the complex social systems and clarified how and under what kind of biases the expert dealers make profits and reduce the risk of loss in a virtual foreign exchange market. For this purpose, we asked real expert dealers to participate in our experiments and compared their performance with novice or general investors' one.

Our two experiments showed that the expert dealers could cut losses so fast that they might prevent the losses from increasing and, at the same time, could take profits (or lock

in profits) after enough profits were made. On the other hand, novices were reluctant to sell currencies that lost value while they were inclined to lock in profits no sooner than the currencies they held were profitable.

We analyzed this result in terms of the prospect theory (Kahneman & Tversky 1979). The prospect theory claims that people tend to sell promising currencies or winners too early and ride unpromising currencies or losers too long, under the influence of deposition effect and sunk-cost effect. This theoretical explanation was clarified to be true of the dealing behaviors of our novice subjects. On the other hand, it was also clarified that our expert dealers could avoid being distorted by the biases. In addition, it was possible that even the expert dealers were not wholly free from the biases because some of them showed stepwise way of profit-taking, which was considered not fully rational.

"Artificial market" research (Arthur, 1991) attracts attention of many researchers in recent years. In the artificial market research, computer programs as virtual market participants are built and simulated, where these computer programs mutually trade, for understanding the phenomena and features of real markets. The key is to build an artificial market with the appropriate features of the actual markets and their participants. The result of this paper is considered to be applicable to this artificial market research, especially to the construction of agents with the way of risk management in the artificial market model by Izumi & Ueda (2001). In this way, the result of this research is also of practical use.

Acknowledgments

We thank to all the subjects for kindly participating in our experiments. This research is partially supported by grant (Grants-in-Aid for Scientific Research, Scientific Research in Priority Areas 2003, No.A06-14) from Japan Society for the Promotion of Science.

References

Arthur, B. W. (1991). Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *The American Economic Review, 81*, 353-359.

Friedman, D. & Sunder, S. (1994). Experimental methods: A primer for economists. Cambridge, UK: Cambridge University Press.

Izumi, K., Nakamura, S. & Ueda, K. (2002). Identification of agents' strategy making process by an experimental market. *Proceedings of the 6th Joint Conference on Information Sciences,* 1081-1084.

Izumi, K. & Ueda, K. (2001). Phase transition in a foreign exchange market: Analysis based on an artificial market Approach. *IEEE Transaction on Evolutionary Computation, 5*, 456-470.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-291.

Kita, H., Sato, H., Mori, N. & Ono, I. (2003). U-Mart system, software for open experiments of artificial market. *CIRA IEEE 2003 Computational Intelligence in Robotics and Automation (CD-ROM).*

Shefrin, H. & Statman, M. (1985). The deposition to sell winners too early and ride losers too long: Theory and evidence, *Journal of Finance, 40*, 777-792.

Smith, K. C. S. (1996). Decision making in rapidly changing environments: Trading in the spot currency market. *Ph. D Thesis,* Department of Information and Decision Science, University of Minnesota, Minneapolis.

Smith, K. C. S. (1997). How currency traders think about the spot market's thinking. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, 703-708. Mahwah, NJ: Laurence Erlbaum Associates.

Ueda, T., Taniguchi, K. & Nakajima, Y. (2003). An analysis of U-Mart experiments by machine and human agents. *CIRA IEEE 2003 Computational Intelligence in Robotics and Automation (CD-ROM).*

Weber, M. & Camerer, C. (1998). The decomposition effect in securities trading: An experimental analysis. *Journal of Economic Behavior and organization, 33*, 167-184.

Cross-Modal Interaction in Graphical Communication

Ichiro Umata (umata@atr.jp)
ATR Media Information Science Laboratories;
Seika Soraku Kyoto, 619-0288 Japan
Yasuhiro Katagiri (katagiri@atr.jp)
ATR Media Information Science Laboratories;
Seika Soraku Kyoto, 619-0288 Japan

Abstract

Cross-modal interaction in graphical communication is observed in collaborative problem-solving settings. Graphical communications, such as dialogues using maps, drawings, or pictures, provide people with two independent modalities: speech and drawing. Although the amount of drawing/self-speech overlap is strongly affected by activity-dependent constraints imposed by the task, the amount of drawing/partner's speech overlap is affected only weakly by these constraints. However, they do affect the function of the utterances in the case of drawing/partner's speech overlap. These results show that activity-level constraints affect the way speech coordinates drawing activities in cross-modal interaction. Furthermore, it suggests that turn-taking in multimodal communication requires general analyses integrating the functions of different modalities.

Introduction

Every joint activity requires coordination among its participants. When a band plays a piece, each member has to work on the same key, keep the same rhythm, and start and end at the same time (Clark (1996)). Some of these coordinating acts can be done across different modalities. In the case of music, a soloist can signal the end of her inprovisation not only with a phrase suggesting the solo's end, but also with eye-contact.

Communication is also a joint activity, and participants must coordinate with each other. One outstanding coordination principle in conversation is sequential turn-taking in speech channels. Several studies have been carried out on speech turn coordination, and some of them analyze cross-modal interaction between speech and nonverbal behaviors such as gaze and posture (Argyle et al. (1976), Kendon (1967)). In this paper, we investigate the interaction between speech and drawing, another powerful communication medium.

Turn-taking in speech involves a wide variety of factors such as sociological principles, the limitations of human cognitive capacity, and so on. One potentially strong factor for sequential turns in speech is the resource characteristics of media: speech media affords only one person's speech sounds at a time. Sacks et al. (1974) regard verbal turns as an economic resource, distributed to conversation participants according to turn organization rules. According to them, one of the main effects of these turn organization rules is the sequentiality of utterances. They observe that one party talks at a time in most cases.

Drawing, on the contrary, has quite different characteristics from speech. First, drawing is persistent whereas speech is not. Drawing remains unless erased, whereas speech dissipates right after it occurs. A drawing can be understood much later than when it is actually drawn, whereas speech must occur in real time. Second, drawing has a much wider bandwidth than speech. Two or more drawing operations can occur at the same time without interfering with each other, whereas simultaneous utterances are hard to understand. These resource characteristics allow for simultaneous drawing. There have been several studies on drawing interaction in the Human Computer Interaction field in the context of computer-supported collaborative work. Some researchers are optimimistic about the possibilities of simultaneous drawing (Stefik et al. (1987), Whittaker et al. (1991)), though others are not (Tatar et al. (1991)).

To approach this problem, Umata et al. (2003) have introduced yet another view based on the activity-dependent constraints imposed by the task performed in the interaction. The analyses show that sequential structure is mandatory in drawing either when the drawing reflects the dependency among the information to be expressed or when the drawing process itself reflects the proceedings of a target event. Further analyses show that speech interaction, which is already restricted by the resource characteristics of media, is not affected by activity-dependent constraints (Umata et al. (2004)).

The relation between drawing and speech modalities is, however, still not quite clear. Takeoka et al. (2003) analyzed face-to-face graphical communication and found that both utterances without drawings and utterances followed by the speaker's drawings behave similary in turn-holding function. They also show that longer silences are allowed while drawing is taking place. These results suggest that turns in communication can be maintained across speech and drawing modalities. This is also supported by the finding that drawing/self-speech overlap is much more frequent than drawing/partner's speech overlap (Umata et al. (2004)). The assumption of continuous turns across modalities is appealing from the viewpoint of modal integration: speech and graphic modalities decribe their target not just independently but also jointly, with linguistic phrases describing the target via graphics (Umata et al. (2000)).

In the following part of this paper, we analyze interac-

tion across these two modalities, focusing on drawing-speech overlap. The results show that the activity-dependent constraints strongly affect the amount of drawing/self-speech overlap, whereas they only weakly affect the amount of drawing/partner's speech overlap. These constraints, however, do affect how their drawing activities are coordinated verbally. We argue that activity-level constraints affect not only drawing-drawing interaction organization but also cross-modal interaction organization.

Drawing Turns and Speech Turns

As we have seen in the previous section, the sequentiality of speech turns has been attributed to the resource characteristics of speech, namely non-persistence and restricted bandwidth. The assumption is that we cannot comprehend two spoken utterances at the same time because of the bandwidth limitation, while we cannot delay comprehending one utterance until later because of the non-persistent characteristic. Drawing, on the contrary, functions quite differently in regard to these assumptions, and it may have potential for parallel turn organization. There have been seemingly contradictory observations of drawing turn organization; one is that drawing turns can be parallel, and the other is that they cannot be parallel. Umata et al. (2003) suggested that there is yet another kind of constraint based on the activities people are engaged in. According to this view, sequential structure is mandatory in drawing in some cases but not in others.

Sequentiality Constraints

1. Drawing interaction occurs in sequential turns under either of the following conditions:

(a) Information Dependency Condition: When there is a dependency among the information to be expressed by drawing;

(b) Event Alignment Condition: When drawing operations themselves are used as expressions of the proceedings of target events.

2. Sequential turns are not mandatory in drawing activities when neither condition holds (and when persistence and certain bandwidths of drawing are provided).

The rationale for the information dependency condition is the intuition that when one piece of information depends on another, the grounding of the former piece of information is more efficient *after* the grounding of the latter has been completed. This should be the case whether a particular speaker is explaining the logical dependency in question to her partners or all participants are following the logical steps together.

Event alignment is a strategy for expressing the unfolding of an event dynamically, using the process of drawing itself as a representation. For example, when you are reporting on how you spent a day in a town by using a map, you might draw a line that shows the route you actually took on the map. In doing so, you are aligning the drawing event with the walking event to express the latter dynamically. Our hypothesis is that simultaneous drawing is unlikely while this strategy of event alignment is employed. Under this condition, the movement or process of drawing is the main carrier of information. The trace of drawing has only a subsidiary informational role. Thus, in this particular use of drawing, its persistency is largely irrelevant. The message must be comprehended and grounded in real time, and the bandwidth afforded by the drawing surface becomes irrelevant. This requirement effectively prohibits the occurrence of any other simultaneous drawing.

An analysis on the corpus gathered from collaborative problem-solving tasks demonstrates that these two activity-dependent constraints can override the resource characteristics of the drawing media, thereby enforcing a sequential turn organization similar to those observed in verbal interactions (Umata et al. (2003)).

These activity-dependent constraints, however, do not affect the speech turn organization that is already affected by resource characteristics. The amount of simultaneous speech shows no difference among different task conditions (Umata et al. (2004)).

In the following part of this paper, we will look into the details of cross-modal overlap, based on the analysis of collaborative problem-solving task data gathered by Umata et al. (2003). We will compare the speech turn organization patterns in different task settings to see whether activity-dependent constraints affect the amount of drawing-speech overlap.

Method

An experiment in which subjects were asked to communicate graphically was conducted to examine the effect of the two factors presented above on their interaction organization. In these experiments, 24 pairs of subjects were asked to work collaboratively on four problem-solving tasks using virtual whiteboards.

Experimental Setting

In the experiments reported here, two subjects collaboratively worked on four different problem-solving tasks. All of the subjects were recruited from local universities and paid a small honorarium for their participation. The subjects were seated in separate, soundproof rooms and worked together in pairs using a shared virtual whiteboard (50 inches) and a full duplex audio connection. The subjects were video-taped during the experiment. They also wore cap-like eye-tracking devices that provided data indicating their eye-gaze positions. The order in which the tasks were presented was balanced between the 24 pairs so that the presentation order would not have an affect on the results. The time limit for each task was six minutes.

At the start of each task, an initial diagram was shown on the subjects' shared whiteboard and the subjects were then free to speak to one another and to draw and erase on the whiteboard. The only limitation to this drawing ac-

tivity was that they could not erase or occlude the initial diagram. All drawing activity on the whiteboard was performed with a hand-held stylus directly onto the screen, and any writing or erasing by one participant appeared simultaneously on the whiteboard in the partner's room. The stylus controlled the position of the mouse pointer and, when not drawing, the positions of both subjects' mouse pointers were displayed on the shared whiteboard.

Tasks

Deduction Task with an Event Answer (1e) A logical reasoning problem with a correct answer. The problem asks the subjects to describe the arrangement of people around a table and the order in which the people sit down. This seating arrangement and order must satisfy some restrictions (e.g., "The fifth person to sit is located on the left-hand side of person B."). A circle representing a round table was shown on the whiteboard at the start of the task. This task has strong informational dependency and strong event alignment.

Deduction Task with a State Answer (1s) A logical reasoning problem with a correct answer asking that the subjects design a seating arrangement satisfying some restrictions (e.g., "S cannot sit next to M."). A circle representing a round table was shown on the whiteboard at the start of the task. This task has strong informational dependency and loose event alignment.

Design Task with an Event Answer (2e) A task with an open-ended answer, asking subjects to make an excursion itinerary based on a given town map. A complete town map was shown on the whiteboard at the start of the task. This task has weak informational dependency and strong event alignment[1].

Design Task with a State Answer (2s) A task with an open-ended answer, asking the subjects to design a town layout to their own liking. An incomplete town map was shown on the whiteboard at the start of the task. This task has weak informational dependency and loose event alignment.

Data

During each task, all drawing, erasing, and mouse movements by each subject were recorded in a data file. Using this data, the amount of simultaneous drawing was calculated as the total time spent drawing simultaneously as a percentage of the total time either subject spent drawing (i.e., the sum of the time intervals in which both subjects drew simultaneously divided by the sum of the time intervals in which at least one of the pair drew on the whiteboard). Speech was recorded with video-data and labeled by hand. As with the drawing data, the amount of simultaneous speech was calculated as the total time spent talking simultaneously as a percentage of the total time either subject talked.

[1]Note that these categories are relative rather than absolute. For example, (2e) also has informational dependency to a certain extent in that each path has to start from the icon of the previous place they decided to visit. However, they can choose the next destination freely. Thus informational dependency is much weaker than in the cases of the seat arrangement tasks where one decision significantly narrows down the subsequent alternatives; e.g., seating a person *M* in a certain position means only *S* or *O* can sit right next to *P*, and so on.

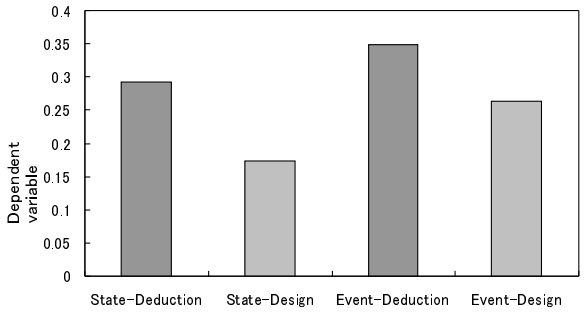


Figure 1: Proportion of drawing/self-speech overlaps

Analysis 1

Drawing/Self-Speech Overlap

As shown in Figure 1, the proportion of drawing/self-speech overlap time to total drawing time was the smallest in the design state (2s) condition. This data was entered into a 2 x 2 Analysis of Variance (ANOVA). Both problem type (deduction and design) and solution type (state and event) were treated as within-subject factors. Analysis revealed a main effect of problem type $F(1,47)=24.968$, $p<.001$ and solution type $F(1,47)=21.783$, $p<.001$ and showed no interaction $Fs < 1$.

Thus, it was shown that the proportion of drawing/self-speech overlap is smaller when the task has either weaker informational dependency or weaker event alignment, or both.

Drawing/Partner's Speech Overlap

As shown in Figure 2, the proportion of drawing/partner's speech overlap time to total drawing time demonstrated a significant, but smaller, difference in each condition compared to the case of self overlap. This data was entered into a 2 x 2 ANOVA. Both problem type (deduction and design) and solution type (state and event) were treated as within-subject factors. Analysis showed a simple main effect of solution type $F(1,47)=4.484$, $p=.04$. No effect was found for the problem type, and analysis showed no interaction $Fs < 1$.

The analysis showed that the proportion of drawing/partner's speech overlap is only weakly affected by the event alignment condition.

Discussion for Analysis 1

The amount of drawing/self-speech overlap is smaller when the task has either weaker informational dependency or weaker event alignment, or both. The activity-dependent constraints work on self-cross-modal overlap

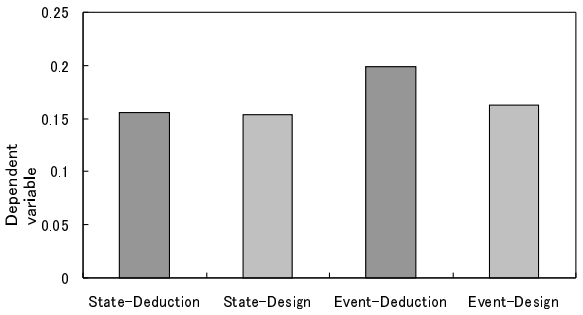


Figure 2: Proportion of drawing/partner's speech overlaps

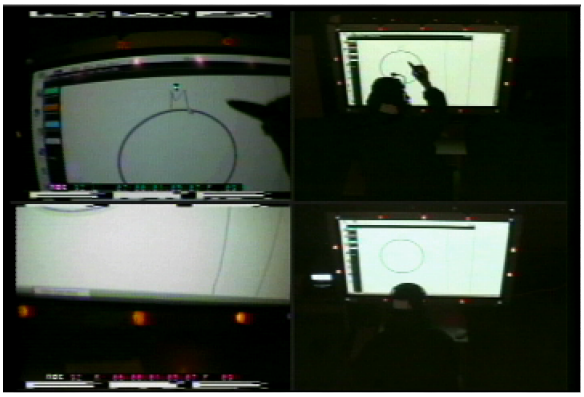


Figure 3: Sequential drawing interaction coordinated verbally (1)

in the opposite way of simultaneous drawing: the amount of simultaneous drawing is smaller when the task has stronger information dependency or weaker event alignment, or both.

This result seems quite reasonable if we consider the way people coordinate their drawing activities verbally. Whittaker et al. (1991) observed verbal coordination of drawing activities through the examination of shared whiteboard communication *with* and *without* the addition of a speech channel. They found that permanent media such as a whiteboard provides users with space for constructing shared data structures around which they can organize their activity. With the addition of a speech channel, people used the whiteboards to construct shared data structures that made up the CONTENT of the communication, while speech was used for coordinating the PROCESS of communication.

As observed in Umata et al. (2004), utterances coordinating drawing activities are also commonly found in our tasks. Figure 3 is a snapshot from the deductive state task (1s). Subjects *A* and *B* have just agreed to fix *M*'s seat first, and *A* suggests "*M*'s seat should be ... here, right?" while drawing the sign M. Then, *B* gives verbal acknowledgement, "Yes." Here, *A*'s utterance serves as a signal for his drawing activity.

Such signal utterances typically preceed drawings, and drawings follow, overlapping them. Signal utterances are expected to occur more often when people feel a stronger need to coordinate their drawing activities; i.e., in cases where activity-level constraints require sequential drawing turns. As expected, drawing/self-speech overlap is most frequent when the task has strong informational dependency or tight event alignment, or both.

There are two other possible explanations for the result. The first is that drawers have to give more verbal explainations of what they are doing as the task increases in difficulty. This does not seem to be the case, though. First, those signal utterances are usually quite simple and short: e.g., "*M* is here," "Station," etc. Second, their drawings are generally simple and easy to understand even in the tasks with stronger constraints. In the seat arrangement tasks ((1e), (1s)), each icon is an alphabetic letter standing for a person. Its position on the table icon simply shows where the person has to be seated, and the sequence of letters beside the table icon means the order of the seating. In the case of the excursion itinerary task (2e), the drawings were mainly route icons and labels showing time of arrival/departure and so on. The meaning of each drawing is also clear to the partner in this case. On the other hand, some icons can be unintelligible to the partners in the case of the town layout task (2s): a box can mean a building icon, a station icon, or anything else. Actually, people sometimes had to ask their partners for more clarification in (2s). Thus, the utterances about what they are drawing are likely to be just signals rather than detailed explanation of their drawing.

The second possible explanation is that simultaneous drawing and cross-modal overlap are affected not by the activity-level constraint but by the symbolic status of the drawing. That is, the drawing requires sequential drawing turn organization in (1s), (1e) and (2e) because they are not just a set of icons but rather a language-like symbolic system. This is also unlikely, since the drawings are almost equally simple throughout the tasks, as described above. It is possible, though, that more complicated symbolic systems require sequential turns and that it is difficult to separate the effect of the activity-level constraint and that of symbolic constraction. More work is required to illuminate the detailed mechanism underlining sequential drawing turn organization.

The activity-level constraints have a much weaker effect on the amount of drawing/partner's speech overlap. Because people cannot precisely predict when and where their partner will start drawing, verbal coordination of drawing activities typically takes the form of signal utterances. This may be why these constraints did not impact strongly on the amount of drawing/partner's speech overlap.

Another possible explanation is that turns in graphical communication tend to be maintained across speech and drawing modalities. Drawing/self-speech overlap is much more common than partner's speech overlap

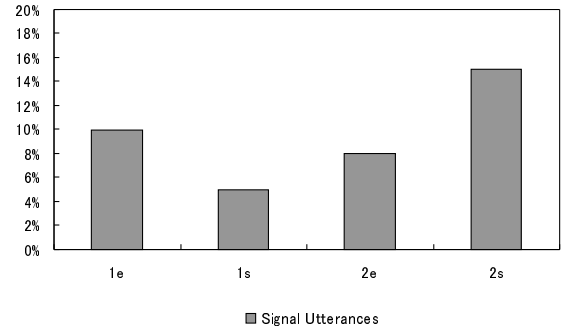


Figure 4: Frequencies of verbal signals for drawing

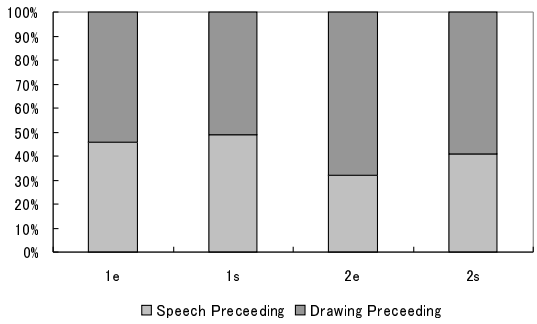


Figure 5: Frequencies of utterance preceeding overlap

(Umata et al. (2004)). The effect of activity-level constraints is much weaker, perhaps because the cross-modal turn organization already blocks speech overlap by partners.

Analysis 2

It was shown that the activity-dependent constraints affect the amount of drawing/partner's speech overlap only weakly, whereas they strongly affect the amount of drawing/self-speech overlap. In this section, we analyze drawing/partner's speech overlap in more detail to determine whether there are any differences among task conditions.

The drawing occurences analyzed above were all recorded as the time duration that the pen is touching the screen. Some drawing activities are divided into segments that are too small under this method. For example, some subjects drew many dots or lines to give colors to some icons. It is unreasonable to divide such an activity into many drawing occurences when we perform closer analysis on each overlapping case of drawing and speech modalities. The drawing occurences within 400 msec gaps are regarded as a *drawing unit* for the analysis below, in the same way as when we divide speech into utterance units. One member of each of the 24 dyads tested was randomly selected for the following analyses.

Verbal Signals for Drawings

The frequencies of verbal signals in all drawing/partner's speech overlap were compared among different task conditions. The analysis showed significantly different proportions among conditions ($\chi^2_{(3)} = 13.775, p < .003$). More concretely, verbal signals in drawing/partner's speech overlap are most frequent in the design state condition (2s), as shown in Figure 4 (adjusted residual: (1e) $= -1.2$, (1s) $= -8.7$, (2e) = -5.4, (2s) $= 15.3$). The design state condition has fewer verbal signals for drawing overall, so their high frequency in drawing/partner's speech overlap is rather outstanding.

Other Findings: Drawing Preceeded Overlaps

We also compared the frequencies of drawing preceeding overlap in drawing/partner's speech overlap among different task conditions. The analysis shows significantly different proportions between deduction conditions (1e, 1s) and design (2e, 2s) conditions ($\chi^2_{(3)} = 7.740, p < .005$). More concretely, the design conditions have fewer drawing preceeding overlap than the deduction conditions, as shown in Figure 5 (adjusted residual: (1e, 1s) $= -19.2$, (2e, 2s) $= 19.2$). People start drawing while their partners are speaking more often in the design condition than the deduction condition.

Discussions for Analysis 2

Drawing/partner's speech overlap includes more verbal signals for drawings in the design state condition (2s) than in any other condition. This reflects the parallel interaction style of drawing in (2s). While verbal signals serve to maintain sequential drawing interaction in the case with stronger activity-level constraints, these signals often serve to coordinate parallel drawing activities when they occur in (2s). Verbal signals also overlap the partner's drawings in some of these cases. Figure 6 shows one such case. Subjects *A* and *B* agreed to divide the design task into two sub-tasks, the design of a station plaza and that of a park. Then, *A* said "Station," and *B* said "I'll make the forest," before starting their respective drawing activities. Here, they verbally coordinated their simultaneous drawing activity, and their verbal signals overlap their partner's drawings.

Drawing preceeding overlap is more frequent in the design condition (1) than in the deduction condition (2). This means only the information dependency constraint affected the frequency of speech preceeding overlaps. Although we cannot give any clear explanation for this phenomenon, we assume this result reflects the different characteristics of these two activity-dependent constraints. The information dependency constraint has a more general nature across modalities: when one piece of information depends on another, the grounding of the former piece of information is more efficient after the grounding of the latter has been completed. On the contrary, event alignment is rather drawing-modality-specific: the drawing process reflects the process of the described event. In this sense, drawing activities are less dependent on the information given in speech modalities

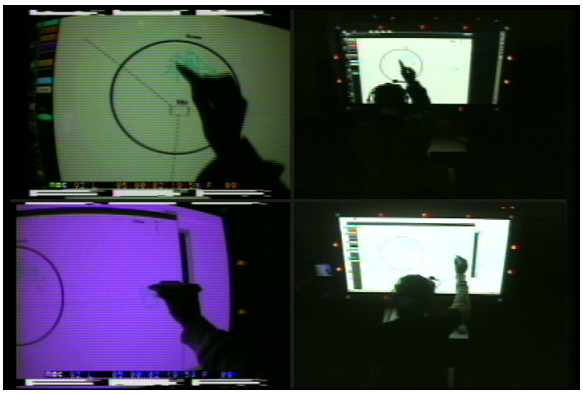


Figure 6: Parallel drawing interaction coordinated verbally

than in cases with strong information dependency. However, the mechanism causing this phenomenon remains unclear. More work is required to demonstrate how the two modalities interact.

Conclusions

Based on the data of collaborative task solving settings, we have analyzed cross-modal interaction in graphical communication. We found that the amount of drawing/self-speech overlap is strongly affected by the activity-dependent constraints, while the amount of drawing/partner's speech overlap is affected only weakly by these constraints.

There are, however, significant differences in the function of the utterances in the case of drawing/partner's speech overlap. Drawing/partner's speech overlap includes more signal utterances for drawing when the activity-level constraints are weaker. This result reflects the parallel interaction style of drawing under weak activity-level constraints.

The precedence of drawing/partner's speech overlap is also affected by the information dependency constraint. Although it is likely that the modality-general nature of this constraint plays a significant role, the mechanism of this phenomenon is still not clear.

These findings indicate that the activity-level constraints affect the way speech coordinates drawing activities in cross-modal interaction and suggest that interaction organization in multimodal communication is a complex phenomenon that requires general analyses integrating the functions of different modalities.

Acknowledgments

This research was supported in part by The National Institute of Information and Communications Technology (NICT) of Japan. We would like to thank the anonymous referees for their insightful comments. Thanks also must go to Takugo Fukaya for his intellectual support of this work.

References

Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Brennan, S. E. (1990). Seeking and providing evidence for mutual understanding. Ph.D. dissertation, Stanford University.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*, 259–294.

Condon, W. S. (1971). Speech and body motion synchrony of the speaker-hearer. In D. L. Horton and J. J. Jenkins (Eds.), *Perception of Language*. Columbus, Ohio: Merill, 150–173

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 32*, 1–25.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50*, 696–735.

Stefik, M., Foster, G., Bobrow, D., Kahn, K., Lanning, S., and Suchman, L. (1987). Beyond the chalkboard: Computer support for collaboration and problem solving in meetings. *Communications of the ACM, 30 (1)*, 32–47.

Takeoka, A., Shimojima A., and Katagiri, Y. (2003). Turn-taking in graphical communication: An exploratory study. In *Proceedings of the Fourth SIGdial Workshop on Discourse and Dialogue*.

Tatar, D., Foster, G., and Bobrow, D. (1991). Design for conversation: Lessons from cognoter. *International Journal of Man-Machine Studies, 34(2)*, 185–210.

Traum, D. (1994). A computational theory of grounding in natural language conversation. Ph.D. dissertation, University of Rochester.

Umata, I., Shimojima, A., and Katagiri, Y. (2000). Talking through graphics: An empirical study of the sequential integration of modalities. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 529–534.

Umata, I., Shimojima, A., Katagiri, Y., and Swoboda, N. (2003). Graphical turns in multimodal communication. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (CD-ROM).

Umata, I., Shimojima, A., Katagiri, Y. (2004). Speech and graphical interaction in multimodal communication. To appear in *Proceedings of Diagrams 2004*.

Whittaker, S., Brennan, S., and Clark, H. (1991). Coordinating activity: An analysis of computer supported co-operative work. In *Proceedings of CHI'91 Human Factors in Computing Systems*, 361–367, New York: ACM Press.

Stylistic and Contextual Effects in Irony Processing

Akira Utsumi (utsumi@se.uec.ac.jp)
Department of Systems Engineering, The University of Electro-Communications
1-5-1 Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Abstract

Irony is perceived through a complex interaction between an utterance and its context and serves many social functions such as to be sarcastic and to be humorous. The purpose of this paper is to explore what role linguistic style and contextual information play in the recognition of irony (i.e., assessing the degree of irony) and in the appreciation of ironic functions (i.e., assessing the degree of sarcasm and humor). Two experiments demonstrated that the degree of irony and sarcasm was affected primarily by linguistic style (i.e., sentence type and politeness), while the degree of humor was affected by both linguistic style and contextual information (i.e., context negativity and ordinariness of negative situation). These results are almost consistent with the predictions by the implicit display theory, a cognitive theory of verbal irony. Discussion of the findings also suggests that the implicit display theory can account for an indirect effect of context on the degree of irony.

Introduction

Irony is an interesting pragmatic phenomenon whose processing involves complex interaction between linguistic style and contextual information. There are also good reasons for probing the mechanism of irony processing in cognitive science. First, irony offers an effective way of accomplishing various communication goals for maintaining and modifying social and interpersonal relationships that are difficult to do literally. Second, irony processing requires higher-order mindreading ability (Happé, 1993), which has been argued to play an important role in the interpretation of ordinary utterances (Wilson and Sperber, 2004). Third, as Gibbs (1994) argues, an ironic way of talking about experiences reflects our figurative foundation for everyday thought.

Recently, many studies have paid much attention to irony processing (e.g., Gibbs, 1994; Sperber and Wilson, 1995; Attardo, 2000; Colston, 2002; Giora, 2003). However, most of these studies focus only on the difference of processing between ironic utterances and literal ones, in spite of the fact that irony is communicated by various kinds of expression (Kumon-Nakamura, Glucksberg, and Brown, 1995; Utsumi, 2000). For example, to your partner who stepped on your feet many times during a dance, you can say ironically in various ways: not only an opposition statement like "You're really a good dancer", but also a true assertion "I love good dancers", a rhetorical question "Could you step on your own two feet?", a circumlocutory utterance "I guess you have a broken leg", and so on. The purpose of this study is to empirically examine how irony processing differs among different kinds of ironic utterances and what role style and context play in causing such differences.

The issue of controversy in irony research is according to what features of irony people distinguish irony from nonirony. Beyond the fallacious view that irony is a meaning opposition or a mere violation, a number of studies have proposed a variety of views of irony: Irony is an echoic interpretation of an attributed thought (Sperber and Wilson, 1995), joint pretense (Clark, 1996), relevant inappropriateness (Attardo, 2000), or indirect negation (Giora, 2003). However, these theories suffer from the same problem that they have attempted to provide necessary and/or sufficient properties for distinguishing irony from nonirony; there appear to be no such properties shared by all ironic utterances. To overcome this difficulty, I have proposed a more comprehensive view of irony, *implicit display theory of verbal irony* (Utsumi, 2000). The implicit display theory takes a comparative view that irony is a prototype-based category, which is the idea underlying cognitive linguistic research. Another point in which the implicit display theory radically differs from the previous views is that it claims a differential role of style and context, whereas the previous theories do not address such a difference or they confuse the different roles. According to the implicit display theory, style of an ironic expression is used to assess to what degree a specific ironic utterance is similar to the prototype of irony, while context motivates the addressee to interpret an expression ironically. The study I present in this paper empirically examined to what degree people perceive an utterance as ironic depending on style of the utterance and its context, and tested whether the claims of the implicit display theory can explain the observed result.

Another heated topic in irony research is the social function of irony, which provides a plausible answer to why people use irony. The functions are divided into negative ones such as to be sarcastic and to criticize, and positive ones such as to be humorous. Previous studies (e.g., Dews and Winner, 1995; Colston, 2002) have compared the degrees of negative effect between ironic utterances and literal equivalent utterances. However, these studies have not addressed how various kinds of ironic utterances differ in negative and positive functions. My study thus examined both negative and positive effects of various ironic utterances by asking people to rate the degree of sarcasm and humor, and tested whether the obtained finding can be explained by the implicit display theory.

Implicit Display Theory

The main claim of the implicit display theory is threefold (Utsumi, 2000). First, irony presupposes *ironic environment*,

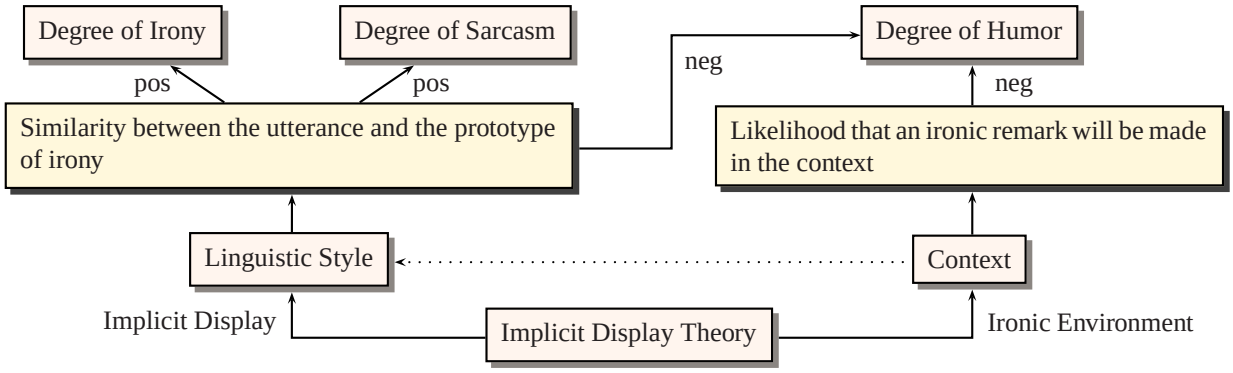


Figure 1: General hypothesis for irony processing elicited from the implicit display theory.

a proper situational setting in the discourse context. Ironic environment consists of (a) speaker's expectation, (b) incongruity between the expectation and the reality, and (c) speaker's negative attitude toward the incongruity. In order for an utterance to be interpreted ironically, the implicit display theory argues, the discourse situation must be identified as ironic environment through the process of checking or inferring these constituents. In the 'dance' example presented above, you have expected that your partner dances well with you but your expectation is not fulfilled, and you gets disappointed or angry at the result. That situation is thus identified as ironic environment.

Second, irony is an utterance that *implicitly displays* ironic environment. Implicit display of ironic environment is achieved by an utterance which (d) alludes to the speaker's expectation, (e) includes pragmatic insincerity by violating one of pragmatic principles, and (f) expresses indirectly the speaker's negative attitude by being accompanied by ironic cues. For example, your utterance "You're really a good dancer" in the above situation satisfies the three conditions of implicit display. First, it mentions, and thus alludes to, your expectation of the partner dancing well. Second, it is a literally false statement that violates the maxim of quality. Third, the hyperbolic word "really" is used to exaggerate the ironic attitude.

Third, as I mentioned in the introduction, irony is a prototype-based category characterized by the notion of implicit display. The prototype of irony is an abstract exemplar which completely meets all the three conditions for implicit display. The degree of irony can be assessed by the similarity between the prototype and a given utterance with respect to the three conditions. Let us consider again the 'dance' example. A circumlocutory statement "I guess you have a broken leg" can be interpreted ironically, but its degree of ironicalness may be much smaller than the typical type of irony "You're really a good dancer". This difference can be explained in terms of to what degree an utterance achieves the implicit display. The circumlocutory statement is only weakly related to the speaker's expectation by a number of coherence relations, whereas the opposition statement directly refers to the expectation. Furthermore, the circumlocutory statement is pragmatically insincere to a much lesser degree than the opposition statement including an apparent violation.

General Hypothesis

The implicit display theory posits the hypothesis for irony processing, which is summarized in Figure 1. On the one hand, style of an ironic sentence, which corresponds to properties of implicit display, governs how similar it is to the irony prototype, i.e., the degree of irony. On the other hand, context determines how likely one is to make an ironic remark, i.e., likelihood of irony, based on to what degree each of the three constituents for ironic environment holds in that context.

This differential role of style and context allows us to draw a general hypothesis about the degree of irony: *The degree of irony is affected by linguistic choice, not by contextual setting, and it is high to the extent that the properties of implicit display are satisfied*. Furthermore, it is reasonably assumed that the degree of sarcasm of ironic utterances proportionally depends on the degree of irony because sarcasm is often conveyed in the form of irony. It is therefore hypothesized that *the degree of sarcasm of an ironic utterance is affected only by linguistic style and it is high to the extent that the properties of implicit display are satisfied*. Note that the hypothesis on the degree of sarcasm does not hold true for victimless irony, which are often perceived as nonsarcastic (Kreuz and Glucksberg, 1989). Because this study attempts to explore the negative function toward a victim of irony, I did not use victimless ironies in the experiments.

Unlike irony and sarcasm, how the degree of humor is determined cannot be directly explained by the implicit display theory. I thereby adopt an incongruity-resolution model of humor (Attardo, 1997), a cognitive model widely accepted in humor research. The incongruity-resolution model argues that humor involves an incongruity between what was expected based on our conceptual pattern and what occurred in the humorous event, which is often expressed by a punch line in humorous texts. When such incongruity is resolved immediately by generating a reinterpretation of a humorous expression, humorous effect takes place. Since we are concerned with interpretable ironic utterances (i.e., they are assumed to be equally resolvable), it is hypothesized that the degree of humor proportionally depends on the degree of incongruity involved in ironic utterances. According to the implicit display theory, ironic utterances involve two kinds of incongruity: (a) incongruity between an expected type of utterance (e.g., ironic or literal) and the actual type of a given utterance (i.e., irony in this paper), degree of which is inversely related

to the likelihood of irony; and (b) incongruity (i.e., dissimilarity) between the irony prototype and a given ironic utterance. If the incongruity-resolution model and the implicit display theory are plausible, a general hypothesis about the degree of humor is as follows: *The degree of humor of an ironic utterance is affected by both linguistic style and context, and it is high to the extent that a discourse context is incongruous to the ironic environment or that the utterance is dissimilar to the irony prototype.*

Experiment 1

The purpose of Experiment 1 is to test the implicit display theory by examining how linguistic style affects the degree of irony, sarcasm and humor. Linguistic style of irony was manipulated by two factors: sentence type and politeness level.

Three sentence types were used in Experiment 1: opposition, rhetorical question and circumlocution. An opposition is a statement whose positive literal meaning is the opposite of the negative situation and thus includes the speaker's expected event or state. A rhetorical question is an interrogative statement by which the speaker rhetorically asks for the obvious fact to the addressee. A circumlocution is a kind of understatement which is weakly related to the speaker's expectation by a number of coherence relations. It is reasonably assumed that an opposition is more related to, and thus more alludes to, the speaker's expectation than a rhetorical question and a circumlocution, and that an opposition and a rhetorical question are pragmatically more insincere than a circumlocution. It follows that an opposition would be the most similar to the prototype of irony and that a rhetorical question would be more similar than a circumlocution.

Politeness is also an important linguistic property which can signal irony. Some experimental studies (Kumon-Nakamura et al., 1995; Okamoto, 2002) found that overpolite utterances are perceived as more ironic. In Experiment 1, politeness level was manipulated by the combination of the use or nonuse of Japanese honorifics (i.e., a system of politeness expressions incorporated into the grammar) and the relationship between the speaker and the addressee (good or bad). The reason for considering speaker-addressee relationship is that whether the use of honorifics shows overpoliteness is determined according to the speaker-addressee relationship (Okamoto, 2002). Generally speaking, when the speaker and the addressee are intimate or on good terms, an utterance with honorifics would be overpolite and unnatural. On the other hand, when they are not intimate or on bad terms, honorifics are usually used for an utterance to be appropriately polite; an utterance without honorifics would be impolite or rude. According to the implicit display theory, overpolite utterances are pragmatically insincere because they can be seen as violating the convention in linguistic politeness. Therefore, other things being equal, overpolite utterances are more similar to the prototype of irony than appropriately polite or impolite utterances.

Prediction

The general hypothesis by the implicit display theory makes the following predictions on the stylistic effect.

(1) Oppositions are the most ironic and the most sarcastic, and rhetorical questions are more ironic and more sarcastic than circumlocutions. On the other hand, circumlocutions are the most humorous, and rhetorical questions are more humorous than oppositions.

(2) Overpolite utterances, i.e., utterances with honorifics by the speaker who is on good terms with the addressee, are more ironic, more sarcastic and less humorous than appropriately polite or impolite utterances.

Method

Participants One hundred and twenty undergraduate students participated for this experiment. All were native Japanese speakers.

Materials and Design Twelve stories were constructed in which the addressee was responsible for the negative situation (and thus a victim) and in which the speaker gave a remark toward the addressee. Each of the stories had two versions: Speaker-addressee relationship is good or bad. Each story was followed by one of the six versions of the final utterance (three sentence types × with/without honorifics). An example of the stories and the final remarks is as follows[1]:

> In the restaurant, the customer was not served the ordered dishes for a while. He said to the master of the restaurant, who is on {good / bad} terms with him:
>
> Opposition: "This restaurant serves the dishes quickly."
> (*Kokoha ryouri wo dasunoga hayai* {*ne* / *desu ne*}.)
>
> Question: "Do you know the recipe for the dishes?"
> (*Ryouri no tsukurikata wo shitteiru* {*no? / no desuka?*}.)
>
> Circumlocution: "I think you are just going to buy recipe ingredients."
> (*Ima zairyou wo kai ni itteiru kato* {*omotta / omoimashita*} *yo*.)

Procedure Each participant was assigned to 12 different stories involving 12 combinations of conditions. The participants read each story and rated the final utterance at the end of the story on the following two 7-point scales: "How sarcastic is the speaker's remark?" (1 = not at all sarcastic; 7 = extremely sarcastic) and "How humorous is the speaker's remark?" (1 = not at all humorous; 7 = extremely humorous). After reading and rating all stories, they read the stories again and rated the degree of irony ("Do you feel the speaker's remark is ironic?") of all the final utterances on a 7-point scale (1 = not at all ironic; 7 = extremely ironic).

Results and Discussion

Type (opposition, rhetorical question, circumlocution) × Honorifics (with honorifics, without honorifics) × Relationship (good, bad) repeated-measures ANOVAs were conducted. In all analyses, the data were analyzed by subjects (F_1) and by items (F_2).

Irony and Sarcasm Ratings The main effect of sentence type was significant both for the degree of irony (only by subject analysis), $F_1(2, 238) = 5.30$, $p < .01$, and for the degree of sarcasm, $F_1(2, 238) = 16.18$, $p < .001$, $F_2(2, 22) = 5.39$, $p < .05$. Post-hoc pairwise comparisons ($p < .05$) revealed that oppositions were significantly more ironic and more sarcastic than circumlocutions, and more sarcastic than rhetorical questions, as shown in Figure 2. Moreover, rhetorical

[1]The original Japanese remarks used in the experiment are indicated by italics and honorific words are indicated by underlines.

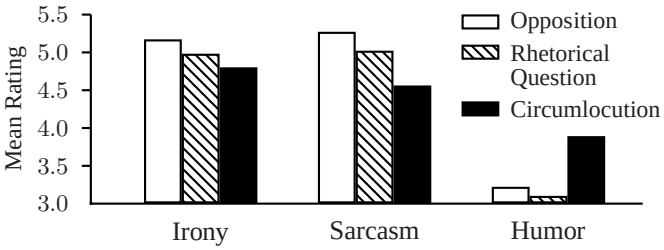


Figure 2: Mean ratings of irony, sarcasm and humor by sentence type.

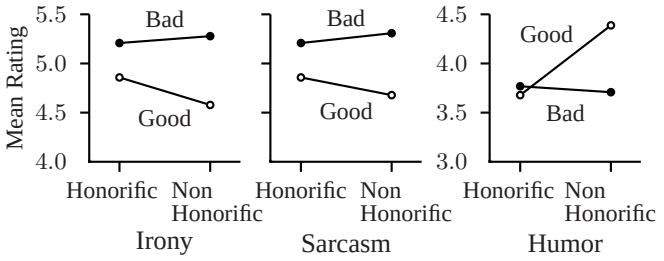


Figure 3: Mean irony and sarcasm ratings for honorific and nonhonorific utterances and mean humor ratings for honorific and nonhonorific circumlocutions in different speaker-addressee relationships.

questions were found to be significantly more sarcastic than circumlocutions. These findings are almost consistent with Prediction (1).

There was also a significant Honorifics $\times$ Relationship interaction only by subject analysis for the degree of irony, $F_1(1,119) = 5.44$, $p < .05$; and for the degree of sarcasm, $F_1(1,119) = 7.85$, $p < .01$. As shown in Figure 3, when the speaker was on good terms with the addressee, honorific utterances were rated as significantly more ironic and sarcastic than nonhonorific ones, but such difference disappeared when the speaker was on bad terms with the addressee. This result is consistent with Prediction (2) in that overpolite utterances are more ironic and sarcastic than appropriately polite utterances. However, the observed higher degrees of irony and sarcasm for the utterances by the speaker who is on bad terms with the addressee are not compatible with the prediction.

This finding against Prediction (2) was due to the significant main effect of speaker-addressee relationship. The final utterances were rated as more ironic and sarcastic when the relationship was bad than when the relationship was good, $F_1(1,119) = 21.73$, $p < .001$, $F_2(1,11) = 17.26$, $p < .01$ for the degree of irony; $F_1(1,119) = 60.55$, $p < .001$, $F_2(1,11) = 41.60$, $p < .001$ for the degree of sarcasm. This finding can be explained as an effect of contextual information (in this case, speaker-addressee relationship) on judgment whether an utterance indirectly expresses the negative attitude, i.e., condition (f) for implicit display. Information about the speaker-addressee relationship may provide an indirect cue to the speaker's negative attitude; the speaker is more likely to have a negative attitude, and thus his/her utterance is perceived as including more indirect cues and as more typical of irony when they have a bad relationship than when they have a good relationship. A number of empirical findings suggest that this explanation is plausible. Especially, in order to explain the finding that the speaker's occupations affected sarcasm ratings, Pexman and Olineck (2002) stated a similar view based on the implicit display theory: " The occupation stereotype influences interpretation because it contributes to the ironic environment. It contributes to that environment by indicating that the speaker is likely to have a negative attitude (tendency to be critical) and that such an attitude is likely to be indirectly expressed" (*ibid.*, 268).

Humor Ratings There was a significant interaction of Type$\times$Honorifics$\times$Relationship, $F_1(2,238) = 4.11$, $p < .05$, $F_2(2,22) = 4.42$, $p < .05$. The nature of this interaction was that the simple interaction of Honorifics $\times$ Relationship was observed for circumlocutions, $F_1(1,357) = 7.64$, $p < .01$, $F_2(1,33) = 7.51$, $p < .01$, but such interaction was not observed for oppositions and rhetorical questions. When the speaker and the addressee had a good relationship, circumlocutions without honorifics were rated as more humorous than those with honorifics but this difference was not observed when the relationship was bad, as shown in Figure 3. This result is consistent with Prediction (2).

There was a significant main effect of sentence type, $F_1(2,238) = 28.55$, $p < .001$, $F_2(2,22) = 19.14$, $p < .001$. Pairwise comparisons ($p < .05$) indicated that circumlocutions were significantly more humorous than oppositions and rhetorical questions, as shown in Figure 2. This result is compatible with Prediction (1).

The main effect of speaker-addressee relationship was also significant, $F_1(1,119) = 22.93$, $p < .001$, $F_2(1,11) = 37.14$, $p < .001$, showing that the utterances were rated as more humorous when the relationship was good than when the relationship was bad. We can consider two possible explanations for why good interpersonal relationship increases the degree of humor. One possible explanation may be that speaker-addressee relationship affects judgment for implicit display and thus the degree of humor, as I described above. Another explanation can be elicited from the motivational condition in which humor is experienced. Wyer and Collins (1992) stated that when the objective of the reader is to understand and enjoy humorous expressions, humor is more likely to be elicited. Therefore, a good relationship may motivate the addressee to enjoy ironic remarks, while a bad relationship may interfere with the addressee's enjoyable attitude toward them.

Experiment 2

The purpose of Experiment 2 is to test the implicit display theory with respect to contextual effect on the degree of irony, sarcasm and humor. In Experiment 2 two independent variables were considered: situational negativity (the situation is weakly or strongly negative) and ordinariness of negative situation (the negative situation is usual or unusual).

Situational negativity manipulates the degree of incongruity between the expectation and the reality, i.e., condition (b) of ironic environment, in such a way that the incongruity is perceived more easily, and thus irony may be more likely to be made, in the strongly negative context than in the weakly negative context. Ordinariness manipulates the manifestness of speaker's expectation, i.e., condition (a) of ironic environment. The expectation is more manifest in the context where an unexpected negative event occurs than in the

context where the same negative event repeatedly happens. Therefore, irony is more likely to be elicited from an unusual context than from an usual context.

Prediction

The general hypothesis by the implicit display theory makes the following predictions on the effect of context.

(3) Neither negativity nor ordinariness has an effect on the degree of irony and sarcasm.

(4) Ironic utterances in a weakly negative context are more humorous than those in a strongly negative context. In the same way, ironic utterances in an usual context are more humorous than those in an unusual context.

Method

Participants Forty-eight undergraduate students participated for this experiment. All were native Japanese speakers. None of them participated Experiment 1.

Materials and Design Eight out of 12 stories used in Experiment 1 were selected, because natural manipulation of negativity and ordinariness was not possible in the other four stories. Each story had four versions: a situation where a weakly negative event is usual or not, and a situation where a strongly negative event is usual or not. The stories of the weakly negative and unusual version were identical to the stories used in Experiment 1 except that the descriptions of the speaker-addressee relationship were deleted. Each story was followed by the final remark identical to the opposition utterance without honorifics used in Experiment 1. An example of the stories is as follows:

> {*In the restaurant* / *In the restaurant where it usually takes a while to serve dishes*}, the customer was not served the ordered dishes {*for a while* / *at all even after a very long time*}. He said to the master of the restaurant,

Procedure Each participant was assigned to eight different stories involving the four versions equally. The procedure was identical to that of Experiment 1.

Results and Discussion

The data was subjected to Negativity (weakly negative, strongly negative) $\times$ Ordinariness (usual, unusual) repeated-measures ANOVAs.

Irony and Sarcasm Ratings There were no significant main effects and no interactions for both ratings, which favors Prediction (3).

However, as I discussed in the result section of Experiment 1, there is a possibility that context (i.e., negativity and ordinariness) has an indirect influence on the degree of irony and sarcasm through its effect on judgment for implicit display. Especially, judgment on allusion to the speaker's expectation highly depends on manifestness of the expectation, because when the addressee does not know the speaker's expectation before interpreting an utterance the expectation must be inferred from the literal meaning of the utterance and contextual information (Utsumi, 2000). It is thus predicted that, other degrees of implicit display being equal, the degree of irony would be affected by context, primarily by ordinariness, when the speaker's expectation is implicit, but that it would not be affected by context when the expectation is explicit.

This prediction was tested by reanalysis of the data of Experiment 2. The stories used in Experiment 2 include two kinds of speaker's expectation: an expectation about a desirable event/state and an expectation about the addressee's belief. Because the speaker's expectation about the addressee's belief presupposes that the addressee does not notice it beforehand, it is assumed to be less manifest to the addressee than other types of expectation. Hence, the eight stories could be divided into two groups — explicit expectation version (n=4) and implicit expectation version (n=4) — according to whether the speaker's expectation is about the addressee's belief or not. An example of the texts including an implicit expectation is as follows:

> To a friend who eats sweets though she is on a diet:
>
> "You eat nothing at all today, are you?."
> (*Kyou ha zenzen tabenai nee*.)

In this case, the speaker's expectation is something like that the addressee (the speaker's friend) should know that her behavior is undesirable for a diet. Then the data of irony and sarcasm was subjected to Negativity $\times$ Ordinariness $\times$ Expectation (explicit, implicit) ANOVAs with repeated measures on the first two factors.

Concerning the degree of irony, there was a significant interaction of all the three factors, $F_2(1,6)=8.94$, $p<.05$. The nature of this interaction was that the simple interaction of Negativity $\times$ Ordinariness was significant for the implicit expectation context where the speaker's expectation was about the addressee's belief, $F_2(1,6)=6.44$, $p<.05$, but such interaction was not observed in the explicit expectation context. This finding is consistent with the prediction that context has an indirect effect on the degree of irony when the speaker's expectation is implicit.

The observed simple interaction of Negativity $\times$ Ordinariness for the implicit expectation was that in the weakly negative contexts the final utterances were rated as more ironic when the negative behavior was unusual ($M = 5.11$) than when it was usual ($M = 4.54$), but that in the strongly negative contexts the final utterances were rated as more ironic when the negative behavior was usual ($M=5.21$) than when it was unusual ($M=4.58$). This result can be interpreted as follows: The addressee is less likely to notice the speaker's expectation about his/her own belief, and thereby perceives an utterance as less ironic when his/her own negative behavior is usual than when it is not usual because of habituation effect. However, once the addressee's usual negative behavior becomes worse, he/she is more likely to be aware of the speaker's expectation because of dishabituation effect.

For the degree of sarcasm, however, there were no significant effects and interactions in the reanalysis. This result suggests that the speaker's expectation may be an important property which distinguishes irony from sarcasm; sarcasm may not need the speaker's expectation.

Humor Ratings Only the main effect of ordinariness was significant by item analysis, $F_2(1,7)=7.81$, $p<.05$. Ironic utterances in the expected contexts in which the addressee's negative behavior was usual ($M=3.23$) were rated as more humorous than the same sentences in the unexpected context

in which the negative behavior was unusual ($M = 3.06$). This result is consistent with Prediction (4). However, the result that the main effect of negativity was not significant suggests that context negativity may have little influence on the likelihood of irony.

General Discussion

As I mentioned in the introduction, the prototype-based view permits the implicit display theory to explain the obtained finding that the degree of irony differs among various utterances and contexts. For example, allusion-based theories such as Sperber and Wilson's (1995) echoic interpretation theory cannot explain why overpolite utterances were rated as more ironic than appropriately polite utterances. On the other hand, insincerity-oriented theories such as Attardo's (2000) relevant inappropriateness view cannot account for the finding that the speaker's expectation affects the degree of irony. (For details of the superiority of the implicit display theory over other theories, see Utsumi, 2000).

Furthermore, the echoic interpretation theory also fails to explain the finding that the degree of irony was affected by contextual information only when the speaker's expectation about the addressee's belief triggered irony. The reason for the difficulty in explaining such effect lies in their view that irony interpretively echoes not only the speaker's expectation but also other sources such as someone's utterances, opinions or even general norms, whereas the implicit display theory assumes that only the speaker's expectation is alluded to by irony. Therefore the echoic interpretation theory need not, and indeed does not, assume the speaker's expectation about the addressee's belief to explain irony like the 'diet' example; it assumes that irony echoes the general norm that teenagers want to be slim by a diet.

Concerning the functions of irony, the implicit display theory is more consistent with the obtained findings than the contrast-assimilation theory recently proposed by Colston (2002). He has claimed that the degree of negative effect of irony can be explained in terms of "contrast and assimilation" effects, which are often observed in perceptual judgment. If the discrepancy between the positive surface meaning of an ironic utterance and its referent negative situation is large, the ironic utterance is perceived as more negative than the literal one because of a contrast effect. On the other hand, if the discrepancy is relatively small, then an assimilation effect is more likely to occur, resulting in that ironic utterances are perceived as less negative. Although the contrast-assimilation theory seems to be compatible with the finding of Experiment 1 that the degree of sarcasm was graded according to the similarity to the irony prototype, the finding of Experiment 2 that situational negativity did not have an influence on the degree of sarcasm may provide evidence against the contrast-assimilation theory. If Colston's theory is right, an utterance should be more sarcastic in the strongly negative context and less sarcastic in the weakly negative context than the literal equivalent utterances, because negativity changes the degree of discrepancy between the utterance and the situation.

To sum up, it can be concluded that the implicit display theory provides a more consistent explanation of the obtained findings on both irony recognition and ironic function than other theories.

Acknowledgments

This research was supported by Grant-in-Aid for Encouragement of Young Scientists (No.14780263), The Ministry of Education, Culture, Sports, Science and Technology, and a grant from Nissan Science Foundation.

References

Attardo, S. (1997). The semantic foundations of cognitive theories of humor. *Humor*, *10*(4), 395–420.

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, *32*(6), 793–826.

Clark, H. (1996). *Using Language*. Cambridge University Press.

Colston, H. (2002). Contrast and assimilation in verbal irony. *Journal of Pragmatics*, *34*(2), 111–142.

Dews, S. and Winner, E. (1995). Muting the meaning: A social function of irony. *Metaphor and Symbolic Activity*, *10*(1), 3–19.

Gibbs, R. (1994). *The Poetics of Mind*. Cambridge University Press.

Giora, R. (2003). *On Our Mind: Salience, Context, and Figurative Language*. Oxford University Press.

Happé, F. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, *48*, 101–119.

Kreuz, R. and Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, *118*(4), 374–386.

Kumon-Nakamura, S., Glucksberg, S., and Brown, M. (1995). How about another piece of pie: The allusional pretense theory of discourse irony. *Journal of Experimental Psychology: General*, *124*(1), 3–21.

Okamoto, S. (2002). Politeness and the perception of irony: Honorifics in Japanese. *Metaphor and Symbol*, *17*(2), 119–139.

Pexman, P. and Olineck, L. (2002). Understanding irony: How do stereotypes cue speaker intent?. *Journal of Language and Social Psychology*, *21*(3), 245–274.

Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition, Second Edition*. Oxford, Basil Blackwell.

Utsumi, A. (2000). Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, *32*(12), 1777–1806.

Wilson, D. and Sperber, D. (2004). Relevance theory. In Horn, L. and Ward, G. (Eds.), *Handbook of Pragmatics*, pp. 607–632. Oxford, Basil Blackwell.

Wyer, R. and Collins, J. (1992). A theory of humor elicitation. *Psychological Review*, *99*(4), 663–688.

A Connectionist Model of False Memories

Saskia van Dantzig (vandantzig@fsw.eur.nl)
Psychology Department, Erasmus University Rotterdam
Postbus 1738, 3000 DR Rotterdam, The Netherlands

Eric O. Postma (postma@cs.unimaas.nl)
Department of Computer Science, University of Maastricht
P.O. Box 616, 6200 MD Maastricht, The Netherlands

Abstract

We present a connectionist model of false memories called the Associative Self-Organizing Network (ASON) model. Four mechanisms underlying the Constructive Memory Framework (CMF) guide the design of the ASON model, a connectionist operationalisation of the CMF. Simulation studies of experiments in the DRM paradigm reveal the ASON model to exhibit false memories. In addition, the effects of Mean Backward Associative Strength and output order on the probability of false recall are simulated. We conclude that the ASON model is capable of simulating and explaining the main findings on false memories.

Introduction

Memory is fallible. Every day people are confronted with the shortcomings of their memory, when forgetting things such as -for example- the phone number of a good friend, the title of a book, or the location of their car keys. Memory can also fail in another way; instead of forgetting things that did happen, people may remember events that never took place. These memories can be just as realistic as memories of real events. Such memories of never-happened episodes are called *commission errors* or *false memories*. False memories may occur in different situations and their severity can range from attributing a memory to the wrong source to confabulating a complete event (Parkin, 1997).

Various studies suggest that false memories are not simply random errors (Gallo & Roediger, 2002; Schwartz et al., 1998). Instead, they appear to be an inevitable consequence of the dynamics of human memory (Schacter et al., 1998). False memories are considered to arise from the very same mechanisms that underlie veridical recall and recognition of true memories. More specifically, we hypothesize that false memories result from the way in which memory representations are stored, processed, and retrieved.

Our approach is to investigate the occurrence of false memories in a connectionist model called the Associative Self-Organizing Network (ASON) model. The ASON model is made up of two associatively connected self-organizing maps, for storing and representing stimuli and the contexts in which they occur. Although the scientific literature on false memories is abundant (e.g. Gallo & Roediger, 2002; Johnson et al., 1993; Schacter et al., 1998), to our knowledge, no connectionist model of false memories has yet been proposed.

The outline of the remainder of this paper is as follows. In the next section we discuss the theoretical background. Then, we present the Associative Self-Organizing Network as a model of false memories. In addition, three simulations are described. Finally, we discuss the results and conclude upon the approach.

Theoretical Background

The common view of memory is that of a (re)constructive process (Roediger & McDermott, 1995; Schacter et al., 1998). This means that memories, rather than being literal reproductions of past events, are considered to be reconstructions that are susceptible to a variety of distorting factors. In this view, memories are distorted by schemes, attribution processes, prior knowledge, assumptions, and so forth. This makes it almost impossible to draw a clear boundary between true and false memories in real life situations. For this reason, our study focuses solely on false memories occurring in the experimental setting of the Deese-Roediger-McDermott (DRM) paradigm. A false memory is formalized as a recollection of a stimulus that is ascribed to the experimental context, whereas it was *not* presented during the experiment. Below, we describe the DRM paradigm in more detail.

The DRM Paradigm

In order to investigate false memories experimentally, Roediger and McDermott (1995) developed the DRM paradigm, which was a variation of a design originally used by Deese (1959). The experimental set up is as follows. Subjects are presented with lists of twelve or fifteen words that are the strongest associates of a "critical lure"; a target word which is not presented. Immediately following the presentation of a list, subjects are instructed to recall as many of the list items as possible and to mention only those words of which they are certain that they appeared on the list. Despite this instruction, subjects are about equally likely to recall the critical lure as the other items on the list (Roediger & McDermott, 1995). After completion of the experiment, which usually involves the presentation of multiple lists, recognition performance of items on all the lists is tested. It is found that subjects identify the critical lure as being a list item as often as or more often than words that were actually presented (Roediger & McDermott,

1995). These results have been widely replicated, using various lists and different variations on the basic paradigm.

The propensity to elicit false recall and false recognition of the critical item varies widely with the type of list used. Roediger et al. (2001) investigated the causes of this variability and found that the strongest predictor of false recall of the critical lure was a variable called Mean Backward Associative Strength (MBAS). MBAS is defined as the average probability that a list item elicits the critical item as its associate. Roediger et al. found that MBAS correlates positively with both false recall (r = + .70) and false recognition (r = +. 43) of the critical lure.

The ASON model is inspired by the Constructive Memory Framework of Schacter et al. (1998). In the following section this framework is discussed in detail.

The Constructive Memory Framework

Many different theories exist that address the topics of memory formation, source monitoring or reality monitoring and false memories (e.g. Gallo & Roediger, 2002; Johnson et al., 1993; Reyna & Brainerd, 1995). The general assumption underlying these different theories is that memory is constructive. This is also the central assumption of the Constructive Memory Framework (CMF) (Schacter et al., 1998). CMF proposes four mechanisms that are involved in a constructive memory system.

First, according to CMF, episodic memories can be viewed as patterns of features, with different features representing different aspects of the episode. The constituent features of a memory representation are distributed widely across different parts of the brain. Forming an episodic memory involves binding together an arbitrary configuration of information from different sources (visual, auditory, affective, semantic etcetera) about a specific episode into a unitary whole (O'Reilly & Rudy, 2001; Rolls & Treves, 1998; Schacter et al., 1998). This process is called *feature binding*.

Second, each episode activates a unique representation that can easily be discriminated from memories of similar events. Even if different memories overlap extensively, the memory system is able to retrieve the unique characteristics of each particular episode, rather than retaining only the general similarities or gist (Reyna & Brainerd, 1995). This requires a process called *pattern separation* (Schacter et al., 1998).

Third, retrieval of memories involves a process of *pattern completion*. At retrieval, a small part of the original memory is used as a retrieval cue. The subset of features representing this part of the memory is activated. Activation spreads from the activated features to the rest of the constituent features that represent that experience, and the complete memory is reconstructed.

Fourth, once a memory is reconstructed, it must be decided whether the retrieved information constitutes a real memory or is derived from internally generated information, such as thoughts or fantasies. This process is called reality monitoring. Source monitoring is a broader concept and refers to determining the source of a retrieved memory. According to Source Monitoring Theory (Johnson et al., 1993), memories from different sources have different qualitative characteristics. Source monitoring decisions capitalize on these differences. When the source monitoring mechanism fails, source amnesia occurs. One is then able to remember specific information, but unable to recall the source of this memory.

The four mechanisms of the Constructive Memory Framework lead to the notion that false memories result from a combination of two factors: (1) memories from different sources (e.g. internal and external) may form overlapping representations, and (2) the source monitoring mechanism fails to distinguish between those representations.

The activation/monitoring framework, (Gallo & Roediger, 2002; Roediger et al., 2001) explains variations in the probability of false remembering in the DRM paradigm in terms of the two factors. According to this framework, two processes, activation and monitoring, take place during the encoding and retrieval of memories. Although activation occurs mostly at the encoding stage and monitoring mostly at the retrieval stage, both processes are at work during both encoding and retrieval. The activation/monitoring framework assumes that the presentation of some items can activate entire knowledge structures or schemata. As a consequence, non-presented items can be activated because they are strongly associated with the presented items (i.e., they are part of the same knowledge structure). The activation may be the result of conscious, deliberate association, or of automatic and unconscious spreading activation. In the case of the DRM paradigm, activation spreads from the list items to related or associated concepts. The critical lure receives much activation because this item is strongly associated to each of the presented list items. This assumption is supported by the high correlation between MBAS and false remembering of the critical lure. The stronger the association between the list items and the critical word, the stronger the activation of this critical word due to automatic or deliberate spreading of activation.

Summarizing, false remembering of the critical lure occurs when the monitoring process fails to correctly attribute its activation to an internal source and the critical lure is falsely ascribed to the learning context. This monitoring process is analogous to the source monitoring or reality monitoring mechanism proposed by Johnson et al. (1993).

Implementing CMF in a Connectionist Network

The CMF acted as a guideline for the design of the ASON model. The four mechanisms of the CMF translate into the following four desired abilities of the ASON model.

(1) Ability to form episodic memories, whereby each episode leaves a unique, distinctive trace that is easily distinguishable from memories of similar episodes (i.e., demonstrate feature binding and pattern separation).

(2) Ability to retrieve or reconstruct a complete representation when cued with only a small part of the original memory (i.e., exhibit pattern completion).
(3) Ability to spread activation among related or associated concepts.
(4) Ability to monitor memory using a mechanism that decides upon the trueness of each retrieved representation.

We incorporate the four abilities in the ASON model as follows.

(1/2) Feature binding and pattern completion. Feature binding is accomplished by using an associative network, or more specifically, an auto-associator. An auto-associator typically consists of one fully-connected layer. The network's task usually is to produce an output that is similar to its input. When an input pattern is presented, the network's connection weights are changed according to a Hebbian learning algorithm. Connections between simultaneously active neurons are strengthened, whereas connections between non co-active neurons are weakened. In this way, the network is able to associate co-occurring input elements. In addition, the auto-associator is able to completely reconstruct a stored pattern, when provided with only a small part of that pattern. In other words, it can also perform pattern completion (McLeod et al., 1998).

(1/3) Pattern separation and spreading activation. In an auto-associative network, pattern separation is obtained by using sparsely distributed representations. A competitive network can be used to transform densely distributed input patterns into more sparse, separated patterns which can be processed by an auto-associator without suffering from interference. A specific kind of competitive network is the Kohonen network or self-organizing network (Haykin, 1999). For our purposes, the self-organizing network has two important advantages over a standard competitive network. First, the self-organizing network creates a topological map of the input space (Haykin, 1999). A distributed, multidimensional input is transformed into a localist representation. The self-organizing principle ensures that the information regarding relations or similarities among input patterns is not lost in this transformation. By creating a topological map of the input space, the similarity between two input patterns is reflected in the lateral distance between the two neurons representing them. This is a biologically plausible way of representing information. There is evidence that at least lower level sensory representations are organized topologically (Haykin, 1999). However, it is still uncertain whether semantic information in higher association areas is represented in a topological way as well. A second important characteristic of a self-organizing network is that there is spreading of activation among neighboring neurons. When a specific neuron in the network is excited, activation spreads to its neighbors. The degree of spreading activation is a function of the distance between the excited neuron and its neighbor. The nearest neighbors receive the most activation, and activation decreases with increasing distance. Since the neighbors of the winning neuron represent concepts resembling the input pattern, there is spreading activation between related concepts. In this way the network resembles a semantic, or conceptual map. It is generally assumed that much of our knowledge is indeed stored in the form of semantic maps or knowledge structures.

(4) Memory monitoring. A memory monitor mechanism may be implemented in the form of a module that modulates the response thresholds or connection weights of neurons in the associative layer.

In the next section the incorporation of the ASON model is described in detail.

The Associative Self-Organizing Network

The ASON model, shown in figure 1, receives two different types of input; *context* input and *stimulus* input. Input is first processed by the input/output layer of the model. This layer is made up of two unconnected parts. One part processes contextual information, the other part deals with stimulus information. Both parts of the input/output layer have I neurons. The input of the model is formed by multidimensional binary patterns. In those patterns, each bit represents the presence or absence of a specific feature by which the stimulus (item) or context is characterized. The input patterns therefore reflect conceptual representations of different stimuli and contexts.

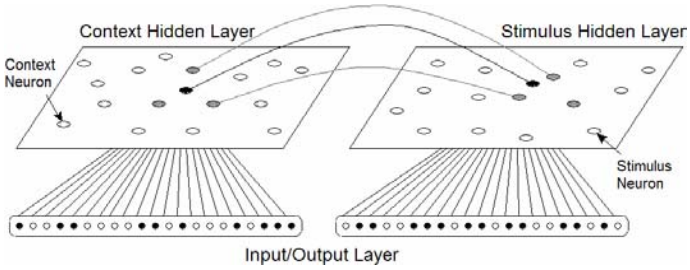


Figure 1: Schematic drawing of the Associative Self-Organizing Network. Each input pattern corresponds to one winning neuron in the hidden layer. Simultaneous presentation of a stimulus input and a context input causes an increase in the connection strength between the hidden neurons representing that stimulus and context, respectively.

Information is propagated from the input/output layer to the hidden layer. The hidden layer consists of two self-organizing maps. These maps are organized as two-dimensional lattices, each having $N \times N$ neurons. The part of the hidden layer that represents stimuli is henceforth called the *stimulus hidden layer*. The neurons making up this layer are called *stimulus neurons*. The other part of the hidden layer is called the *context hidden layer* and the constituent neurons are called *context neurons*. Both hidden layers are fully connected to each other via two-directional modifiable associative connections.

The Four Processing Stages

The processing of information in the ASON model proceeds in four stages: (i) the initialization stage, (ii) the topological-mapping stage, (iii) the learning stage, and (iv) the performance stage. Below, we discuss each of these stages in detail.

In the initialization stage, the connection weights between the input/output layer and the hidden layers, and those between both hidden layers are set to small random values.

In the topological-mapping stage, contexts and stimuli are presented to the input layer of the network and, using the Kohonen learning algorithm or SOM algorithm (Haykin, 1999), a topological organization in the hidden layers is created: semantically related concepts (overlapping input patterns) are represented by neurons that lie close to one another in the two-dimensional grid that makes up the hidden layer. In addition, associations are formed between stimuli and contexts. Whenever a particular stimulus co-occurs with a particular context, there is simultaneous activation of the winning context neuron and the winning stimulus neuron. Following an associative learning algorithm, the (associative) connection between these two hidden neurons is strengthened. Simply said, the stimulus is coupled to the context.

The learning stage simulates the learning phase of the DRM task. It refers to the presentation of the list items. During this stage, a number of stimuli are presented in one specific context -the learning context- and associations between the presented stimuli (the list items) and this context are formed. Due to spreading of activation, not only the connections between the winning context neuron and the winning stimulus neurons are strengthened, but also those between the context neuron and the neighbors of the winning stimulus neuron. The connections between the context neuron and even further neighbors of the winning stimulus neuron are actually decreased.

During the performance stage, the network can either perform a recall task or a recognition task. When performing a recall task, a context input is presented to the network as a recall cue. The winning context neuron in the hidden layer is determined and activation is propagated forwards through the associative connections towards the stimulus hidden layer. The stimulus neuron that is most strongly associated to the winning context neuron is activated and propagates its activation to the stimulus input/output layer. The weights of the connections between the winning stimulus neuron and the input/output layer have changed during the topological-mapping stage so that they have come to resemble the input pattern to which this neuron responds most strongly. Therefore, propagating activation through these connections will result in an output that resembles the original input pattern to a large degree. In other words, reconstruction of the stimulus that is most strongly associated with the presented context takes place. Subsequently, the connection between the winning context neuron and the activated stimulus neuron is 'blocked', the stimulus neuron with the second-strongest association to the context is determined and the next stimulus is recalled.

Most of the time, the stimulus recalled is one that was actually presented during the learning stage (a list item). Occasionally, however, the network recalls a stimulus that has not been presented. In other words, it has false memories. Clearly, false memories occur whenever there exists a strong association between the non-presented stimulus and the context, caused by spreading activation.

When performing a recognition task, the network is presented with a number of stimuli, both list items and a number of non-presented distractors (including the critical item). Based on the stimulus input, the winning stimulus neuron in the hidden layer is determined. The strength of the association between this winning stimulus neuron and the learning context is determined. If the strength exceeds a certain threshold, the stimulus is marked as a target and as a distractor otherwise. The decision whether to accept or reject a retrieved item is based on the strength of its association to the learning context. Raising the threshold reduces the probability of falsely recognizing the critical item, but it also decreases the hit rate. On the other hand, lowering the threshold leads to more hits, but also to more false alarms. This process is a formalization of the memory monitoring or source monitoring mechanism in various theories of memory (Gallo & Roediger, 2002; Johnson et al., 1993; Schacter et al., 1998).

To evaluate the ability of the ASON model to exhibit the false-memory performance as observed in the DRM paradigm, we performed a number of simulations that are described in the following section.

Simulations

Our simulations focus on three aspects of false memories in the DRM paradigm: the DRM effect, the role of association strength and the output order effect. All simulations were performed with the following parameter values: $I = 30$ and $N = 10$. The results reported do not depend critically on these choices.

The DRM effect The simulation of the DRM effect has two conditions; the DRM condition and a control condition. In the DRM condition, the network learns six list items that are semantically related to the critical lure. Their input patterns closely resemble the input pattern of the critical lure. In the control condition, the six list items are randomly chosen from the input set. After learning the six list items, the network performs a recall task. The results of the simulation are shown in figure 2. As is evident from the graph, the probability of recalling the critical lure is much higher in the DRM condition ($P = 0.65$) than in the control condition ($P = 0.05$), but it is lower than the average recall rate of the list items ($P = 0.78$). Hence, the ASON model simulates the DRM effect faithfully.

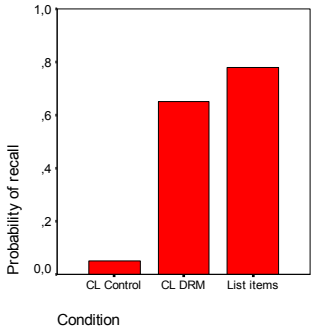


Figure 2: Probability of recalling the list items and the critical lure in the DRM (CL DRM) and control (CL Control) conditions.

The effect of association False recall and recognition of the critical lure is affected by a number of factors. As stated in the introduction, the most important factor is the Mean Backward Associative Strength (MBAS). A stronger association between the list items and the critical lure is correlated with stronger false recognition and false recall effects (Roediger et al., 2001). In the Associative Self-Organizing Network, spreading activation from the list items to the critical lure causes false recall and recognition of the latter. It is important to realize that in the ASON model concepts are related semantically instead of associatively. In contrast to what is proposed by the activation/monitoring framework, activation spreads along semantic relations rather than along associative connections. However, if we disregard this difference, we can define the Backward Associative Strength between two concepts as the lateral distance between the winning neurons that represent these concepts. The smaller the distance, the more related the two concepts are. In the network, the degree of spreading activation is a function of the distance between the excited neuron and its neighbor. Consequently, the smaller the average lateral distance between the list items and the critical lure, the stronger the activation of this critical item due to spreading activation from the list items will be. This stronger activation leads to a stronger association of the critical lure to the context, and therefore to an increasing likelihood of falsely recalling the critical lure. Figure 3a shows the results of a simulation in which the average lateral distance from the list items to the critical lure is varied. As can be seen, the probability of recalling the critical lure decreases sharply with increasing distance. The correlation between lateral distance and probability of recalling the critical lure is -.76. We compare our results with the results from a multiple regression analysis done by Roediger et al. (2001) where MBAS was found to be the strongest predictor of false recall of the critical lure (with the correlation between MBAS and probability of recalling the critical lure being +.73). Figure 3b shows the probability of recalling the critical lure as a function of MBAS, as found in the study of Roediger et al. (2001). Clearly, the results of the ASON model agree very well with those of Roediger et al.

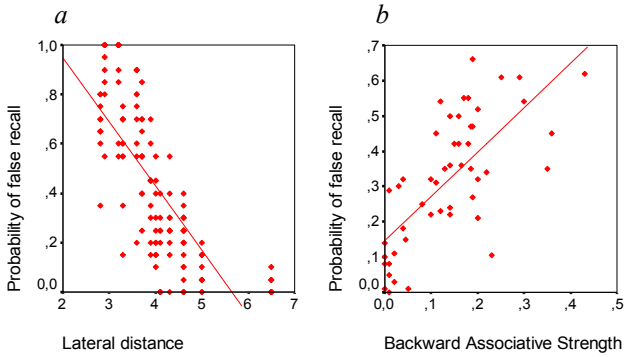


Figure 3: The probability of recalling the critical lure as a function of (a) average lateral distance between list items and the critical lure: r = -.76, and (b) MBAS: r = +.73.

The effect of output order In the third simulation we investigated the effect of output order on the probability of false recall. The *output order effect* (Schwartz et al., 1998) refers to the finding that the probability of a false memory increases with the position of items in the recall sequence.

The output order effect can be explained by the variation in association strength of presented and non-presented stimuli. False memories occur when non-presented stimuli , become strongly connected to the learning context through the processes of spreading activation and association. The association strength of those stimuli to the context is usually smaller than that of the most strongly associated targets, but larger than that of the most weakly associated targets. Since memories are generated in the order of their association strength, the probability that a false memory is generated increases with the position in the recall sequence. In our third simulation, the network performed a simple recall task, rather than a DRM task. The network learned twenty stimuli in a single context. Afterwards it performed a recall task. As can be seen in figure 4a, the probability of a false memory is largest in the last quartile of the output. Figure 4b shows the results of Schwartz et al. (1998), in which subjects performed a similar task. Evidently, our results have a striking similarity to the experimental results of Schwartz et al.

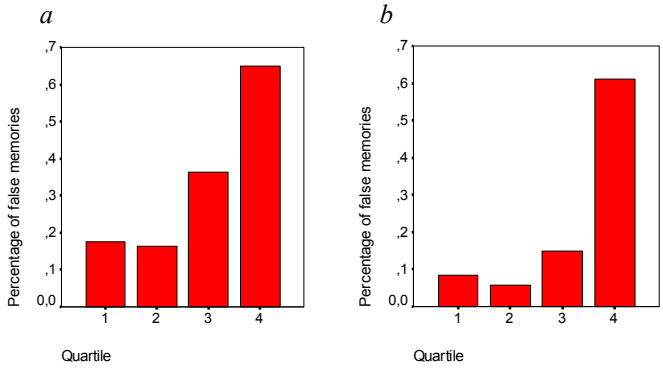


Figure 4: Number of false positives as a function of output order. Results of (a) our simulations, and (b) Schwartz et al. (1998).

Discussion and Conclusion

The ASON model demonstrates how the essential features of a constructive memory system, as put forward by CMF, can be translated into a connectionist model. Specifically, the ASON model incorporates the encoding processes of feature binding and pattern separation, as well as the retrieval processes of pattern completion and memory monitoring. In addition, it explains how spreading activation leads to high false memory scores for the critical lure in the DRM paradigm. The remaining question is to what degree the model's architecture resembles that of brain structures that are involved in the processes of storing, retrieving and monitoring of memory.

The brain structure that is considered to be responsible for the storage of episodic memories is the hippocampus (Rolls & Treves, 1998). The hippocampus is not thought to be the site of storage itself. Rather it is regarded as the mechanism that binds together the sensory features of a situation or episode to create a unitary representation of the experience. In other words, it is the structure that performs feature binding. The hippocampus receives, via the adjacent parahippocampal gyrus and entorhinal cortex, inputs from virtually all association areas in the neocortex. In addition, it gets input from the amygdala and from cholinergic and other regulatory systems (Rolls & Treves, 1998). It thus receives highly elaborated, multimodal information from various sensory pathways. Within the hippocampus, information is processed along a mainly unidirectional path, consisting of three major stages; the Dentate Gyrus (DG), the Cornu Ammonis 3 (CA3) and the Cornu Ammonis 1 (CA1). From CA1, backprojecting pathways lead via the subiculum and the entorhinal cortex back to the neocortex.

The hippocampus shares two essential characteristics with our model. First, there is a large degree of interconnectivity among neurons in the CA3 area of the hippocampus. This interconnectivity makes this area perfectly suited to perform auto-association. In fact, the idea that the CA3 area serves as an auto-associator that binds together the various elements of an episode is a core assumption in a number of computational models (O'Reilly & Rudy, 2001; Rolls & Treves, 1998). Second, the hippocampus receives a load of multimodal information from various cortical areas. The forward pathways to the hippocampus are thus characterized by strong convergence. It is hypothesized that these pathways, and the DG in particular, serve as a competitive network, transforming the widely distributed information in the cortex into more sparse, orthogonal and separated patterns that can be processed by the auto-associator without much interference (O'Reilly & Rudy, 2001).

Instead of a standard competitive network, the ASON model features a self-organizing map. The specific characteristics of this type of network, its ability to form a topological map of the input and spreading activation among neighboring neurons, can provide an explanation of false memories. Specifically, according to our model, false memories arise when activation spreads from the list items to the critical lure, causing a faulty association between this non-presented item and the learning context. This explains how false memories occur in the DRM paradigm, and gives an account of the effect of MBAS on false recall of the critical lure.

By incorporating the four mechanisms of the CMF, the ASON model is able to simulate the occurrence of false memories in the DRM paradigm and the effects of MBAS and output order on the probability of false recall of the critical item. Furthermore, its architecture is compatible with that of the hippocampus, the brain area that is widely acknowledged as being involved in the storage and retrieval of episodic memories. We conclude that this connectionist operationalisation of the CMF is able to simulate and explain the main findings on false memories.

References

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.

Gallo, D. A., & Roediger, H. L. (2002). Variability among word lists in eliciting memory illusions: evidence for associative activation and monitoring. *Journal of Memory and Language, 47*, 469-497.

Haykin, S. (1999). *Neural networks, a comprehensive foundation* (2nd ed.). Upper Saddle River: Prentice-Hall.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin, 114*, 3-28.

McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press.

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological Review, 108*(2), 311-345.

Parkin, A. J. (1997). The neuropsychology of false memory. *Learning and Individual Differences, 9*, 341-357.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: some foundational issues. *Learning and Individual Differences, 7*, 145-162.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 803-814.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: a multiple regression analysis. *Psychonomic Bulletin and Review, 8*, 385-407.

Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 48*, 289-318.

Schwartz, B. L., Fisher, R. P., & Hebert, K. (1998). The relation of output order and commission errors in free recall and eyewitness accounts. *Memory, 6*, 257-275.

A Method for Studying Representation of Action and Cognitive Distance

Davi Vann Bugmann (dbugmann@plymouth.ac.uk)
School of Psychology, University of Plymouth
Devon, PL4 8AA England (UK)

Kenny R. Coventry (kcoventry@plymouth.ac.uk)
School of Psychology, University of Plymouth
Devon, PL4 8AA England (UK)

Abstract

Past studies examining the effects of action on memory for route distance have overlooked the problem of the control of visual information. A new methodology was developed to investigate the involvement of action on the representation of route distance information in two experiments which eliminated the possible confounding effects of visual cues. In both experiments the number of turns was manipulated. Blindfolded participants learned new environments through verbal descriptions by imagining themselves walking in synchronization with metronome beats preset to match their natural walking speed. During turns, they were carefully moved. Following instructions, they performed an action at mid-route. Upon reaching the destination, their memories for the newly learned environments were tested through recall and measured again (with metronome beats representing footsteps). In Experiment 1 participants were exposed to the environment only once, and in Experiment 2 they were exposed to the environment twice. The results were consistent across the experiments and showed the influence of number of turns on remembered distances. Our data support the segmentation hypothesis with regard to the perception of the segment length and the influence of the number of turns on path distance estimates. However, our data point to a more parsimonious explanation in terms of body movement that triggers attentional processes which signal memory for events.

Introduction

When asked how far it is from one place to another there is much evidence that people do not give very accurate distance estimations. Investigations into the relationship between physical distance and cognitive distance have shown that the two differ. Furthermore, the differences between actual and cognitive distance are not random; cognitive distance is systematically distorted from the physical distance (Golledge, 1987).

The disparity in distance estimations has been explained as a function of the hierarchical organization of memory (e.g., Hirtle & Jonides, 1985; McNamara, 1986; Steven & Coupe, 1978), the organization of reference points (e.g., Sadalla, Burroughs, & Staplin, 1980), the modes of acquisition at learning (e.g., Thorndyke & Hayes-Roth, 1982), the contexts of learning (e.g., Gauvain & Rogoff, 1986; Taylor & Naylor, 2002), or the environment complexity (e.g., Sadalla & Magel, 1980; Thorndyke, 1981). Hence there is a disparate range of explanations for biases in distance estimation.

One possible explanation for bias in distance estimation, which has not been explored in detail, is that it may be a function of the actions we perform in the environment, and how those actions are cued on retrieval. This may provide a means of incorporating these multiple accounts of distance bias within a single unified framework. The view that cognition is grounded in the individual bodily interaction with the environment (e.g., Barsalou, 1999; Glenberg, 1997) is widely supported. Empirical evidence supporting the embodiment framework can be found across a range of domains. There is a tight coupling between visual perception and action. It has been shown that the representation of a visual stimulus generated from pictures or from purely linguistic descriptions can activate motor affordance, i.e., merely viewing an object, an image of an object, or hearing a description of an object results in the activation of the motor patterns necessary to interact with it (e.g., Richardson, Spivey, & Cheung, 2001; Tucker & Ellis, 1998). In language comprehension, understanding a sentence may call upon the same cognitive mechanisms as those used in planning and executing actions (Glenberg & Kaschak, 2002). It has also been shown that the representation of action or motor representation shares the same neural mechanisms as those that are responsible for the preparation and programming of actual movements (Decety, Jeannerod, & Prablanc, 1989). This evidence indicates that motor activation can occur as part of a cognitive process.

A number of studies have examined the effects of turning during route navigation. Sadalla and Magel (1980) found that paths containing several turns were perceived as being longer than paths of equivalent objective length with fewer turns, and the segmentation hypothesis has been used to explain this effect. The segmentation hypothesis claims that a right angle turn divides a pathway into segments and that the perceived lengths of the segments are combined to produce an estimate of total pathway length. Given two pathways of the same length but differing in the number of turns contained in each, the pathway with fewer turns will necessary have longer segments. These segments will be psychologically compressed to a greater extent than shorter segments (longer segments are underestimated relative to shorter segments). Therefore, the combination of a number

of compressed segments will yield an underestimation of total pathway length. This underestimation will be greater for the pathway with fewer turns. However, this study does not separate out a range of possible explanations for these effects, such as visual cues, or the rate of motion (stepping up or down, or turning) that influence the perception of traversed distance (Hermann, Norton, & Klein, 1986; Rieser, Pick, Ashmead, & Garing, 1995). For example, Hermann et al. (1986) found that the size of the effect of turns on memory for distance is affected by the number and complexity of visual cues in the environment. Therefore, to examine whether action is implicitly part of cognitive processes, it is important to have strict control over the visual information that participants could perceive and extract from the environment during navigation, and the performance of action (walking and turning). A new methodology was developed that considered all these factors in order to allow us to adequately measure whether action exerts an effect on distance estimation during navigation. In the present study, we manipulated the influence of turns on traversed distances to assess more precisely the mental mechanisms that mediate why complex routes (with many turns) were estimated differently from less complex ones (with fewer turns).

Experiment 1

The methodology was designed to control for the confounding factors present in previous studies, while maintaining realism for participants. In order to do this, a blindfold methodology was developed where participants heard linguistic descriptions describing environments over headphones, and had to imagine themselves walking around the environment in time with a series of metronome clicks preset to control for speed of walk and size of step (number of clicks heard). The aim was for participants to listen and visualize the landmarks' descriptions (thus minimizing the risk of participants gauging distances by counting steps). The environmental descriptions were formulated as guided tours, and were read by a female colleague and tape recorded for use in the experiments.

The linguistic descriptions used were controlled for number of words and detail presented. Typically the environments included five landmarks (e.g., a school, a museum, a post-office, a bank, a library, etc.). Each landmark was described by specifying its physical or historical features. Following is an excerpt of a typical description of an environment, used in the study (landmarks are in bold): "*You are in a place called Charlestown, a typical New England town. Your starting place is Victoria Park. I am going to take you on a walk from Victoria Park to St John's Basilica. It is quite a nice walk with lots of things to look at on the way. You are now standing at the gate of a place called* ***Victoria Park****. Victoria Park is renowned for its formal and shrub gardens. They are of interest and beauty in all seasons. During summer, Victoria Park hosts a Folk Music Festival. … You are now at the entrance of a place called the* ***Central Library****. Built of silvery-grey stone, the front of the building has columns and triple arches with elaborated decoration at the tops. Inside the Library, there is an intricately carved oak staircase. You are standing directly in front of the book return box. Now I will let you post the book in the return box. You can actually feel the return box in front of you. So feel the box and post the book. …*".

To encourage participants to visualize only the described scenes, a blindfold was used in order to eliminate visual information that they could have gathered from the test laboratory. Furthermore, to examine the influence of action the actual walking was replaced by mental walking. A metronome pre-set to each participant's natural walking speed (stride length and frequency of stepping) emitted beats to simulate their walking rhythm. So instead of actually walking, participants heard a certain number of metronome beats, which corresponded to the exact measure of the distance to be traversed. When the distance was mentally traversed the metronome beats ceased. However, during the simulated navigation through the environment, participants performed an action (e.g., put an object into a box) which occurred at mid-route. This manipulation allowed us to determine whether there was any difference in the perception of distance before versus after performing the action on the representation of distance. Participants also experienced the change in angular displacement when he/she arrived at 90-degree turn in the mental walk. In this instance, they were rotated to face in the appropriate direction. Once participants reached the destination landmark, their memories for the newly learned environments were measured through recall. Participants were told that they were now at the starting landmark again and had to "walk" on their own towards the destination (still wearing the blindfold). They had to describe what they "saw" on the way, and to instruct the experimenter to engage/disengage the metronome to signal the start of the mental walk or to stop walking. The dependent variables were the remembered traversed distances, which were again measured by metronome clicks. The experimental arrangement is shown in Figure 1.

Figure 1: Arrangement during tests. The participant is on the right, with the experimenter behind her.

Environment Characteristics. Participants learned two routes (Route A and Route B). They were not aware that Route B was the mirror image of Route A.

As each route contained 5 landmarks, there were 4 paths in each (denoted P1 to P4). Each path measured 64 meters which meant the total route length measured (64 m x 4) 256 m. Ninety-degree turns divide a path into segments. Each route contained 11 segments. The segment lengths were fixed at 8, 12, 16, 24, 32, and 40 m. These distances were combined to make up the length of 64 m for each path. Figure 2 shows a schematic representation of one of the environments used in the experiment. Note that in Route A, P1, P2, and P3 contained 1 turn each followed by P4 with 4 turns; in Route B, P1 contained 4 turns followed by P2, P3, and P4 with 1 turn each. The performance of action occurred at mid-route (at the middle landmark); P1 and P4 were located at the outer positions of each route, while P2 and P3 were located at the inner positions of each route.

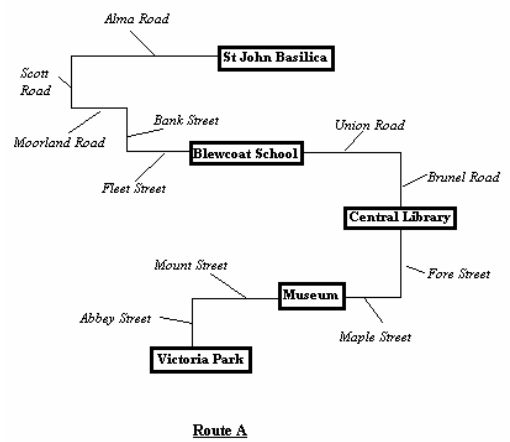


Figure 2: Configuration of Route A, in Experiment 1

Pilot Study. Before we ran the study, we tested the methodology on two pilot subjects in order to check whether they felt any discomfort during the test given that they had to wear a blindfold, and had to be physically turned during the testing procedure. However, the subjects commented that they were perfectly comfortable and relaxed during the test. We then proceeded to the first experiment using the new methodology.

Presentation. The study was presented to the participants as an investigation into people's memory for described places. They were told that they were going to listen to descriptions of imaginary walks through new environments, and were told that during the simulated walks they had to visualize the described landmarks. Additionally, they were asked to return a book or a parcel at some point en-route. The participants were not aware that their memory for distances was being tested.

Experimental Design. To examine the influence of action and the effect of number of turns on traversed distances, the experimental design used was a 2 route (Route A vs. Route B) x 2 position (inner vs. outer) x 2 action (before action vs. after action) within-subjects design.

Participants. Twenty-nine undergraduate students agreed to participate in the experiment in exchange for course credit. They were between 18 and 35 years old (mean age = 20.50, SD = 4.80). By agreeing to participant in the experiment, they were aware that they would wear a blindfold during the test.

Procedure. Participants were tested individually in a session lasting about 45 minutes. Initially, participants were instructed to walk round the room (following a pre-designated path) at their own natural walking speed so that step length and speed of walk could be established. Next, they were asked to put on the blindfold and headphones, and to stand comfortably at the centre of a circle marked on the floor. The experimenter familiarized the participants with the turning procedure: she spun the participant around on the spot, finishing by positioning him/her facing a box that was sitting on a table. At this time, the experimenter gave the participant the book or the parcel to carry with him/her. Then the participants were instructed to visualize the landmarks when they heard the descriptions, and to imagine walking in synchronization with the metronome clicks, and to stop imagining walking when the metronome ceased clicking. The experimenter then started the tape player and both listened to route descriptions through headphones. At the appropriate times, the experimenter stopped the player and engaged the metronome to implement the mental walking. During turns, the experimenter intervened by physically rotating the participants on the spot. Note that all turns were 90 degrees turns. At mid-route, participants performed the dispatch task as instructed, i.e., he/she extended his/her arm to reach the box, touched it to find the slot, and then dropped the objects into the box. Once the destination was reached, the experimenter spun the participant around again and positioned him/her in front of the box. Still blindfold, the participant's route memory was tested through recall. After the recall of the first route, the second route was immediately presented which was followed straight away by the recall.

For the recall, participants were told that they were taken back to the starting place from which they had to re-walk the routes. They were asked to describe back as accurately as possible what they "saw" en-route. They had to tell when they wanted to walk away from the landmarks and when they wanted to stop walking, so that the experimenter could engage and disengage the metronome. At turns, they had to rotate themselves on the spot and to indicate verbally the direction of turns. Once it was established that participants understood the recall instructions, the experimenter switched on the recorder that participants carried with them.

Data Treatment. The participants' recalls were transcribed. Then we proceeded to check the order of landmarks recalled

by the participants. In order to ensure that participants had a good understanding of the environments they learned, only responses with the correct sequence of landmarks were used in the analyses.

Data were obtained by first translating the number of metronome clicks (= steps) into traversed distances expressed in meters. The accuracy of turns with regard to amplitude and direction was not recorded in the present experiment.

Results

Responses from 13 participants (45%) were excluded. Twelve of these produced incorrect sequences of landmarks for one or both routes, and the remaining participant was eliminated because of poor English. Responses from 16 participants were used in the analysis (55%).

To check whether participants were not gauging distance by counting the number of steps a correlation between the total number of steps to walk Route A and Route B and the re-walked distances of both routes across participants was performed. The results showed no significant correlation, indicating that participants were not counting clicks and remembering the number of clicks on recall. As both Route A and Route B contained 11 segments each, in total there were 22 segments. For each segment, we averaged the remembered distances across participants in order to examine the correlation with the corresponding actual distances. We found an overall significant correlation between actual and remembered distances, $r_{(22)} = 0.68$, $p < 0.001$ (1-tailed), which indicates that longer segments were associated with remembering walking longer distances on recall. To examine the influence of action and the effect of number of turns on traversed distances, a 2 route (Route A vs. Route B) x 2 position (inner vs. outer) x 2 action (before action vs. after action) within-subjects ANOVA was performed on path distances. The results of the 3-way ANOVA are displayed in Table 1.

Table 1: Results of the 3-Way ANOVA on Path Distance Estimation in Experiment 1.

Source	df and F value	MS (error)	Significance
Route (R)	$F_{(1, 15)} = 1.89$	442.53	ns
Position (P)	$F_{(1, 15)} = 8.88$	1922.00	**
Action (A)	$F_{(1, 15)} = 0.93$	94.53	ns
R x P	$F_{(1, 15)} = 0.90$	105.12	ns
R x A	$F_{(1, 15)} = 0.85$	195.03	ns
P x A	$F_{(1, 15)} = 0.30$	128.00	ns
R x P x A	$F_{(1, 15)} = 8.44$	430.13	*

Note. ns: $p > .05$; *: $p < .05$; **: $p < .01$

No main effects of route, or action were found. However, there was a main effect of position on remembered path distances. Overall, participants remembered walking significantly longer distances on the outer paths (one of which contained 4 turns) than on the inner paths (which contained one turn). There was also a significant 3-way interaction between route, position, and action on remembered path distances. Follow up analyses indicated that in Route A, after the performance of action the outer path (i.e., P4 contained 4 turns) was remembered as being significantly longer than the inner path (P3 contained 1 turn), $F_{(15)} = 6.16$, $p < 0.05$. In Route B, the reverse was the case; before the performance of action the outer path (P1 contained 4 turns) was remembered as being significantly longer than the inner path (P2 contained 1 turn), $F_{(15)} = 6.64$, $p < 0.05$. This result confirmed that the influence of number of turns was a robust effect on remembered distances.

Discussion

We developed a new procedure in order to allow us to adequately measure whether action exerts an effect on distance estimation. During the experiment, none of the participants expressed any discomfort during or after the task, indicating that the methodology was appropriate.

That said, there was a large dropout rate (45%) due to participants not being able to reproduce the landmarks in the correct order (or to remember all the landmarks completely). This may have been because the task was too difficult, or because participants were exposed to the environment only once.

Despite the high dropout rate, we found that within the same routes, distance estimation was influenced by the number of turns contained in a path; paths containing four turns were remembered as being longer than paths with one turn. This result is in line with evidence from other studies (Sadalla & Magel, 1980), but with more control over visual information and action. Our procedure allowed us to observe the effect of number of turns on the same route through auditory simulated navigation, while Sadalla and Magel (1980)'s result was on separate paths, and involved actual walking. However, taking together both studies indicate that the influence of number of turns on memory for distance is a robust effect.

The absence of the effect of performing an action may be due to the salience of the action itself. The movement of dispatching (dropping) an object into a box may be perceived as a simple and routine activity therefore was not salient enough to exert an effect on spatial representation. A sequence of more pronounced movements to perform the dispatch task may make the action more memorable. For the moment, we were concerned by the high dropout rate. For that reason, in Experiment 2 we exposed participants to the same environments twice before their memories were tested using exactly the same methodology as in Experiment 1.

Experiment 2

Method

The method used was the same as in Experiment 1, except that this time participants were exposed to each environment twice before recalling routes.

As in Experiment 1, participants learned two different routes (Route A and Route B), and then they had to

reproduce each route trip in free recall. Route A and Route B were presented to participants in counterbalanced order.

Participants. Twenty-three undergraduate students agreed to participate in the experiment in exchange for course credit. Participants were between 18 and 46 years old (mean age = 24.17, SD = 7.84). They were tested individually.

Procedure. The procedure was exactly the same as in Experiment 1, however here participants were guided through each route twice before their memories for each route were tested through free recall. The tests lasted about one hour.

Results

As in Experiment 1, to be included in the analyses participants' responses must show the correct sequences of landmarks in both routes. Responses from 18 out of 23 participants (78%) were used in the analyses. Responses from 5 participants (22%) were eliminated (4 incorrect sequences of landmarks, 1 poor quality recording). The exposure to the environment twice seemed to work as the rate of data inclusion has much improved, although there is still quite a high rate of exclusion.

On average, we found that short distances were overestimated, whereas longer distances were underestimated. The overall correlation between actual and remembered distances was highly significant, $r_{(22)} = 0.68$, $p < 0.001$ (1-tailed). This result indicates that if the actual distances were longer, participants remembered walking longer distances as well.

To examine the influence of the number of turns, position, and action on path distance estimates, a 2 route (Route A vs. Route B) x 2 position (inner vs. outer) x 2 action (before action vs. after action) within-subjects analysis of variance was performed on path distance estimates. There were no significant effects of route, or action. However, there was a main effect of position on path distance estimates. Overall, participants walked significantly longer distances at the outer paths (one path contained 4 turns) than the inner paths (1-turn paths). There was a significant 2-way interaction between route and action; before action, remembered distances were shorter in Route A than in Route B; however after action, remembered distances were larger in Route A than in Route B. This effect was observed because of the influence of number of turns. There was also a significant 3-way interaction between route, position, and action. As in Experiment 1, the follow up analyses indicated that in Route A, after the performance of action the outer path (i.e., P4 contained 4 turns) was remembered as being significantly longer than the inner path (P3 contained 1 turn), $F_{(17)} = 4.09$, $p = 0.05$. In Route B, the reverse was the case; before the performance of action the outer path (P1 contained 4 turns) was remembered as being significantly longer than the inner path (P2 contained 1 turn), $F_{(17)} = 9.41$, $p < 0.01$. This result confirmed the robust effect of number of turns on remembered distances; the inner paths (P2 and P3) were not remembered significantly differently from one another.

Discussion

The fact that participants were exposed to the environments twice in order to acquire route knowledge substantially improved the data collection. Although the rate of exclusion was still high (22%) suggesting that some participants' memories for routes were imprecise, the majority of participants produced the landmarks in the correct order, and therefore distance estimates could be analyzed.

The results replicated those in Experiment 1. As expected, the effect of number of turns was also observed in this experiment; paths with more turns were remembered as being longer than paths with fewer turns. The absence of the influence of action may be due to the salience of the action itself. A more pronounced sequence of movements to perform the dispatch task may make the performance of action more memorable thereby the prediction of a difference between remembered distances before and after the performance of action would stand more of a chance of being found if present.

General Discussion

The new procedure was developed with the aim of controlling confounding factors, such as visual cues and the speed of walk in order to adequately investigate whether action exerts an effect on distance estimation during simulated navigation.

To begin with, in general during tests participants claimed they felt comfortable and relaxed with the task, which indicated that the methodology was an appropriate and sensitive procedure, especially given that participants had to wear a blindfold for the whole duration of the test that lasted about one hour. However, despite the relatively high dropout rate, the data we collected across both experiments indicated nevertheless that the methodology was successful. Future studies could present the environment a third time, which might improve the inclusion rate further.

Let us now consider how our data fit with current theories of environmental knowledge. Our results are in line with the segmentation hypothesis with regard to the perception of the segment lengths and the influence of the number of turns on path distance estimates. However, we found the same effect of number of turns on remembered distances without actually traversing any distance. Our data actually point to an interpretation in terms of attention processes that signal memory for events. Participants heard the metronome clicks representing their footsteps during mental walks. It was clear that they had internalized distance and direction as well as turns information for use during recall that had enabled them to get from the starting landmark towards the final destination. As they were not walking any distance, they seemed to have been encoding the action of turning. In the absence of direct visual information, the body movement triggers the retrieval process; i.e., the

participants' attention would focus on memory for events (actual turning). However, this form of representation is available for limited periods only; as time went on, memory faded and decayed (Thompson, 1983). The attention process then must be shifted in order to attend to the next event that came to mind. To proceed still further, the attention process had to be re-initialized. When walking naturally one average footstep measures about 70 cm, and there are two footsteps forward per second. Therefore, it will take 10 sec to walk 14 m. It is not surprising in terms of the attentional process that people remember only a certain distance (14 m) given that they can focus their attention only for the first 10 sec during retrieval. The fact that participants remembered walking longer distances in paths containing 3 turns than paths containing 1 turn corresponded to the fact that they were actually moving (turning) more often in paths with several turns as well. Consequently, the more turns in a path the more attention shifts were required and the longer the perceived distance. The cognitive mechanism uncovered in the present study is different from that of the segmentation hypothesis. We attributed the fact that paths with more turns were remembered as being longer than paths with fewer turns to the attention shifts during the retrieval process, and suggested that the function of body movement was to re-initialise the retrieval process.

Although the new procedure permits a more precise examination of processes involved in spatial judgment, work needs to be done regarding the large drop out rate. Maybe repeating the simulated walk three times would improve data collection. Additionally, the influence of action at the midpoint would stand more of a chance to be found if present by making the action more pronounced (through more extensive turning or walking on the spot).

More importantly, further work needs to be done in order to establish whether our results can be generalized. For example a comparison between the present study and a study where actual walking takes place is desirable.

Despite these limitations, the new procedure has allowed control over action and visual information during testing, and provides a means for future investigation of a range of possible action manipulations that have hitherto evaded controlled experimental procedures. It also provides important indication that basic processes underlying mental distance estimation seem to persist even in rather extreme sensory deprivation conditions.

References

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences,* **22**, 577 - 609.

Decety, J., Jeannerod, M., and Prablanc, C. (1989). The timing of mentally represented actions. *Behavioural Brain Research,* **34**, 35 - 42.

Gauvin, M., & Rogoff, B. (1986). Influence of the goal on children's exploration and memory of large-scale space. *Developmental Psychology*, **22**, 72 – 77.

Glenberg, A.M. (1997). What memory is for. *Behavioral & Brain Sciences,* **20**, 1 - 19.

Glenberg. A.M., & Kaschak, M.P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, **9**, 558 – 565.

Golledge, R.G. (1987). Environmental cognition. In D. Stokols, I. Altman (Eds). *Handbook of Environmental Psychology*. John Wiley & Son.

Herman, J.F., Norton, L.M., & Klein. C.A. (1986). Children's distance estimates in a large-scale environment. A search for the route angularity effect. *Environment & Behavior*, **18**, 533 – 558.

Hirtle, S.C., Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory & Cognition*, **13**, 208 – 217.

McNamara, T.P. (1986). Mental representations of spatial relations. *Cognitive Psychology,* **18**, 87-121.

Richardson. D.C., Spivey, M.J. & Cheung, J. (2001). *Proceedings of the Twenty-third Annual Meeting of the Cognitive Science Society*, 867 – 872. Erlbaum: Mawhah, NJ.

Rieser, J.J., Pick, Jr., H.L., Ashmead, D.H. & Garing, A.E. (1995). Calibration of human locomotion and models of perceptual-motor organization. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 480 – 497.

Sadalla, E.K., & Magel, S.G. (1980). The perception of traversed distance. *Environment and Behavior*, **12**, 65-79.

Sadalla, E.K., Burroughs, W.J., & Staplin, L.J. (1980). Reference points in spatial cognition. *Journal of Experimental Psychology : Human Learning and Memory*, **6**, 516-528.

Steven, A., & Coupe, P. (1978). Distortion in judged spatial relations. *Cognitive Psychology,* **10**, 422 - 437.

Taylor, H. A. & Naylor, S. J. (2002). Goal-directed effects on processing a spatial environment. Indications from memory and language. In: Coventry, K. R. & Olivier, P. (Eds.). *Spatial Language: Computational and Cognitive Perspectives.* Kluwer, Dordrecht.

Thompson, J.A. (1983). Is continuous visual monitoring necessary in visually guided locomotion? *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 427 - 443.

Thorndyke, P.W. (1981). Distance estimation from cognitive maps. *Cognitive Psychology*, **13**, 526 – 550.

Thorndyke, P.W., & Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology,* **14**, 560-589.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, **24**, 830 – 846.

The Context Dependent Sentence Abstraction model

Matthew Ventura (mventura@memphis.edu)
Department of Psychology/ Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152-3115 USA

Xiangen Hu (xhu@memphis.edu)
Department of Psychology/ Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152-3115 USA

Art Graesser (a-graesser@memphis.edu)
Department of Psychology/ Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152-3115 USA

Max Louwerse (mlouwers@memphis.edu)
Department of Psychology/ Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152-3115 USA

Andrew Olney (aolney@memphis.edu)
Department of Computer Science/ Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152-3115 USA

Abstract

The Context Dependent Sentence Abstraction (CDSA) model and Latent Semantic Analysis (LSA) were compared in their ability to predict sentence similarity. Evidence supports the conclusion that the CDSA model better predicts human ratings for short phrases and sentences than does LSA. Alternative theoretical reasons are given for this finding.

Introduction

Researchers in many disciplines within cognitive science have proposed and tested theoretical claims about the meaning of natural language expressions. One of the contemporary models is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA is a statistical, corpus based technique for representing world knowledge. It computes similarity comparisons between words or documents by capitalizing on the fact that words are similar when they are surrounded by similar words (i.e., the company a word keeps).

LSA takes quantitative information about co-occurrences of words in documents (paragraphs and sentences) and translates this into a K-dimensional space. The input of LSA is a large co-occurrence matrix that specifies the frequency of words in documents. LSA reduces each document and word into a lower dimensional space by using singular value decomposition. This way, the initially extremely large word-by-document co-occurrence matrix is typically reduced to about 300 dimensions. Each word ends up being a K-dimensional vector. The semantic relationship between words can be estimated by taking the cosine (normalized dot product) between two vectors. Although LSA performance has been shown to be impressive at the paragraph level (Foltz, Gilliam, & Kendall, 2000; Landauer, Laham, Rehder, & Schreiner, 1997), other research has found limitations of LSA at the sentence level (Kintsch, 2001). In this paper we will present the Context Dependent Sentence Abstraction (CDSA) model, a corpus-based model that builds sentence meanings based on combinations of pooled adjacent neighbors of individual words. We will first discuss a weakness with vector representational systems (e.g., LSA) in handling sentence comprehension and then turn to a description of the CDSA model, with evidence supporting it.

A weakness with LSA

One major strength of LSA is its versatility and simplicity in handling word meaning and sentence meaning by the use of vector representations. It could be argued, however, that there are potential theoretical problems with combining word vectors to form sentences. For example, the meaning created from a sentence in LSA is a linear combination of word vectors, without eliminating information for any word. Consider the sentence *the cow ate in the field.* In LSA all information about *cows* (e.g., animal, milk, burger), *ate* (e.g., food, grocery, digest), and *field* (e.g., grass, baseball, football) may be included in the sentence representation. It could be argued that this assumption is not theoretically plausible because much of this associated information is not relevant to the word in context. There must be constraints that narrow down the vast array of information that may be "primed" in the first stages of sentence comprehension. Indeed, Kintsch's construction-integration model (1998) has attempted to explain this convergence of activated

information by principles that guide the integration mechanisms.

Whereas the standard use of LSA is based on the assumption that a sentence's meaning is the sum of all the individual word meanings, there are extensions. Kintsch's predication algorithm (2001) tries to build meaning of a sentence by using syntactical information and LSA to create dependencies between subjects, predicates, and objects. For example, consider the sentences *the horse ran* and *the color ran*. The context established by *ran* has different meanings in these two sentences. Therefore, in the predication algorithm, constraints are made on what *ran* means in these sentences. The first step is to find the near neighbors of the word *ran* (i.e., words that give the highest cosine to *ran*). For the *horse* example, all the neighbors of *ran* are compared to the word *horse*. This provides words like *walk, gallop, crawl, rode,* etc. These neighbors of *ran* that are closest to *horse* (i.e., highest cosine) are then included into the vector for the sentence *the horse ran*. The same is done for the color example, resulting in different overall meanings. Including this additional information has been shown to more accurately capture the meaning of a sentence when we consider metaphor and causal inferences (Kintsch, 2001).

Kintsch's predication algorithm (2001) therefore imposes augmentations and constraints on the standard use of LSA. However, this algorithm still may not go the distance in solving the problem of information overload mentioned earlier. That is, predicating the verb *ate* to *cow* does give relevant information like *graze*, but all information about *cows* and *ate* are also included. To successfully implement context in the given example, we would want to include only information about "*cows eating"*, not about "*cows* and *ate* and *graze* and *field* and *pasture"*. While the predication algorithm solves some problems by adding information, it also may be limited by not taking any information away.

The need for contextual constraints

Computational representations like LSA go beyond general word meanings, but may not adequately handle contextual constraints. LSA may go some distance in handling proposition meanings that constrain words in context (Kintsch, 1998), but there still is a large landscape of representations and algorithms for combining information from words. We propose a new way of implementing contextual constraints. These contextual constraints are first built from simple individual word meanings that get established over time from their occurrences in the environment. But as sentences are constructed, similarities between the words in the constrained construction build a new meaning different from the sum of its parts.

The CDSA Model

Associationist frameworks (Landauer, 2002; Louwerse & Ventura, in press; Smith, Jones, & Landau, 1992) assume that it is critically important to measure and model the correlations between occurrences or events in the environment. We pursued a corpus-based model of word and sentence meaning, called the Context Dependent Sentence Abstraction (CDSA) model. In the CDSA model, semantic information within any word *w* is the pooled words that co-occur with word *w* in every context. One of the goals of this model is to try and capture the associations between words under a new level of specificity that considers the pool of their surrounding words.

In order to implement this model, it was necessary to make decisions about the learning rule and training set to be used. For this model, the deciding factor in each of these cases was psychological plausibility. That is, this model considers a corpus of prior experiences with words in context and the theoretical weights between words that change with experience, as opposed to a priori sets of features that are dictated by a brittle, symbolic model. The central question is how these weights change with experience. The proposed CDSA claims that they change by accumulating specific sentence exemplars.

Consider two words *chair* and *table*. The central question to be asked is what are all the possible relevant or useful relations that can exist between these two concepts? Each word has a neighborhood set that includes all words that co-occur with the target word. These words are the extensional meaning of the target word and serve as the basis for all associations. The neighborhood intersection is the relation that occurs when two words share similar co-occurrences with other words. Much like LSA, words become associated by their occurrence with many of the same words. For example, *food* and *eat* may become associated because they both occur with words such as *hungry* and *table*. Therefore the neighborhood set ***N*** for any word *w* is all the information we have in the exemplars for a word.

Neighbor weights

The neighborhood set for any word is intended to represent the meaning of a word from a corpus. But there were several theoretical challenges that arose when we developed the model. One dealt with how to differentially weight neighborhood words. We assigned *neighborhood weights* to each neighborhood word *n* of word *w* according to Equation (1).

$$\lambda_{n|w} = \frac{f(n \mid w)}{\sqrt{f(w)f(n)}} \quad (1)$$

The expression $f(n|w)$ designates the frequency of occurrence of the neighbor word *n* to target word *w*, whereas $f(n)$ is the total frequency of the neighbor word *n*, and $f(w)$ is the total frequency of the target word *w*. This formula essentially restricts the weights for the neighbor words as being between 0 and 1 in most cases. We adopted this simple assumption but we acknowledge that there are

other ways to guarantee the range of the weights being within 0 and 1.

Therefore, the weighting function was aimed at giving more importance to words that consistently co-occur and less importance to words that occur frequently in the corpus. Additionally, rare co-occurrences may be given low weights because they do not consistently co-occur with the target word.

Some important assumptions had to be made in order to build relevant associations to target words most effectively. The next section will explain the procedures of the algorithm written to perform these operations.

Neighborhood Intersection Algorithm

In order to construct the neighborhood set for any word, an algorithm was written that pooled all words ***N*** that co-occurred with the target word *w*. We used the Touchstone Applied Science Associates (TASA) corpus because of its size (750,000 sentences) and diversity of topics (reading a diversity of texts up to college level). Each sentence in the corpus served as the context for direct co-occurrence. So for entire set of sentence sentences ($s_1...s_C$) that target word *w* occurs in, every unique word in ($s_1...s_C$) is pooled into the neighborhood set ***N***. For example the neighborhood of *chair* may consist of: *table, sit, leg, baby, kitchen, talk*, etc. This represents the neighborhood ***N*** of each target word *w*. Each word in the set ($n_1...n_K$) of ***N*** is weighted by the function described in equation (1). To evaluate the relation between any two words w_1 and w_2, we follow the following algorithmic procedure:

1. Pool neighborhood sets for w_1 and w_2 ($\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ respectively), computing the weights for all the neighbor words using Equation (1).
2. Calculate neighborhood intersection as follows:

$$\frac{\sum_{n \in N_1 \cap N_2} \left(\lambda_{n|w_1} + \lambda_{n|w_2} \right)}{\sum_{n \in N_1 \cup N_2} \left(\lambda_{n|w_1} + \lambda_{n|w_2} \right)} \qquad (2)$$

The numerator is the summation of weights over the intersection of the neighborhood sets ($\boldsymbol{N}_1$ and $\boldsymbol{N}_2$) whereas the denominator is the summation of weights over the union of the two neighborhood sets. This formula produces a value between 0 and 1.

In the next section we will discuss how the CDSA model was evaluated.

CDSA Model Evaluation

In four experiments we evaluated the CDSA model against LSA and human raters. The estimations of word and sentence meanings in the CDSA model and LSA were trained on the TASA corpus. Ratings in all four experiments were made by 10 undergraduate psychology students who were instructed to rate the similarity of various pairs of words (i.e., primarily from words from Spellman, Holyoak, & Morrison, 2001) on a 6-point scale that varied from 1 (very unrelated) to 6 (very related). A rating of 1 or 2 meant the rater could not easily find a functional or physical relationship between the word pairs (e.g. fish-office). The mean among the raters for each pair was taken as the basic data to test the models.

Experiment 1

Word Pairs A total of 64 word pairs was constructed that had a frequency over 10 in the TASA corpus. Some of the words were expected to be unrelated (e.g., chair-hear) and some related (e.g., chair-sit) in order to provide a sensitive range of values.

Results and Discussion

Human ratings (M = 3.57, SD = 2.20) were significantly correlated with the values produced by the CDSA model, $r = .71$, $p < .001$, and with LSA cosines, $r = .78$, $p < .001$. So both models fared quite well in accounting for the ratings of word pairs.

Neighborhood intersection estimation shared a relation to human ratings, so we might conclude that this type of association between words is used in human judgments. That is, by using all the co-occurrence information about a word, one can capture the meaning of a word. As can be seen, LSA was slightly more predictive of word relations than the CDSA model, although the difference was not statistically significant.

The lack of difference between models may be due to the construction of neighborhood sets for a single word in the CDSA model. Since there are many neighbors that exist for any particular word, there are many degrees of freedom that exist for determining the meaning for a single word. For instance, if one is asked to give an association to the word *cow*, there are many possible associations (e.g., *animal, milk, burger*, etc), which will lead to a very general non-specific representation of a single word.

The purpose of Experiment 2 is to try to use the model to represent the meaning of word-pairs. This involves imposing constraints on the neighbors for each pair in order to more accurately represent the contextual meaning of the pair. For instance, *cow-graze* should give a more specific representation of *cow* than *cow* without a context because constraints are built on the meaning of *cow*. These constraints initially involve measuring the neighborhood overlap between the neighbors of *cow* and the neighbors of *graze*, which then are used to compare to another set of information (e.g., word, sentence).

Experiment 2

A central theoretical assumption in Experiment 1 was the idea that neighborhood intersection plays a prominent role in the relation between words. But how can the current

model account for conceptual relationships beyond the word level? Figure 1 gives an illustration of how this could be done. If two pairs are being compared, the neighborhood overlap of each pair is pooled into F_1 and F_2. Then the intersection (Equation 2) is calculated to access the similarity between the two pairs. This constrains the degrees of freedom for the pair, which eliminates any information that is not mutually shared by both words in the pair (i.e., the problem found in Kintsch's predication algorithm). Therefore, each word is always dependent on the context in which it appears. As the context for a word becomes more specific (i.e., as reflected by the number of unique words it appears with), the less likely that the same context will be associated with any random word. For instance, *chair-sit* has a smaller neighborhood set than the sum of neighbors for *chair* and the neighbors for *sit*. This assumption therefore states that word pairs, or even sentences, are different than the sum of its parts, an assumption quite different from current models of associative learning like LSA.

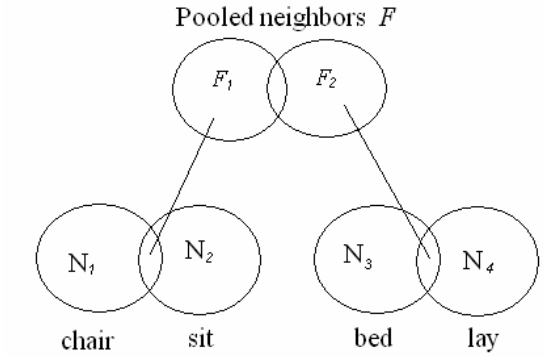


Figure 1: The recursive nature of neighborhood overlap. The neighborhood overlap (F_1) of *chair* (N_1) and *sit* (N_2) is intersected with to the neighborhood overlap (F_2) of *bed* (N_3) and *lay* (N_4).

Model modifications

Neighborhoods were first built on words within the pair as described in equation 1. As described in Figure 1, the neighborhood N_1 of *chair* is intersected with the neighborhood N_2 of *sit* to yield a new neighborhood F_1 that represents the "chair sit" neighborhood. Since any shared neighbor in N_1 and N_2 each have a separate weight, the average of the two weights (Equation (3)) is calculated to represent the new weight for each neighbor in F_1.

$$\lambda_{n|w_1 w_2} = \frac{1}{2}\left[\lambda_{n|w_1} + \lambda_{n|w_2}\right] \quad (3)$$

In the same manner, we obtain $\mathbf{F_2}$. Once F_1 and F_2 have been calculated for both word-pairs, the neighborhood intersection is calculated (as described in Equation (2)), to access the relationship between the 2-word pairs.

Additionally the union of the neighborhood weights (i.e., the entire neighbor weights of all words in each pair) was calculated for F to compare the effectiveness of intersecting the neighborhoods.

Word Pairs We constructed 53 word pairs that had a frequency over 10 in the TASA corpus. Separate sets of pairs were intended to be unrelated (e.g., bear/cave—pen/write), related by analogy (e.g., bear/cave—fish/pond), or related by both analogy and semantic relation (e.g., teeth/bite—leg/kick).

Results and Discussion

Human ratings (M = 3.46, SD = 1.62) were significantly correlated with CDSA intersection, r = .60, p < .001, and union, r = .51, p < .001. LSA cosines were also related to human similarity ratings, r = .64, p < .001.

It appears that imposing context reduced the correlation with rated similarity of 2-word pairs, compared with single-word pairs. As can be seen LSA performance also drops. Most notably, the union of neighbor sets does not perform as well as the intersection version of the CDSA model.

The purpose of Experiment 3 was to examine how performance would be affected by implementing more context through comparison of 3-word phrases.

Experiment 3

The process we used in building constraints on three-word combinations involves a multinomial neighborhood overlap (N-O) among all neighborhood pairs. Each neighbor that is shared by at least two neighborhoods is then pooled into F. Figure 2 gives an illustration of how this can be achieved..

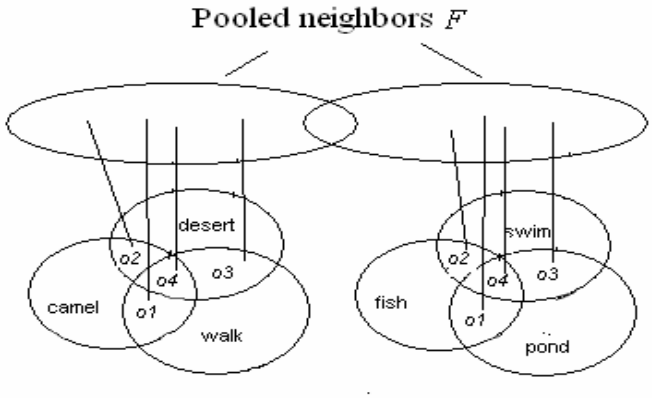


Figure 2: 3-word neighborhood overlap for two 3-word combinations. The neighbors of each combination are compared for neighborhood overlap, which are then pooled into a neighborhood F for each pair.

Model Modifications

Equation 3 is used to compute weights for the words that are in the intersection of any two neighborhoods (o_1, o_2, and o_3).

By making all possible intersections between each neighborhood N_1, N_2, and N_3, a select number of words may be counted three times. Therefore neighbors shared by all three sets (o_4; see figure 2) will be averaged (i.e., divided by 3) and eliminated from any other N-O to avoid multiple counts. The computation for the weights in the intersection of the three neighborhoods is simply an extension of equation 3, where the average is taken with three weights instead of two. Additionally each neighbor in o_4 is a special neighbor because it is shared by all neighbor sets. Therefore these neighbors are multiplied by a constant of 3 (i.e., since there are three sets) to give greater importance to these context bound neighbors. Once F_1 and F_2 have been pooled together by all the N-O for each pair, the neighborhood intersection is calculated (Equation 2), to access the relationship between the 3-word pairs.

Additionally the union of the neighborhood weights (i.e., the entire neighbor weights of all words in each pair) was calculated for F to compare the effectiveness of intersecting the neighborhoods.

Short phrases We constructed 58 three-word phrases that had a frequency over 10 in the TASA corpus. Some pairs were intended to be unrelated (e.g., bird/nest/fly—brush/paint/art), related by analogy-like relations (e.g., gun/shot/bullet — axe/chop/wood), and related by both analogy and semantic relation (e.g., dog/loud/bark—cat/quiet/meow).

Results and Discussion

Human ratings (M = 3.10, SD = 1.78) were significantly related to the CDSA intersection, $r = .63$, $p < .001$, and union, $r = .44$, $p < .001$. LSA cosines were related to human similarity ratings, $r = .47$, $p < .001$. The results give evidence that imposing context improves performance in calculating similarity. Furthermore, LSA performance continues to drop as more word context is introduced. This in part could be due to the lack of constraints that are put in the sentence representation in LSA.

Experiment 4

The purpose of the present experiment is to test the CDSA model to sentences of varying lengths (i.e., sentences ranging from 4 to 6 words). One challenge that arises in calculating sentence similarities is how to handle all the possible intersections between word neighbors within one sentence. Therefore three conditions were tested on how to calculate the final sentence neighbor set F. First, a multinomial approach entailing N-O among all neighborhood sets was pooled to get F. Weightings were computed between any N-O words as an extension of Equation 3, where the neighborhood intersections could entail 2-6 neighborhoods.

Second, a word-chunking maximum likelihood approach was used that calculated a set P for every three words in a sentence (Johansson, 2000). This chunking approach using a 3-word context to any target word was found to give equal performance to 5-word and 7-word contexts in syntactic tagging. So if a sentence had five words, a multinomial N-O calculation between the 1st 2nd and 3rd word neighborhoods would produce P_1 (i.e., as described in Experiment 3), then N-O would be calculated between the 4th and 5th neighborhood words to produce P_2 (as described in Experiment 2). Then the N-O between P_1 and P_2 would be calculated to produce the final neighborhood F for the sentence. The intersection between F_1 and F_2 (Equation 2) will give the final similarity between the two sentences. This word-chunking hypothesis is consistent with the intuition that adjacent words in a sentence constrain meaning more than nonadjacent words in a sentence.

Finally, the union of the neighborhood weights (i.e., the entire neighbor weights of all words in each pair) was calculated for F to compare the effectiveness of intersecting the neighborhoods.

Sentences We constructed 42 sentences whose words had a frequency over 10 in the TASA corpus. Sentences pairs were constructed of varying length (e.g., *blue bird fed babies nest tree -- bear protected cubs den*; articles, pronouns and prepositions were removed). Sentences were constructed so that about half were considered related and half unrelated.

Results and Discussion

Human ratings (M = 2.12, SD = 1.48) were significantly correlated to CDSA model 3 word chunking intersection, $r = .69$, $p < .001$, CDSA union, $r = .65$, $p < .001$, and the CDSA multinomial intersection, $r = .56$, $p < .001$. LSA cosines were also related to human similarity ratings, $r = .50$, $p < .001$.

The results give evidence that imposing context may be important when calculating sentence similarity. By applying an arbitrary rule set to sentences of varying lengths seems to yield better performance than just making all possible intersections among neighbors. Alternatively, the union of all the neighbors seems to perform just as well as a rule based intersection procedure. Possible reasons for this will be discussed next.

General Discussion

In sentence comprehension, comprehenders must understand the nature of word context and the constraints one word places on another (Kintsch, 1998). In other words, comprehenders will have to ask themselves: how does the meaning of one word affect the meaning of another word? The most straightforward relationship among words is an additive one, where the meaning of one word has no influence on the meaning of another word. In contrast, in the case of sentence comprehension, the levels of a one word can dramatically change the effects of another word. In this model, context refers to N-O among words in a sentence. That is, changing levels of one word can dramatically affect the meaning of another word. Thus, without structural constraints involving processes similar to

N-O, sentence meanings proceed in a radically different manner. Many relations shared between the pairs in the 4 experiments were abstract relations, ones that were only clearly established by filtering the individual word meanings and keeping shared information among words.

The word-chunking N-O approach appears to perform better than the multinomial N-O approach among all neighbors. Making all possible intersections among neighbors does not seem to be very psychologically plausible since it would involve making many comparisons between words that may not be relevant. For instance, comparing the first word to the last word in a sentence may not be important in evaluating the meaning of a sentence.

Possible improvements

As can be seen in experiment 4, N-O did not seem to help predict sentence similarity to a great extent over the union of all the neighborhoods in a sentence. This may be due to the arbitrariness of the rules used to calculate N-O for varying sentence lengths. For instance, if the sentence was 6 words long, N-O would be calculated for the three words and the last three words. With these two pools we would then calculate ***F***. This type of rule makes the assumption that all 6-word sentences follow the same syntactic structure. This obviously will not do for all 6-word sentences. Therefore, it seems likely that if the CDSA model was implemented with a syntactic parsing mechanism, it could give the correct word pairs to calculate N-O for any sentence.

Conclusion

The computational model presented here captures both word and sentence meaning. There are several reasons why using the CDSA model is advantageous. First, it uses simple mechanisms that are psychologically plausible. Second, it gives the freedom to add more information to the corpus at any time. Since the measures derived are computed on-line on the corpus, dynamically adding text to the corpus is not a problem. Essentially, many weights are changed between words as soon as text is added.

The proposed computational model captures word and sentence meaning by appealing to constraints reflected in a corpus analysis. Embodiment theorists (Glenberg & Robertson, 2000) may claim that there is no meaning derived from a corpus analysis because the words are not grounded in sensory-motor experience. In principle, one could have a more grounded corpus with units extensively embedded in sensory and motor experience. The TASA corpus was simply readily available. Whether the episodic experiences are reflected in TASA or in sensory-motor experience, the theoretical assumptions of the CDSA model are that, specific exemplars and associative processes are sufficient to account for the judgments of meaning similarity. The CDSA model uses simple mechanisms that rely on co-occurrences of words in exemplars.

One additional advantage of the CDSA model is that it allows more information to be added to the corpus at any time. Since the measures derived are computed on-line on the corpus, dynamically adding text to the corpus is not a problem. Essentially, weights are changed between words as soon as text is added.

Acknowledgments

This research was supported by the National Science Foundation (REC 0106965, ITR 0325428) and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF.

The TASA corpus used was generously provided by Touchtone Applied Science Associates, Newburg, NY, who developed it for data on which to base their Educators Word Frequency Guide.

References

Foltz, P.W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111-127.

Johansson, C. (2000). A Context Sensitive Maximum Likelihood Approach to Chunking. In: *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal.

Kintsch, W. (1998) *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. (2001) Predication. *Cognitive Science* 25, 173-202.

Landauer, T. K., & Dumais, S. T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E., (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.), *The psychology of learning and motivation,* 41, 43-84.

Louwerse, M.M. & Ventura, M. (in press). How children learn the meaning of words and how computers do it (too). *Journal of the Learning Sciences*.

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition*, 60, 143-171.

Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory & Cognition, 29*, 383-393.

Similarity and Taxonomy in Categorization

Timothy Verbeemen (timothy.verbeemen@psy.kuleuven.ac.be)
Departement Psychologie, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Gert Storms (gert.storms@psy.kuleuven.ac.be)
Departement Psychologie, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Tom Verguts (tom.verguts@ugent.be)
Vakgroep Experimentele Psychologie, Ghent University, H. Dunantlaan 2
B-9000 Ghent, Belgium

Abstract

In this paper, a two by three approach to modeling categorization is presented. Similarity representations based upon a geometric, an additive tree and an additive cluster model are combined with an exemplar model and a prototype model in a single approach. The six models are applied to the categorization of pictorial known and unknown fruits and vegetables (Smits et al., 2002). For novel stimuli, the geometric exemplar model and the cluster models gave the best account, indicating a strategy where people compare stimuli with stored members on more general continua or a limited set of features. For well-known stimuli, the tree-based models gave the best account of the data, suggesting more elaborate taxonomic knowledge. More generally, the results show that different categorization models may perform better for different sets of stimuli, and that a systematic empirical comparison of such models is needed.

Introduction

A major contribution of categorization research over the last decades has been to establish the relation between similarity and categorization. Rosch and Mervis' (1975) seminal paper on the graded structure of categories showed that categories are ill-defined, and that the extent to which an instance of a category is seen as a typical member is positively related to similarity towards the category in question and inversely related to similarity towards relevant contrast categories (e.g., Verbeemen et al., 2001). Given the importance of similarity in categorization, a formal model should take a clear stance on two issues: The nature of similarity computation and the relevant objects of comparison in this calculation. First, the model must make assumptions about the nature of similarity, especially when the structure of the stimuli under investigation is not experimentally controllable. There are two main approaches to similarity, geometric and feature-based. The geometric approach (e.g., Carroll & Arabie, 1980; Shepard, 1964) represents stimuli in abstract space where similarity is inversely related to the distance between stimuli. In the feature-based approach (e.g. Shepard & Arabie, 1979; Tversky, 1977), similarity is considered a function of feature overlap, where commonalities increase and differences decrease overall similarity. Second, a model should specify the objects used in similarity calculation. In particular, information employed in making category decisions may be stored at the category level, or it may be stored at the level of individual instances of a category. The former approach is known as the prototype view (e.g., Hampton, 1979; Smith & Minda, 1998), and the latter as the exemplar view (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). In this paper, we argue for a systematic evaluation of these formal models in a two by three approach that compares prototype and exemplar models on the one hand, and geometric and feature representations on the other hand.

The Generalized Context Model and a Geometric Prototype Model

In the generalized context model (GCM; Nosofsky, 1984, 1986, 1992), an exemplar model, categorization is assumed to be a function of similarity towards all relevant stored exemplars. In case (physical) dimensions are unavailable, the GCM fitting procedure starts with a multidimensional scaling procedure (MDS; Borg & Groenen, 1997) on proximity measures of all stimuli involved. The coordinates of these stimuli are then used as input for the model. In the case of two categories, A and B, the probability that stimulus x is classified in category A is given by:

$$\mathrm{P(A \mid X)} = \frac{\beta_A \eta_{XA}}{\beta_A \eta_{XA} + (1-\beta_A)\eta_{XB}} \qquad (1)$$

where β_A lies between 0 and 1 and serves as a response bias parameter towards category A. The parameters η_{XA} and η_{XB} denote the similarity measures of stimulus x toward all stored exemplars of category A and B, respectively:

$$\eta_{XA} = \sum_{j \in A} \exp\left[-c \left(\sum_{k=1}^{D} w_k \left| y_{xk} - y_{jk} \right|^r \right)^{1/r} \right] \qquad (2)$$

with y_{xk} and y_{jk} as the coordinates of stimulus x and the j-th stored exemplar of category A (or B for η_{XB}, respectively) on dimension k. The weight of the k-th dimension is denoted by w_k, with all weights restricted to sum to 1. The power

metric, determined by the value of r, is usually given a value of either 1 or 2, corresponding to city-block and Euclidean distance, respectively.

A prototype model can be constructed with the GCM as a start (Nosofsky, 1986, 1987, 1992; Smits et al., 2002). With the prototype defined as the central tendency of a category (Malt & Johnson, 1992; Malt & Smith, 1984; Rosch & Mervis, 1975), the object created by taking, on each dimension, the average coordinate over all members of the category, is a good way to define a prototype. The similarity function changes to:

$$\eta_{XA} = \exp-\left[c\left(\sum_{k=1}^{D} w_k \left| y_{xk} - \overline{y_{.k}} \right|^r \right)^{1/r} \right] \quad (3)$$

where $\overline{y_{.k}}$ denotes the mean value of all stored members of category A on dimension k. We will refer to (3) in combination with (1) as the Geometric Prototype Model (GPT).

A number of studies have already been conducted that compared prototype and exemplar models (e.g., Nosofsky, 1992; Nosofsky & Zaki, 2002; Smith & Minda, 1998, 2000; Smits et al., 2002). In many, the GCM performed better than prototype models. In the next section we elaborate on the major alternative to geometric similarity models, the contrast model (Tversky, 1977).

The Contrast Model and Categorization

In the contrast model, similarity between two stimuli is defined as a function of the features that these stimuli possess:

$$Sim(a,b) = \theta g(A \cap B) - \alpha f(A - B) - \beta f(B - A) \quad (4)$$

where $g(A \cap B)$ is a function of the features shared by objects a and b (the *common features*), and $f(A-B)$ and $f(B-A)$ are functions of the features that belong to one stimulus but not the other (the *distinctive features*). Different models have been proposed, mostly focused on either the common feature component or the distinctive feature components.

Pruzansky, Tversky and Carroll (1982) reanalyzed 20 data sets taken from various published studies, divided into two groups depending on the hypothesized structure of the stimuli used: conceptual (e.g., vegetables) and perceptual (e.g., polygons) stimuli. For 10 out of 11 studies of conceptual stimuli, analyses of proximity data proved better when performed by ADDTREE, a distinctive features approach to similarity. Seven out of nine studies of perceptual stimuli showed a clear advantage for low-dimensional MDS solutions.

A number of studies have been conducted that compared geometric and featural exemplar models. Lee and Navarro (2001) used additive clustering to extract common features from similarity data and provided excellent accounts of an artificial learning experiment with ALCOVE (Kruschke, 1992). Takane and Shibayama (1992) analyzed identification data of digits taken from Keren and Baggen (1981) and they too obtained excellent results for a featural version of the similarity-choice model (Luce, 1962) based on ADDTREE (Corter, 1982; Sattath & Tversky, 1977). Whereas clustering provides a very flexible way of representing similarity, allowing for overlapping clusters, additive trees are more restrictive in that they impose a hierarchy. There are, however, reasons to apply tree models, especially in the case of conceptual knowledge. A tree model produces, in general, a higher amount of features for the total set than additive clustering. But the amount of shared features is lower in general, and most weight is given to idiosyncratic features. This may be appropriate for well-known stimuli (McCrae & Cree, 2002), as people can be expected to have a fair amount of background knowledge about these stimuli, but it is unclear whether this is plausible in the case of novel stimuli.

The implementation of the feature structure in the GCM yields the *featural exemplar model* with similarities as:

$$\eta_{XA} = \sum_{j \in A} \exp-\left[c\left(\sum_{k=1}^{F} w_k \left(y_{xk}(1 - y_{jk}) + y_{jk}(1 - y_{xk}) \right) \right) \right] \quad (5)$$

where y_{jk} = 1 if stimulus j has feature k and y_{jk} = 0 otherwise. Therefore, the term y_{xk} (1-y_{jk}) is 1 if and only if the target stimulus x possesses the feature and the "reference" stimulus j does not, and vice versa for y_{jk} (1-y_{xk}). Each feature has a weight w_k that corresponds to the length of the segments in the tree. We will refer to this model as GCM-F (generalized context model – featural).

The *featural prototype model* will be illustrated using Figure 1, for an additive tree solution for birds and mammals. Distances between objects are defined by the sum of the horizontal segments on the shortest path between two stimuli (vertical segments are added for visual ease only). Each segment represents a feature that applies to all of its children with more general features closer to the left ("root") and more specific features located towards the right (endpoints) of the tree. The model is again formally similar to the featural GCM with the prototype treated as a pseudo-exemplar. The distance function equals:

$$\eta_{XA} = \exp-\left[c\left(\sum_{k=1}^{F} w_k \left(y_{xk}(1 - y_{pk}) + freq_{pk} y_{pk}(1 - y_{xk}) \right) \right) \right] \quad (6)$$

where y_{pk} is 1 if the prototype of A has the feature, and 0 otherwise. The frequency weighting term corresponds to the relative frequency or proportion with which the feature

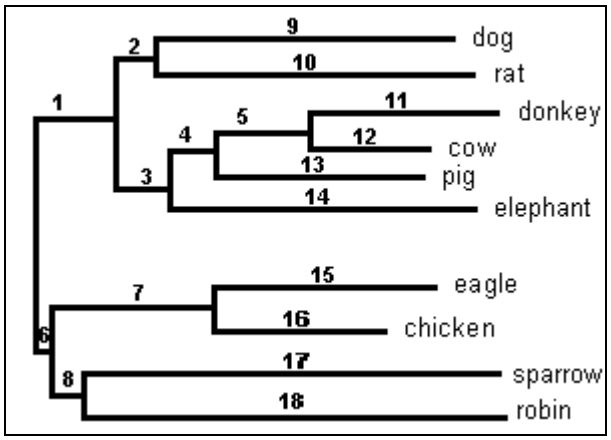


Figure 1: Example of a rooted additive tree.

occurs within *A*, i.e. the proportion of stored members of category *A* that possess that particular feature. This corresponds to the idea that the impact of the features in the prototype, which is seen as a pseudo-exemplar, depends on the prevalence of those features in the category[1]. We will refer to this model as FPT (featural prototype model).

The application of the featural models to an additive clustering solution, i.e. a common features model, is straightforward, as the feature structure is defined. This is not the case for additive trees as they produce distinctive features: a feature that adds to the difference between two stimuli may belong to one or the other. This is not a problem for exemplar models as distances between objects remain unchanged, but it will be required for a prototype model. To define a particular structure, one needs to define a root. If the root in Figure 1 is placed anywhere else it would imply that some members of one category possess some of the most general features of the other category but share none of the features belonging to members of their proper category. This is implausible as it would imply that some stimuli are seen as members of a category on a purely idiosyncratic basis and not because they share any features with that category, even though these stimuli would possess the most general features of a related contrast category. Therefore, in the remainder of this article, we will assume that the root is placed on the segment or path that best approximates this linearly separable structure (Medin & Schwanenflugel, 1981) for *stored* (well known) categories[2]. (This implies that one has to decide, *a priori*, which objects are considered to be stored members of a given category, as is the case for all other models. The choice of a root that best approximates separability serves to define the feature structure and not category membership for stored items, which is already determined.)

Analysis of Smits et al.'s data

Smits et al. (2002) analyzed a stimulus set consisting of pictures of 79 well-known items, retained after an exemplar *generation* task for the categories *fruit* and *vegetables*, and 30 fruits or vegetables, mostly exotic, that were completely unknown to participants. Ten participants completed a feature applicability task for all stimuli, for the 17 most frequently generated features for *fruit* and *vegetables,* generated by a different group of thirty participants. (Taking the most frequently generated features ensures that the analysis is not clouded by potentially unreliable features that are important to only a few subjects.) A similarity matrix was then obtained by correlating the feature applicability vectors for all 109 stimuli, after summing over participants. A different group of thirty participants classified the well-known stimuli as belonging to either *fruit* or *vegetables*. A group of twenty different participants did the same for the novel stimuli. Smits et al. then predicted category decisions based on the geometric versions of the GCM and the GPT and found a clear advantage of the GCM over the prototype model. Since their data from the categorization task had a fair amount of variance in the categorization proportions even for the well-known stimuli, it is possible to fit the respective models to old and novel stimuli separately. Therefore, we will analyze the data in a 2 × 3 × 2 framework, where the last factor is added to assess the fit of the models for old and novel stimuli separately.

Generating Similarity Representations

In order to obtain dimensions, similarities between 109 old and novel fruits and vegetables were reanalyzed with ALSCAL (Takane, Young & De Leeuw, 1977), using the BIC criterion (Schwarz, 1978)[3] to determine the optimal

[1] As the similarity structure was derived from the presented stimuli, we assume that a *presented* exemplar has all of its features to the full extent. Because a prototype is a construction after encountering exemplars, we assume that the activation and impact of its features is dependent on the frequency of those features in its own category (Kellogg, 1980). Because the prototype is treated as a pseudo-exemplar, we assume that its features can be no more than fully active (in the case of a feature that applies to all of its members), resulting in a factor of 1. We assumed that the features of *stored* exemplars in the *exemplar model* were not weighted dependant on the frequency of occurrence in *other* exemplars. This was confirmed post hoc: fit values were much worse when weighted for frequency.

[2] ADDTREE starts by grouping together the closest pair of objects, and then creates a dummy object with the average of the distance of the original objects to all other objects. This procedure is repeated until there are three objects left, and the root is placed so as to minimize the variance to the last three objects. Here we minimize the variance on the path that provides the best linear separation for *well-known* stimuli in a similar way.

[3] BIC = -2 ln(L) + k ln(n), where L is the likelihood value (the probability of the data given a certain model), k is the number of free parameters, and n is the number of data points. *Lower* means better, and only the difference in free parameters needs to be taken into account. The first term decreases with increasing model fit, the second is a penalty term that increases with the number of free parameters and data size. As such, the measure is a trade-off between model fit and model complexity. The statistic is most appropriate when the information provided by the data is relatively large as compared to any prior information, as is the case in all the

dimensionality (Lee, 2001b). A three-dimensional solution was chosen that explains 96 percent of the variance.

Following the same procedure for additive clustering, an analysis with ADCLUSGROW (Lee, 2001a) resulted in 32 clusters that explained 96 percent of the variance.

Finally, the same similarity matrix was reanalyzed using ADDTREE/P (Corter, 1982). The explained variance was 84 percent. The algorithm does not readily lend itself to the BIC-guided approach and usually fits to maximum complexity, in this case with 209 arcs (features). To the extent that this may cause the fitting of error, it may cause a drawback for the categorization models as the error would be "plugged" into the model, clouding the explanatory power of the underlying feature structure. At first sight this, and the lower fit value, would indicate that these models are less appropriate.

(It is important to note that these analyses were based on correlation patterns and not on the rough feature vectors, so the similarity algorithms, especially in the featural approach, are in no way restricted to have as much or less features than the original feature vectors.)

Fitting the Similarity Representations to the Categorization Models

The geometric models were fitted with the Euclidean distance metric ($r = 2$), as this resulted in clearly better fit values. The GCM and GPT were fitted as discussed previously with four free parameters: the bias parameter β, the sensitivity parameter c, and two dimension weights, as weights are restricted to add to 1. Feature weights for the tree- and cluster-based models were taken from the original solutions, however, to keep the number of parameters feasible for estimation, hence there are two free parameters, β and c. Stored members are the same in all models and are based on the earlier exemplar generation task. (Note that the tree-based models were based on the original ADDTREE solution after placing the root so as to provide the best linear separation, for known stimuli, between fruit and vegetables. Compared to the actual *generation* task for well-known stimuli, only one item, rhubarb, was generated as a fruit but closer to vegetables according to the ADDTREE solution.[4])

Results and Discussion All models were fitted by maximizing the binomial likelihood. Correlations between predicted and observed category decisions ranged from .85 to .93 with the best performing models ≥ .92, indicating a fair but not perfect amount of explained variance. The models were evaluated using BIC. Results are summarized in Table 1.

Analyses of the 30 novel stimuli separately are presented in the first panel of Table 1. The lowest (best) BIC value was obtained for the GCM, but the difference with the cluster-based exemplar and prototype models is small. All ADDTREE-based models performed clearly worse. The current results do not clearly favor exemplar or prototype models for novel stimuli, but it appears that the geometric approach to prototypes provides less explanatory power as compared to the clustering approach.

Table 1: -ln(L)[5] and BIC (only the difference in parameters taken into account) for the category *fruit* for all models.

	MDS		ADCLUSGROW		ADDTREE/P	
	GCM	GPT	GCM-F	FPT	GCM-F	FPT
1.*New*						
-ln(L)	73.95	83.27	81.38	82.12	90.83	93.01
BIC	160.69	179.33	162.76	164.24	181.66	186.02
2.*Well-known*						
-ln(L)	351.75	405.29	364.87	388.51	334.75	330.18
BIC	719.04	826.12	729.74	777.02	669.50	660.36

Analyses of the well-known stimuli resulted in a very different pattern. BIC values for the analyses of the 79 well-known stimuli separately are presented in the second panel of Table 1. The BIC values for the geometric and the cluster-based models were clearly higher (worse) than for the tree models. Clearly, the data from the well-known subset is best accounted for by the tree-based models that assume more elaborate taxonomic knowledge. The difference between the exemplar and prototype model is rather small and should be interpreted with caution. This result appears to contradict the earlier fit values where the MDS and additive clustering solutions provided a substantially better fit to the *similarity* data. In fact, a better fit to similarity data need not imply a better fit of the categorization model: those aspects of stimuli that are activated in a similarity task may very well be different from what is activated in a categorization task, especially after a concept has become well-elaborated. In other words, the less flexible and hierarchic structure of trees may not have captured all aspects of similarity, but the aspects it did capture may be more relevant for categorization of well-known concepts. Indeed, every aspect of similarity that is not used in categorization can be considered error in the model.

In fact, the most interesting pattern that emerges from these data is the fact that categorization of novel stimuli is best explained by those models that are based on the flexible representations that best explain similarity. These models have either a limited number of dimensions or a limited number of features, with little idiosyncratic features in the

analyses presented here. It is also fit to compare nonnested models. (For an extensive discussion, see Kass & Raftery, 1995.)

[4] In the actual classification task (not the *generation* task), the proportion of classifications for *rhubarb* as a fruit was only .33.

[5] This value is the most "democratic" measure as it only incorporates model fit, (incorrectly) disregarding the penalty term for free parameters. The measure is equal to the sum of minus the log likelihoods of the individual data points and is therefore sensitive to the size of the data set; hence differences in fit *between* the two datasets are *not* directly interpretable (the same is true for the BIC measure).

latter case. Categorization of well-known stimuli, on the other hand, is best explained by the models that use a representation that is less close to the similarity data but that impose a more elaborate taxonomy and more idiosyncratic features.

An interesting interpretative property of additive trees in this respect is the fact that, at each node of the tree, a feature that applies to all of its children is linked to a limited number of alternatives. Second, branches in the tree tend to have a higher weight as one goes down in the tree. This implies that the number and weight of commonalities decreases with the number of nodes between stimuli. It also means that those features that add most weight to the difference are less likely to be related as the number of nodes increases and vice versa. A similar argument was made by Markman & Gentner (1993) who presented stimuli with different ontological distances and found a similar pattern when subjects listed commonalities and alignable and nonalignable differences. A possible explanation for the good results of tree-based models could be that, as a concept becomes more elaborated, people tend to gravitate to an alignable structure that might dominate other, presumably less alignable, aspects of similarity.

Conclusion

The goal of the present paper was two-fold. First we presented a general framework, in which different models (i.e., exemplar and prototype models, embedded in either dimensional or featural similarity representations) could be systematically formulated, compared and tested. Given the framework, one can investigate precisely in what situations which model aspects perform best. Second, the framework was applied to categorization data of well-known and novel stimuli in the context of familiar natural language concepts. The results indicate that, depending on the amount of knowledge and mastery of the stimuli, different representational structures and different decision processes may operate.

One may wonder how these results relate to the findings from the category learning literature (e.g., Nosofsky, 1992; Smith & Minda, 2000; Stanton, Nosofsky & Zaki, 2002). In most of these studies, exemplar models embedded in multidimensional representations have been shown to account very well for the categorization data. However, in these studies, artificial categories are used almost invariably, with stimuli that vary along a limited number of salient dimensions. Formal models, such as the ones described in our paper, have seldom been applied to natural language concepts, which are far more complex than the stimuli used in the artificial category literature, and of which our participants arguably have a much richer and more elaborate knowledge than even the best trained participants have of artificial stimuli. (For other attempts to apply formal models to natural language concepts, see Bailey & Hahn, 2001; Smits et al., 2002; Storms, De Boeck, & Ruts, 2000, 2001; Verbeemen et al., 2001.) However, in spite of participants' extensive knowledge of such concepts, determining the relevant underlying dimensions or features for categorization with natural language concepts is perhaps the most crucial problem in modeling natural language categories (see, e.g., Murphy & Medin, 1985). The two by three framework that was presented here may serve as a valuable tool in this endeavor.

Acknowledgments

The first author is a research assistant of the Fund for Scientific Research – Flanders. This project was in part sponsored by grant OT/01/15 of the University of Leuven research council to Gert Storms.

References

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, **44**, 568-591.

Borg, I., & Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications.* Springer-Verlag, New York.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology*, **31**, 607-649.

Corter, J. E. (1982). ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath & Tversky's ADDTREE algorithm. *Behavior Research Methods & Instrumentation*, **14**, 353-354.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **18**, 441-461.

Kass, R. E. , & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kellogg, R. T. (1980). Simple feature frequency versus feature validity models of formation of prototypes. *Perceptual and Motor Skills*, **51**, 295-306.

Keren, G., & Baggen, S. (1981). Recognition of alphanumeric characters. *Perception and Psychophysics*, **23**, 234-246.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Lee, M.D. (2001a). On the complexity of additive clustering models. *Journal of Mathematical Psychology,* **45***,* 131-148.

Lee, M. D. (2001b). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology*, **45**, 149-166.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin and Review*, **9**, 43-58.

Luce, R. D. (1962). An observable property equivalent to a choice model for discrimination experiments. *Psychometrika*, **27**, 163-167.

Malt, B. C., & Johnson, E. C. (1992). Do artifact concepts have cores? *Journal of Memory and Language*, **31**, 195-217.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, **23**, 250-269.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, **32**, 517-535.

McRae, K., & Cree, G. S. (2002). Factors underlying category-specific semantic deficits. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category-specificity in brain and mind* (pp. 211-249). East Sussex, England: Psychology Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 355-368.

Murphy, G. L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289-316

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13**, 87-108.

Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honour of William K. Estes*, Vol. 1. Lawrence Erlbaum, Hillsdale, NJ.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **28**, 924-940.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika*, **47**, 3-24.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, **42**, 319-345.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**, 54-87.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, **86**, 87-123.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **24**, 1411-1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **26**, 3-27.

Smits, T., Storms, G., Rosseel, Y., & De Boeck, P. (2002). Fruits and vegetables categorized: An application of the generalized context model. *Psychonomic Bulletin and Review*, **9**, 836-844.

Stanton, R. D., Nosofsky, R. M., & Zaki, S. R. (2002). Comparisons between exemplar similarity and mixed prototype models using a linearly separable category structure. *Memory and Cognition*, **30**, 934-944.

Storms, G., De Boeck, P., & Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, **42**, 51-73.

Storms, G., De Boeck, P., & Ruts, W. (2001). Categorization of unknown stimuli in well-known natural language concepts: a case study. *Psychonomic Bulletin and Review*, **8**, 377-384.

Takane, Y., & Shibayama, T. (1992). Structures in stimulus identification data. In F. G. Ashby (Ed), *Multidimensional models of perception and cognition*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, **42**, 7-67.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

Verbeemen, T., Vanoverberghe, V., Storms, G., & Ruts, W. (2001). The role of contrast categories in natural language concepts. *Journal of Memory and Language*, **44**, 618-643.

Everyday Conditional Reasoning with Working Memory Preload

Niki Verschueren (Niki.Verschueren@psy.kuleuven.ac.be)

Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)

Gery.d'Ydewalle (Gery.d'Ydewalle@psy.kuleuven.ac.be)

University of Leuven, Lab of Experimental Psychology, Tiensestraat 102
3000 Leuven – Belgium

Abstract

There are two accounts explaining how background information can affect the conditional reasoning performance: the probabilistic account and the mental model account. According to the mental model theory reasoners retrieve and integrate counterexample information to attain a conclusion. According to the probabilistic account reasoners base their judgments on likelihood information. It is assumed that reasoning by use of a mental model process requires more working memory resources than solving the inference by use of likelihood information. We report a thinking-aloud experiment designed to compare the role of working memory for the two reasoning mechanisms. It is found that when working memory is preloaded participants use less counterexample information, instead they are more inclined to accept the inference or to use likelihood information. The present results add to the growing evidence showing that working memory is a crucial determinant of reasoning strategy and performance.

Introduction

There is evidence for a general link between working memory capacity and performance in a range of reasoning tasks (see e.g., Barrouillet, 1996; Gilhooly, Logie, & Wynn, 1999; Kyllonen & Christal, 1990). Previous studies showed that skilled reasoners generally give more normative answers and follow a high demand reasoning strategy (see e.g., Copeland & Radvansky, in press; Gilhooly, Logie, & Wynn, 1999). It is assumed that these normative answers are obtained by an analytic reasoning mechanism that hinges on working memory capacity (Klauer, Stegmaier, & Meiser, 1997; Meiser, Klauer, & Naumer, 2001). The present research continues this line of research and concerns causal conditional reasoning with everyday sentences.

Without labeling conclusions as (in)valid, we will investigate how people solve the following two conditional inferences with everyday causal sentences:

Modus Ponens (MP)
If cause, then effect
Cause occurs.
Does the effect follow?

Affirmation of the Consequent (AC)
If cause, then effect
Effect occurs.
Did the cause precede?

Examples of everyday 'if cause, then effect' sentences are: If you phone someone, then his telephone rings.
If you eat salty food, then you will get thirsty.
If someone has a high income, this person will be rich.
If a dog has fleas, then it will scratch constantly.

Abundant research established that when people reason on everyday conditionals, they spontaneously bring relevant background knowledge into account (for a review see Politzer & Bourmaud, 2002). This contextualization process is characteristic for common-sense reasoning and is responsible for our ability to adaptively cope with everyday situations. The current study focuses on how background knowledge is used for deriving conditional inferences.

There are two reasoning mechanisms describing how background information is used during reasoning. First, according to the probabilistic account reasoners derive the probability that the conclusion follows given the categorical premise and use this probability to draw a gradual conclusion (Lui, Lo, & Wu, 1996; Oaksford, Chater, & Larkin, 2002). For MP, reasoners will confine their knowledge base to the situations where the cause occurs. Based on this range of situations they then determine the likelihood that the effect follows. If they can induce that a particular effect always or mostly follows the cause, they conclude that the effect will (probably) follow. The endorsement of MP is thus directly proportional to L(effect|cause). AC is solved in analogy with MP. Reasoners activate all relevant situations where the effect occurs. Within this subset they infer the likelihood that the cause preceded the occurring effect. This likelihood L(cause|effect) directly reflects the AC acceptance rate.

According to the second reasoning mechanism the conclusion is attained by taking possible counterexamples into account. There is a strong and reliable effect of the number of available counterexamples on inference acceptance (see e.g., Cummins, Alksnis, Lubart, & Rist, 1991). The mental models theory describes how participants reason with counterexample information (Johnson-Laird & Byrne, 1991; Markovits & Barrouillet, 2002). When given a problem based on a causal rule, for instance, '*If you water a plant well, the plant stays green*', reasoners will start by representing the content of the conditional as a possibility: It is possible that a plant is well watered and green. Active consideration of the problem

content will then lead to an automatic activation of relevant background information. This information is used to complement the initial model. For MP and MT, the categorical premise triggers the retrieval of disablers. Some examples of disablers are: 'the plant caught a disease' or 'the plant was deprived of sunlight'. When reasoners retrieve at least one disabler, they do not conclude that the effect follows. For AC an automatic search for alternative causes starts, for example, 'the lack of water was compensated by adding fertilizer' or 'the plant is a succulent'. When reasoners retrieve an alternative cause, their mental models inform them that there are two conclusions possible (watered and not watered). As a result, they do not accept the default conclusion.

It is clear that the probabilistic and the mental model reasoning mechanisms both rely on available background information, but they focus on a different *type* of background knowledge: probabilities versus exemplars. Both information types have already been brought together by, e.g., Weidenfeld & Oberauer (2003); Verschueren, Schaeken and d'Ydewalle (2003; 2004a) integrated the two theories that explain how the information is taken into account in a dual process perspective. They label the probabilistic mechanism as heuristic and the mental model mechanism as analytic. Heuristic processes are generally considered as fast, automatic mechanisms that operate at a low cognitive cost and at the periphery of awareness. Analytic processes are generally slower, more demanding reasoning mechanisms that operate in a conscious and strategic manner (Stanovich & West, 2000). Verschueren et al. (2004a) manifest three reasons for linking the two reasoning processes to a heuristic-analytic polarity. (1) The heuristic reasoning process is mainly implicit - reasoners have no recollection of the range of situations that are taken into account to calculate a likelihood estimate whereas people reasoning by use of mental models are conscious of the counterexample(s) they retrieve. (2) The process based on likelihood information yields relatively fast results whereas using counterexamples requires a sequential thus slower reasoning process. (3) The heuristic conclusion is overwritten when a more analytical conclusion can be produced (see Verschueren, et al., 2004a for experimental evidence for 2 and 3). At present we will investigate whether both reasoning mechanisms differ in their working memory demands. If indeed the mental model account describes an analytical reasoning mechanism it should pose more demands on working memory capacity than the heuristic likelihood process.

Experiment

It is assumed that reasoning with counterexample information draws heavily on working memory resources, whereas the use of mere likelihood estimates imposes a far lesser demand on working memory. When participants reason based on counterexample information, the problem content as well as all models of relevant situations have to be represented. The larger the number of mental models that participants have to represent and maintain, the heavier the load on working memory during reasoning (Barrouillet & Lecas, 1999). Additionally, it has been found that counterexample retrieval efficiency suffers from dual task loads, which indicates that working memory is also involved in the retrieval of counterexample information (De Neys, 2003). In case the reasoners have a representation of both the conditional sentence and at least one counterexample, they subsequently have to integrate this information to see that there are two different conclusions for the same problem. This information manipulation and integration is considered as a crucial task of working memory.

For the reasoning process based on likelihood information, the demands on working memory are far less. The situations used for attaining a likelihood estimate are not actively represented in working memory, but rather briefly accessed. There is neither an active controlled search process nor a need for premise integration. The likelihood estimate is based on all relevant situations at a time and the final conclusion directly mirrors the obtained likelihood estimate.

When reasoners are asked to think aloud during reasoning, we can monitor which information they use for deriving conclusions. By concurrently checking the information that people use we get a direct indication of the underlying reasoning process. Only in case where people do not provide extra information but accept the conclusion without further argumentation, this procedural aspect is unclear. It can be that participants did use their background knowledge and found that the likelihood that the conclusion follows is sufficient to grant acceptance or that there are no counterexamples available. Or else it can be that they did not rely on background information and just satisfied the conclusion by restating the given information.

In a previous thinking-aloud study Verschueren, Schaeken and d'Ydewalle (2004b) showed that participants with low working memory capacity more often use likelihood estimates to solve an inference, whereas participants with a larger working memory capacity rather use counterexample information. These results can be considered as an indication for the difference in working memory demands of the heuristic and analytic process. This setup provides however only *correlational* evidence. Indeed, it is still possible that a third factor (e.g., general intelligence, motivation, etc.) explains both the performance on working memory tests as well as on reasoning tasks. The following experiment was designed to test whether there is a difference in the actual working demands of the two processes.

In this experiment we examined the effect of secondary task interference on the applied reasoning mechanisms. In the dual task methodology, a secondary task chosen to burden working memory capacity has to be carried out concurrently to the criterion task. The degree of disruption in the criterion task under dual

task conditions – as compared to single task conditions – is taken to reflect the dependence of the criterion task on working memory. The criterion task we used was a thinking-aloud conditional reasoning task. Concurrent verbalization allows us to monitor the information that reasoners consult for deriving conclusions. By checking the information that people refer to (likelihood or counterexample information) we get a direct indication of the underlying reasoning process.

Because the criterion task entails spontaneous verbalization, the choice of secondary tasks is limited. Pilot work revealed that concurrent motor, auditory or articulatory activity interfered with the participants verbalization. We therefore opted for a preload paradigm. Because a spatial load is less likely to interfere with verbalization than a verbal or numerical load, we worked with a spatial preload set-up. The evidence that spatial storage tasks tap a working memory feature crucial for reasoning is twofold: Klauer, et al.(1997) report that a concurrent spatial load led to a significant disruption of propositional (including conditional) reasoning. Second, in the visuospatial domain simple storage tasks have a similar correlation with executive functioning and reasoning as classic processing-and-storage tasks (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001; Suess, Oberauer, Wittman, Wilhelm, & Schultze, 2002). We can thus assume that the preload task taps working memory resources that are needed for reasoning, while at the same time minimizing a possible interference with the verbalization process. The dot memory task we used is a classic simple storage task (adapted from Miyake, et al., 2001; Oberauer, Suess, Wilhelm, & Wittman, 2003). We briefly presented a 3x3 matrix with 4 dots forming a complex pattern, afterwards participants were asked to reproduce this dot pattern.

In the preload-condition participants had to memorize the pattern of the dots while solving a reasoning problem. We will verify whether the use of counterexample information decreases when working memory is preloaded, compared to performance in the control condition. The decreement in the use of likelihood information should be significantly smaller than the decreement in counterexample use.

Method

Participants A total of 52 first year psychology students participated in the study.

Procedure and Design The participants were tested individually. The experiment was run on computer. Participants started by reading the instructions. They were told that they will be asked to think aloud while solving conditional inference problems. The reasoning instructions read that they should answer the question as in an everyday setting. Each participant then solved two test problems, e.g.,

> If someone catches a cold, then he will cough.
> Someone coughs.
> Did this person catch a cold or not?

The participants read the premises aloud and answered immediately. When they found that they had completed their answer, they pressed a key to go to the following problem. After the presentation of the reasoning instructions, the participants either reasoned with or without working memory preload. In the preload conditions participants started by practicing two dot patterns: A pattern was presented for 500ms and participants were immediately asked to reproduce this pattern. The overall performance on the test problems was nearly perfect. After these dot pattern practice trials, participants were given instructions for reasoning under preload. First, a dot pattern was presented for 500ms, next the reasoning problem occurred, participants read the premises aloud and answered immediately. The answers participants gave were recorded on tape. When they finished their answer, they pressed a key and a blue screen appeared where they were asked to reproduce the dot pattern. When they completed the dot pattern, they pressed a key to start the next trial. It was explicitly mentioned that they had to memorize the dot patterns correctly; they were told that an incorrect reproduction rendered the trial invalid. This was done to make sure that participants actively attended the dot pattern and tried their best in memorizing it. In the control condition, the dot patterns were presented for 500ms before the premise presentation. Participants were told that these dot patterns are presented as a control condition, they were asked to look at the dot patterns but not to memorize them. They read the premises and pressed a key when their answer was complete, the next trial started immediately. The time that participants needed to read and solve the reasoning problem was measured.

Materials and Design Based on previous research we selected 12 sentences with a maximally varying necessity and sufficiency of the cause (maximal variation in L(effect|cause), L(cause|effect), and in the number of available disablers and alternatives). We made sure that the reading time of all 12 sentences was comparable ($M_{\text{number of words}} = 9.5$, $SD = .314$). Twenty-six participants solved 12 AC inferences; the others solved 12 MP problems. The 12 sentences occurred always in the same order; the causes of the first six sentences and the last six sentences were equally necessary and sufficient. For both reasoning forms, half of the participants solved the first six problems under preload; the other six problems were solved without preload (control condition). For the other half of the participants the order of the preload/control conditions was reversed. Because we used 12 different sentences, transfer effects between the two conditions could be excluded.

Results

The obtained reasoning answers were literally transcribed. Next, the condition-codes were removed and the answer types were rated. It was indicated whether the answer reflected a simple acceptance of the

default conclusion or whether there was reference to a counterexample or to a likelihood estimate. There was no overall difference in the average response time for the preload (18.19s) and the control condition (18.53s). In the control condition, there was 26% inference acceptance, in 22% of the trials participants used likelihood information and in 64% they referred to counterexamples. These results are similar to those observed by Verschueren et al. (2004a; 18%, 18% and 66% respectively).

In the preload condition there were 6.4% combination trials (in a 'combination trial' participants refer to counterexample and likelihood information) whereas in the control condition there were 23.1% combination trials. The observation that combining the two types of information becomes less prevalent when working memory is preloaded, suggest that the information integration process that is characteristic for combination answers taps on working memory resources. For comparing the relative importance of both reasoning processes, we confined the analysis to trials where participants either referred to a likelihood or to counterexample information. Combination trials were excluded from the analysis (14.4%).

Task interference. Only 69% of the dot patterns were reproduced correctly. There was an effect of answer type on the correct reproduction of the dot patterns, $F(2, 21) = 6.696$, $p < .01$ (*Wilks' lambda* = .611). This interaction is displayed in Figure 1. When the dot patterns were correctly reproduced, there were fewer counterexamples mentioned than when the dot patterns were incorrectly reproduced, $F(1, 22) = 11.96$, $MSE = .458$, $p < .05$. On the correctly reproduced trials, there were more answers where participants referred to likelihood information, $F(1, 22) = 5.21$, $MSE = .037$, $p < .05$. There was no significant effect on the inference acceptance rates. These results reflect a *task interference*. When participants rely on a reasoning process that puts only a minor demand on working memory there are enough resources left to maintain and reproduce the dot pattern. In contrast, when participants rely on retrieval, manipulation and integration of counterexample information, working memory capacity is severely burdened. There are then not enough resources left to actively maintain the dot patterns, resulting in an incorrect reproduction. These results support the idea that using counterexample information draws heavily on working memory resources.

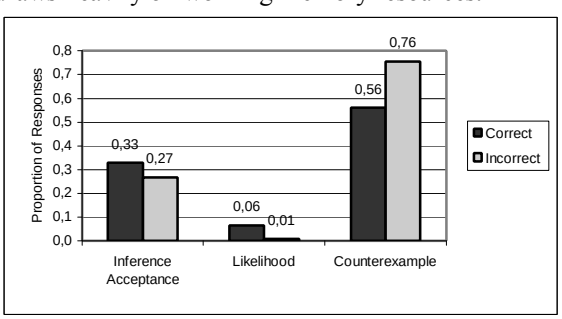


Figure 1: Difference in the proportion of the three types of answers for preload trials where the dot pattern was correctly versus incorrectly reproduced.

Effect of preload on the reasoning process. For examining the effect of preload on the types of answers, we only included the preload trials where the dot pattern was correctly reproduced. All analyses were run on proportions; the number of times each answer type occurred was divided by the total number of correctly reproduced trials. We ran an analysis of variance with sentences as the unit of analysis, and a 2 (inference type, between subjects) * 2 (preload, within subjects) * 3 (answer type, within subjects) design. We found a main effect of answer type. There were more answers referring to counterexample information (60.1%) than there was plain inference acceptance (27.7%) or likelihood information used (5.6%), F(2, 21) = 102, 72, p < .001 (*Wilks'lambda* = .08). The interaction between answer type and preload condition was marginally significant, $F(2, 21) = 3.120$ $p = .065$ (*Wilks'lambda* = .771). Figure 2 illustrates this interaction. There was a clear yet marginally significant decrease in the use of counterexample information when working memory was preloaded, $F(1, 22) = 3$ 304, $MSE = 0.078$, $p = .082$. There were significantly more inferences accepted in the preload condition, $F(1, 22) = 8.255$, $MSE = 0.131$, $p<.01$ while there was no significant increase in the use of likelihood information. No other interaction effects reached significance. The observation that there is more inference acceptance under preload corroborates previous effects of secondary task load on the conditional reasoning performance (De Neys, 2003).

The explanation provided by De Neys (2003) is that under preload, the resources available to participants are insufficient to retrieve counterexample information. The currently observed decrease in counterexample use is in line with this explanation. The increase in inference acceptance can also be - at least partially - related to an enhanced matching heuristic. We can assume that some reasoners do not engage in an active reasoning process based on counterexample retrieval, but simply restate the information from the conditional and blindly accept MP and AC. In this case the preloading should cause more participants to accept all conclusions, even on sentences where counterexamples can be automatically retrieved and likelihood estimations are high. In the preload condition, there were indeed more participants (13.5%) who accepted at least 75% of the inferences than in the control condition (7.7%). Even for sentences with many available counterexamples – for these sentences counterexamples can be retrieved automatically and likelihood estimations are very low - we found an increase in the inference acceptance rates (7.1% control vs. 19.8% preload). This shows that it is unlikely that participants consulted their background knowledge for deriving the conclusion and lends support for the hypothesis that the working memory preload led to an enhancement of the computationally low demanding matching heuristic.

In sum, as expected the resource dependent use of counterexample information decreased under preload, while the use of likelihood information was unaffected

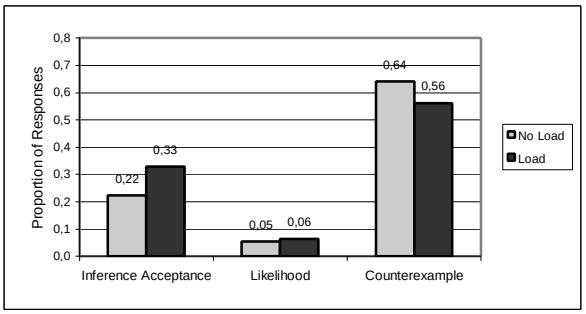


Figure 2: Proportion of answers of the three types for the preload versus control condition (only preloaded trials with correctly reproduced dot patterns).

by preload conditions. The decrease in use of the counterexample based reasoning process is at least partly compensated by shifting to inference acceptance.

Number of counterexamples used. Does the decrease in the use of counterexample information under preload reflect a decrease in a strategic validation tendency? If participants retrieve counterexample information to merely check whether the default conclusion can be falsified (see e.g., Schroyens, Schaeken, & Handley, 2003) they would need to retrieve only one counterexample to falsify the given conclusion. However, we did not find a difference in the number of trials where participants referred to only *one* counterexample (preload: 73% vs. control-condition: 82.4%). This raises doubt on the validation-hypothesis. In contrast, we observed a decrease in the proportion of trials where *more than one* counterexample was mentioned, $t(23) = 2.77$, $p < .05$ (preload: 17% vs. control-condition: 26%). This underscores the idea that in tasks without deductive instructions reasoners retrieve counterexample information to provide an adequate and informative conclusion rather than to merely falsify a default conclusion. When looking at the *total number* of specific counterexamples used, there were significantly more counterexamples used in the control condition (1.09) than when working memory was preloaded (0.86), $t(23) = 3.97$, $p < .01$.

If counterexample retrieval, representation and integration demand effort, we should observe an effect of counterexample retrieval on the secondary task performance. We tested whether there was a difference in the number of counterexample answers for the trials where the dot pattern was correctly versus incorrectly reproduced. We included the number of available counterexamples (few/many; measured by the generation task) because it is a strong predictor of counterexample use. There was a marginally significant interaction between the number of counterexamples used and the (in)correct reproduction of the dot pattern, $F(1.20) = 4.120$, $MSE = 2.866$, $p = .056$. Pairwise comparisons revealed that for sentences with many available counterexamples there were significantly more counterexamples produced when the dot patterns were not recalled correctly, $F(1, 20) = 6.946$, $MSE = 4.832$, $p<.05$ (not significant for few-sentences). This converges with the observed interference of counterexample use and correct dot pattern recall.

In general, these results sustain the idea that using counterexample information draws heavily on working memory resources whereas using likelihood information or matching is less resource demanding.

Discussion

Correlational studies revealed that differences in working memory capacity relate to differences in the conditional answer patterns. A possible explanation is that differences in reasoning performance do not simply relate to differences in a single reasoning predisposition, but are mediated by differences in the working memory demands of the active reasoning mechanisms. Highlighting the distinction between more heuristic strategies (such as matching and likelihood use) and more cognitively demanding analytical strategies (relying on counterexamples) may provide a more differentiated picture of the specific role of working memory in conditional reasoning. We found evidence for two conditional reasoning mechanisms with a differing working memory demand: a probabilistic account relying on likelihood information and a mental model account relying on counterexample information.

The results reveal that using counterexample information to attain a conclusion taps heavier on working memory resources than deriving the conclusion based on likelihood information. This provides additional support for considering the reasoning process based on likelihood information as heuristic and the reasoning process based on counterexample information as analytic. The differences in use of counterexamples/likelihood on participants with varying working memory capacity observed by Verschueren et al. (2004b) may thus be attributed to the working memory demands of the two reasoning mechanisms.

We found a large effect of working memory preload on the inference acceptance rates. When relating inference acceptance to the two reasoning strategies, it can reveal that either no counterexamples can be retrieved or that the likelihood estimation is sufficiently high. However, because we also observed an increase in inference acceptance on sentences for which pretests revealed many available counterexamples as well as likelihood estimates that are well below 1, it rather seems that the inference acceptance rates show that under preload some reasoners do not consult their background knowledge. When working memory capacity is burdened by preload, these participants are discouraged to engage in a demanding retrieval process. Instead they provide an answer that satisfies the inference question, simply by restating information from the premises. This strategically placed escape hatch can explain the increase in inference acceptance rates under preload.

Taken this together, we found evidence for the involvement of working memory in conditional reasoning. By analyzing the answers participants gave we were able to pinpoint which information participants used to attain their conclusion. We found

support for distinguishing two heuristic reasoning strategies -use of likelihood information and matching- and for an analytic strategy that takes counterexamples into account. Working memory preload yielded an increase in the use of heuristic strategies whereas the use of the analytical strategy decreased.

The present study is one of the first to combine a secondary task paradigm with a verbalization criterion task. Using a preload-paradigm is probably the best way to investigate the working memory demands of tasks involving verbalization. Although we cannot be entirely conclusive on a possible secondary task interference on verbalization processes (the answers were structurally similar to baseline results) this procedure enabled us to experimentally test the difference in working memory demands.

The effect of working memory capacity on inference making is at present only discussed on an intensive level: We investigated the *global* effect of a working memory dependent secondary task on the use of likelihood and counterexample information. Whether the working memory demands of the two processes coincide with the assumed differences in representation, retrieval and manipulation cost cannot be decided upon based on the present results. The data may also reflect the cost of determinacy: Giving a gradual uncertain answer may be overall less demanding than providing a determinate conclusion. There is also no information about the relative functional involvement of the different working memory components. Specific research with different types of well-chosen secondary tasks may reveal this crucial information.

In sum, distinguishing different reasoning mechanisms that can be used to solve conditional inferences can enhance our comprehension of how working memory mediates the reasoning performance. The specific working memory demands of different reasoning strategies co-determine the robust effect of working memory capacity on the conditional reasoning performance.

Acknowledgments

This research was conducted thanks to funding of the Fund for Scientific Research (F.W.O-Vlaanderen).

References

Barouillet, P., Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Thinking and Reasoning, 5*, 289-302.

Copeland, D. E. & Radvansky, G. A. (in press). Working memory and syllogistic reasoning. *Quarterly Journal of Experimental Psychology.*

Cummins, D.D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19,* 274-282.

De Neys, W. (2003). *In search of counterexamples: A specification of the memory search process for stored counterexamples during conditional reasoning.* Unpublished doctoral dissertation, University of Leuven, Belgium.

Gilhooly, K.J., Logie, R.H., & Wynn,V. (1999). Syllogistic reasoning tasks, working memory and skill. *European Journal of Cognitive Psychology, 11, 473-498.*

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Laurence Erlbaum.

Klauer, K. C., Stegmaier, R. & Meiser, T. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking and Reasoning, 3*, 9-47.

Liu, I., Lo, K., & Wu, J. (1996). A probabilistic interpretation of 'if-then'. *Quarterly Journal of Experimental Psychology, 48*, 828-844.

Markovits, H. & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review, 22*, 5-36.

Miyake, I., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning and spatial abilities related? A latent variable analysis. *Journal of Experimental Psychology, 130, 621-640.*

Meiser, T., Klauer, K. C., & Naumer, B. (2001). Propositional reasoning and working memory: the role of prior training and pragmatic content. *Acta Psychologica, 106, 303-327.*

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology, 26,* 883-899.

Oberauer, K., Suess, H.-M., Wilhelm, O., & Wittmann, W. W. (2003) The multiple facets of working memory: Storage, processing, supervision and coordination. *Intelligence, 31, 167-193.*

Politzer, G. & Bourmaud, G. (2002). Deductive reasoning from uncertain conditionals. *British Journal of Psychology, 93*, 345-981.

Schroyens, W. Schaeken, W., & Handley, S. (2003). In search of counterexamples: Deductive rationality in human reasoning. *Quarterly Journal of Experimental Psychology*.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioural and Brain Sciences, 23*, 645-726.

Suess, H.-M., Oberauer, K., Wittmann, W., Wilhelm,, O., & Schultze, R. (2002). Working memory explains reasoning ability – and a little bit more. *Intelligence, 30, 261-288.*

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2003). Two reasoning mechanisms for solving the conditional 'fallacies'. *Proceedings of the 25rd Annual Conference of the Cognitive Science Society* Mahwah: Erlbaum.

Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2004a). *A dual process theory on causal conditional reasoning.* Manuscript submitted for publication.

Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2004b). Working memory capacity determines which reasoning process is used for solving conditional inferences. Accepted for publication in *Memory and Cognition*.

Weidenfeld, A. & Oberauer, K. (2003). Reasoning from causal and non-causal conditionals: Testing an integrated framework. *Proceedings of the 25rd Annual Conference of the Cognitive Science Society* Mahwah: Erlbaum.

Grammatical Gender and Meaning

Gabriella Vigliocco (g.Vigliocco@ucl.ac.uk)
Department of Psychology, University College London, Gower Street
London WC1H 6BT, England

David P. Vinson (d.vinson@ucl.ac.uk)
Department of Psychology, University College London, Gower Street
London WC1H 6BT, England

Federica Paganelli (f.paganell@ucl.ac.uk)
Department of Psychology, University College London, Gower Street
London WC1H 6BT, England

Abstract

Two experiments assessed whether grammatical gender of Italian nouns referring to animals and tools affects conceptual representations of the corresponding objects, comparing results from Italian and English. In the first experiment, we elicited semantic substitution errors (e.g., saying "hammer" when "axe" is intended), finding language-specific gender effects (more errors in Italian than English for words sharing gender) for words referring to animals but not for words referring to tools. In the second experiment, words sharing gender were judged as more similar in meaning by Italian speakers than English speakers, again only for animals and not for tools. Moreover, no such gender effect was observed for pictures of the same animals.

Introduction

As Roman Jakobson (1959) put it: "Languages differ essentially in what they must convey and not in what they may convey" (p.236). That is, languages differ in which conceptual or formal properties must be realized in sentential form. For example, in English the word "friend" does not indicate the sex of the friend, while in Italian the corresponding word is differentially inflected for a man ("amico") or a woman ("amica"). In English, adjectives used as predicates (e.g., "tall" in "The boy is tall" and "The girl is tall") do not agree in gender with the subject of the sentence, while they must in Italian (e.g., "Il ragazzo e' alto" or "La ragazza e' alta"). Such differences in obligatory expression may imply that speakers of different languages pay more or less attention to those dimensions of meaning. For example, Italian speakers may pay more attention to the sex of referents than English speakers. By extension, Italian speakers may tend to think of objects in the world as more male- or female-like on the basis of the words' grammatical gender (as suggested by the work of, e.g. Boroditsky, Schmidt & Phillips, 2003; Sera, Elieff, Forbes, Burch, Rodriguez, & Dubois, 2002). But how strong and pervasive can these effects be?

Here we present experiments investigating the conditions under which effects of a language-specific property (grammatical gender of Italian nouns) are present, contrasting performance by Italian and English speakers on translation-equivalent nouns. Grammatical gender allows a conservative test of language-specific effects on cognition because it is largely arbitrarily linked to meaning (although see Foundalis, 2002).

How could grammatical gender affect conceptual representations for objects? Effects of grammatical gender could arise as a consequence of general language-learning mechanisms based on similarity. According to this hypothesis (to which we will refer as "Similarity and Gender"), words that are similar to each other on any linguistic dimension (including but not limited to grammatical gender) may become more semantically similar as a consequence of the fact that words of the same syntactic class (e.g., same gender, same grammatical class, etc.) appear in the same syntactic contexts. For example, in languages with grammatical gender, nouns are used in sentences along with gender-marked determiners and adjectives, whether the nouns refer to sexuated entities or not. Sensitivity to shared sentence context could allow children to bootstrap properties of similarity in meaning from the syntactic contexts in which the words occur during language acquisition (Landauer & Dumais, 1997). This hypothesis does not require any explicit associations between grammatical gender and sex of human referents; instead it predicts that any effects of grammatical gender on semantic representations should be found in any gendered language (no matter how many gender classes are in the language), and that they should be found for all words (whether the referents are sexuated or not).

However, mechanisms mediating such effects may be more specific and limited. According to this other hypothesis (to which we will refer as "Sex and Gender"), effects of grammatical gender could arise because children would treat all grammatical categories as revealing specific semantic properties (Boroditsky, et al., 2003). In the case of grammatical gender, these effects would require linking the grammatical gender of nouns referring to humans to the sex of referents. Across languages there is a core correspondence between grammatical marking of gender and biological sex (Corbett, 1991), although the consistency of this mapping differs across languages. According to this view, children learning a gendered language would first notice the core correspondence between the gender of nouns and male/female semantic properties of human referents (and some animals). They would then generalize this correspondence to other nouns for which there is no clear conceptual foundation of gender,

assigning male or female features to referents in agreement with the grammatical gender of the corresponding words. Thus, words of the same gender would be more similar among themselves than words of different gender because they share male or female-like properties. Such a mechanism could be strongest for languages with the greatest degree of correspondence between the gender of nouns referring to humans and the sex of referents. This is the case in Romance languages which have only two genders and few exceptions to the consistent mapping between the gender of nouns referring to humans and sex of the referents. It could be weaker (if present at all) in languages with multiple genders and/or in which nouns referring to humans fall into more than two classes. Moreover, any effect of gender could be stronger for words referring to sexuated entities (e.g. animals) than for words referring to objects and abstract concepts, because semantic properties of sex are less relevant in these latter domains. Most of the studies investigating language-specific effects of grammatical gender (e.g. Boroditsky, et al., 2003; Sera, et al., 2002) have tested this hypothesis, either implicitly or explicitly.

In the experiments below we tested some predictions stemming from these views, considering grammatical gender of Italian nouns. As in other Romance languages, all nouns in Italian are marked for gender, either masculine or feminine. For nouns referring to humans and some animals, the gender depends on the sex of the referent (e.g., "ragazzo/ragazza" [boy/girl]; "leone/leonessa" [lion/lioness]), while for other animals gender does not depend upon the sex of the referent (e.g., "lupo" refers to both male and female wolves, although it is possible to mark the gender in some cases). For words referring to objects and abstract entities, instead, there are no such clear semantic correlates (with certain exceptions not addressed here). We investigate two semantic fields, animals and tools, to test the hypotheses outlined above. Both predict language-specific effects of grammatical gender on meaning, such that word pairs sharing Italian gender will show greater semantic similarity effects than the same word pairs in English translation. The two hypotheses make different predictions, however, suggesting that these effects may have different breadth. The Sex and Gender hypothesis predicts that greater language-specific gender effects should be observed for animals than for tools (as animals are sexuated entities), while the Similarity and Gender hypothesis predicts no category difference.

Experiment 1

Here we assessed whether grammatical gender affects online linguistic tasks such as picture naming. We begin with a linguistic task, as finding language-specific gender effects is not only evidence for a "thinking for speaking" view of language-specific effects on cognition (Slobin, 1996), but is also a pre-requisite to testing for broader language specificity in tasks less tightly tied to linguistic encoding. We focus upon *semantic substitution errors* (i.e., instead of producing a target word, speakers produce another word related in meaning, e.g., saying "hammer" when "saw" is intended) in picture naming. In previous work we have introduced a continuous picture-naming paradigm to elicit such errors (Vigliocco, Vinson, Lewis & Garrett, in press). Here we investigate whether semantic substitution errors in Italian tend to preserve the gender of the target word (i.e., masculine nouns are more likely to substitute for other masculine nouns, and feminine nouns for other feminine nouns). It is generally agreed upon in the language production literature that semantic substitutions arise during the process of retrieving the lexical entry corresponding to a concept, thus tapping into the interface between linguistic and conceptual knowledge (e.g., Garrett, 1984; Levelt, 1989). Moreover, these errors are sensitive to fine-grained semantic similarity: the likelihood of errors increases with greater semantic similarity (Vigliocco et al., in press). Thus, other factors being equal, if grammatical gender has a semantic effect, it should increase the likelihood that words of the same gender substitute for each other in a language such as Italian. However, other factors may not necessarily be equal, particularly non-language-specific factors such as semantic similarity not related to gender in a language-specific manner, or visual similarity among pictorial referents. In order to provide the tightest controls, we selected English as a baseline comparison language, using the same items (translation equivalent words) and the same tasks. This allows us to test for language-specific effects of Italian gender while avoiding concerns related to general semantic or visual similarity among the items used in our experiments. We investigate whether Italian errors tend to preserve gender above the English baseline level (based upon assigning Italian gender to English translations). It should also be noted that substitution errors are sensitive to phonological similarity between target and intruder (Dell & Reich, 1981), and Italian gender does have strong and reliable phonological correlates. To minimize the possibility that any observed language-specific effects of gender are due to phonological similarity, we also conducted analyses in which we excluded all errors with substantial phonological similarity to the target word.

Method

Participants

Participants were 27 native speakers of Italian from the London community and 20 English speakers from the UCL subject pool. The Italian participants had only rudimentary knowledge of English, and none of the English speakers reported moderate or better competence in any Romance language.

Materials

We selected 27 black and white line drawings of animals, avoiding those animals for which the gender of the noun strictly depends upon the sex of the referent. We further selected 50 black and white line drawings of tools. Most pictures came from Snodgrass and Vanderwart (1980), with additional ones prepared for the experiment. Name agreement was established for each of our participants during the experimental session (see also Vigliocco et al., in press); in general there was very strong name agreement in both languages, further ensuring that the words are suitable translation equivalents. Because in previous work we have established that substitution errors in this paradigm do not cross

semantic fields, we used a blocked presentation design, analyzing the data for the animals and the tools separately. We presented 77 blocks of 10 pictures each to every participant. Each block contained only animal or tool pictures (presented in random order within the block), and each picture appeared 10 times in the course of the experiment.

Procedure

The experiment began with a name agreement phase in which each picture was presented and participants were asked to name them. This phase allowed us to ensure name agreement across participants and also to identify specific naming preferences by individual participants (which might otherwise have been considered errors). Next, a practice series of blocks were presented in which the speed of presentation of each picture was adjusted for each participant (between 600ms and 1100ms) in order to render the task difficult but manageable for each speaker. After the training, the experiment started. Participants were told that their task was to name each picture as it appeared on a computer screen as quickly as possible.

Scoring Criteria

Participants' responses were transcribed and scored in the following categories: *Correct responses*: participants uttered the correct target word. *Different label*: participants used a different word than our intended target (e.g. "stag" for "deer"), but this different label was consistently used by that participant and did not refer to another item in the experiment. *Lexical errors*: participants produced a word that differed from the target and that did not qualify as a "different label". Lexical errors were further classified as "out of set" (intruding words that are not among the experimental items) and "within set" (those items from within the present response set). Because of the repeated presentation of a limited set of pictures to be named, most lexical errors tended to be other items from the response set. Because of this, and to minimize the possibility of linguistic variability beyond this particular set of item, analyses were performed only upon within-set lexical errors items. *Miscellanea*: other responses not included above, such as dysfluencies, incomplete utterances, inaudible responses, omissions and self-corrections. Table 1 reports a breakdown of the proportions of responses in the different scoring categories.

Table 1: Response Types
(IT: Italian, EN: English; A: Animals, T: Tools)

Response type	IT: A	EN: A	IT: T	EN: T
Correct & Different Label	.876	.935	.948	.910
Lexical errors				
Within-set	.021	.024	.013	.018
Out of set	.004	.001	.002	.002
Miscellanea	.100	.04	.133	.062

Results and Discussion

All analyses were carried out on within-set lexical errors. First, we eliminated all those items for which the average correct performance was not above 75% in both languages or for which the average correct performance differed more than 15% across the two languages in order to exclude additional cross-cultural differences. For each semantic field (animals and tools) we carried out two 2x2 ANOVAs. In all ANOVAs, proportion of errors was the dependent variable with target-error pair as a random factor. Independent variables were language (Italian, English) and Italian gender (shared between target and intruder or not shared). English words were assigned Italian gender for the purpose of the analysis. The first ANOVA was carried out on the within-set errors remaining after we excluded the cases discussed above (for *animals*: 103 errors in Italian and 73 errors in English; for *tools*: 90 in Italian, and 117 in English). For *animals*, this analysis showed a significant interaction between language and Italian gender, such that errors sharing gender with the target were more common in Italian (68%) than in English (41%); $F(1,63) = 8.03$, $p = .006$. Neither main effect was significant ($F < 1$). The results of the analysis for *tools* were similar; only the interaction between language and gender was significant: gender preservation was greater in Italian (61%) than in English (36%); $F(1,79) = 4.6$, $p = .04$; main effect $Fs < 1$).

In the second analysis we excluded all errors for which the target and the intruder shared phonological similarity. Phonological similarity between target and intruder was assessed as in Vigliocco et al. (in press). In this second analysis, only errors for which either of two measures of phonological overlap did not exceed the average + one standard deviation of that measure (in either language) were considered (for animals, leaving 64 errors in Italian and 42 in English; for tools, leaving 39 errors in Italian and 42 in English). This analysis for *animals* also showed a significant interaction between language and gender; such that even among target-intruder pairs with low phonological similarity, Italian target-intruder pairs tended to share gender (77%) more often than English pairs (43%) (interaction $F(1, 37) = 5.88$, $p = .020$; main effect $Fs < 1$). However, this interaction was not significant in the analysis for *tools* (all $Fs < 1$). Thus, for the tools, the language x gender effect observed in the complete set of errors may just be a consequence of greater phonological similarity in Italian for words sharing the same gender.

To summarize the results of this experiment, we found language-specific effects of grammatical gender; gender affects the likelihood of producing a lexical error for Italian speakers, compared to the errors produced by English speakers naming exactly the same pictures. This language-specific effect of grammatical gender, however, survives only for words referring to animals once phonological similarity is taken into account. This result suggests that language-specific effects are constrained even in a linguistic encoding task such as picture naming.

Experiment 2

The results of the error induction task in Experiment 1 show that language-specific effects of grammatical gender can be observed in an on-line task requiring lexical retrieval. In this

second experiment we sought to obtain converging evidence using a very different task. Moreover, we assessed the generalizability of these effects beyond linguistic materials by performing the same experiment using pictures as well as words as stimuli. As in Experiment 1, we contrast responses from Italian speakers for words and pictures referring to animals and tools with responses from English speakers. In this experiment we used the triadic similarity judgment task. Speakers of Italian and English were presented with triplets of words or pictures (translation equivalents in Italian and English) and their task was to judge which two of the three were more similar in meaning. This task has been successfully used in previous studies investigating semantic organization and its impairments (Fisher, 1994; Garrard, Carroll, Vinson & Vigliocco, in press). Particularly relevant here are the following facts. First, because all possible combinations of triads of a relatively small set of items are presented to the participants, this task allows us to consider semantic similarity at a very fine-grained level. Second, this task has been shown to be sensitive to linguistic variables at the interface between meaning and syntax. For example, Fisher (1994) showed that English speakers' judgments reflected differences in the subcategorization requirements of semantically related verbs; Garrard et al (in press) showed that English speakers' judgments reflected the distinction between "count" and "mass" nouns for words referring to food items (for which the semantic divide between entities and substances is less obvious). Thus, if grammatical gender of Italian nouns exerts influence upon semantic similarity, we should observe language-specific effects in this task. If this effect extends beyond the use of linguistic materials we should observe it also with pictures.

Because all possible triads within a category are to be presented to the participants, and in order to maximize the opportunity of observing grammatical gender effects, which could be masked by extreme semantic diversity in the item set, all participants were presented with words from only one of two categories (land animals in Experiment 2a, and tools in Experiment 2b), reported separately below.

Experiment 2a: Animals

Method

Participants

Participants were 24 native speakers of Italian from the London community, and 24 native English speakers from the University College London participant pool. The Italian participants had only rudimentary knowledge of English, and none of the English speakers reported moderate or better competence in any Romance language.

Materials

Words (and corresponding pictures) referring to 20 animals were selected for the experiment. Words were translation equivalents in Italian and English. The words and the pictures used were a subset of those used in Experiment 1.

Triads for the Italian and English conditions were created by first assembling all possible three-word combinations of the 20 items in the experimental set (for a total of 1,140 triads). The order of words in each triad was randomized; then the order of triads was randomized across participants. A separate set of picture triads were then created by replacing each word with its corresponding picture (this set was identical for Italian and English participants). Twelve participants from each language were assigned to the word condition and twelve to the picture condition. In each modality (word or picture) and language (Italian or English) condition, the 1,140 triads were divided into three lists, each containing 380 triads of words or pictures.

Procedure

All participants were told that the experiment concerned participants' judgments of meaning similarity among groups of words (or pictures), and that their task was to choose the two words (or pictures) out of the three which were more similar in meaning and to delete the odd one out. Instructions emphasized that the decision was to be made on the basis of meaning and not other types of similarity between the items (e.g., phonological similarity among the words or visual similarity among the pictures). After completing the task, participants were asked to describe the strategies they may have used to perform the task, to list the easiest and most difficult decisions, etc. For the purpose of the present study, the most important aspect of these questions was whether any Italian participants mentioned grammatical gender as an overt basis for making their decisions.

Design and Analysis

The dependent variable was similarity ratio: the number of times that a given pair of words/pictures was selected as "similar", divided by the number of triads in which those two items appeared in the experiment. Four participants completed each list of 380 items; thus each triad (either words or pictures) was judged by four different speakers of a language. Results were analyzed using a three-way mixed ANOVA with item pairs as a random factor. Independent variables were language (English or Italian, manipulated within item pairs), modality (words or pictures, manipulated within item pairs) and Italian gender (same Italian gender; different Italian gender, manipulated between item pairs). As in Experiment 1 this latter factor refers to the gender of the Italian translation or label.

Results and Discussion

No Italian participant indicated that they used grammatical gender in their similarity judgments in the post-experimental questionnaire. Table 2 reports the average similarity proportions for items of same vs. different Italian gender as a functions of language and modality.

Table 2: Average similarity ratios in Experiment 2a (Standard errors in brackets)

Language	Modality	Grammatical Gender Same	Different
Italian	Words	.336 [.022]	.331 [.021]
	Pictures	.315 [.026]	.351 [.025]
English	Words	.311 [.026]	.354 [.025]
	Pictures	.315 [.025]	.351 [.024]

Results were analyzed using a three-way mixed ANOVA investigating the effect of language, modality and Italian gender on similarity proportions. Only the three-way interaction (Language x Modality x Gender) reached significance ($F(1,188)=6.539$, $p=.013$); no other main effects and interactions were significant (all Fs < 1). Analysis of simple interaction effects within each modality revealed that the similarity proportion for same-gender items was relatively higher for Italian word judgments than for English word judgments (with the corresponding difference in the opposite direction for words differing in Italian gender), while there was no such language difference for picture judgments by speakers of either language.

Thus, in this experiment we found language-specific effects of grammatical gender for words referring to animals, but not for pictures referring to the same animals; Italian speakers' judgments of meaning similarity seem to be affected by shared grammatical gender. Experiment 2b assessed whether such a gender effect is present for tools.

Experiment 2b: Tools

Method

Participants

Participants were 48 native speakers of Italian from the London community, and 48 native English speakers from the University College London participant pool. The Italian participants had only rudimentary knowledge of English, none of the English speakers reported moderate or better competence in any Romance language. None of them had participated in Experiment 2a.

Materials

Words (and corresponding pictures) referring to 24 tools were selected for the experiment. Words were translation equivalents in Italian and English. The words and the pictures used were a subset of those used in Experiment 1. Word and picture triads for the Italian and English conditions were created as in Experiment 2a. In this experiment all possible three-word combinations of the 24 items in the experimental set yielded a total of 2,024 triads. These triads were divided into six lists each containing 337 or 338 words or pictures. This experiment was otherwise the same as in Experiment 2a.

Results and Discussion

No Italian participant indicated that they used grammatical gender in their similarity judgments in the post-experimental questionnaire. Table 3 reports the average similarity proportions for items of same vs. different (Italian gender as a functions of language and modality.

Table 3: Average similarity ratios in Experiment 2b (Standard errors in brackets).

Language	Modality	Grammatical Gender Same	Different
Italian	Words	.318 [.018]	.348 [.017]
	Pictures	.314 [.020]	.352 [.019]
English	Words	.316 [.017]	.348 [.017]
	Pictures	.308 [.019]	.357 [.018]

Results were analyzed using a three-way mixed ANOVA investigating the effect of language, modality and Italian gender on similarity proportions. No main effects or interactions were significant (Fs < 1), with the exception of the main effect of gender which was marginal ($F(1,274) = 2.49$, $p = .115$). This indicates an underlying tendency for tools sharing Italian gender to be more similar than items with different Italian gender. Because no interaction between language and Italian gender was observed, this main effect cannot reflect language-specificity. Thus, whereas grammatical gender affected Italian speakers' judgments of meaning similarity for words referring to land animals, no such effect was observed for either words or pictures referring to tools.

General Discussion

In the experiments reported above, we explored language-specific effects of Italian grammatical gender on semantic representations for the corresponding objects. These experiments combined on-line and off-line methodologies, assessing effects for two different semantic fields: one for which associations between grammatical gender and sex can be plausibly built (animals) and one for which they cannot (tools). Moreover, we further explored gender effects within the same task, using both words and pictures as stimuli, in order to establish the generalizability of any effect.

We found that language-specific effects of Italian grammatical gender are present, but highly limited. They are limited to a semantic field (animals) in which entities have biological gender, and in which the gender of some nouns can depend on the sex of the referent. However, this effect does not extend to a field for which distinctions in grammatical gender have no conceptual foundation (tools). These effects are further limited to tasks that recruit linguistic knowledge (picture naming or similarity judgments for words, but not similarity judgments for pictures).

Our results suggest a far more limited role of grammatical gender on semantic representations than it has been suggested in previous studies. For example, Sera et al (2002) showed that grammatical gender of Spanish and French nouns influenced speakers' assignment of a male or female voice to inanimate (and animate) objects, regardless whether the task was carried out using words or pictures as stimuli. However, speakers might have used gender in a strategic manner in this task. Boroditsky, et al. (2003) report studies suggesting that grammatical gender may have implicit effects. However, these studies are reported without enough methodological detail to address possible reasons for the different results.

Our results suggest that these gender effects are linked to assigning male- or female-like semantic properties to referents in agreement with the gender of the nouns. They provide evidence for a very constrained version of the Sex and Gender hypothesis described in the introduction. According to this view, language-specific grammatical gender effects should be stronger for semantic fields in which there is a conceptual motivation for establishing a link between gender of words and sex of the referent (such as animals), than for fields for which there is not such a clear conceptual motivation (such as tools). This hypothesis also predicts that language-specific effects of gender on meaning should be stronger for languages such as Italian that have strong transparent links between gender of nouns and sex of referents (male entities strongly tend to have masculine gender, and female entities strongly tend to have feminine gender). Although the present experiments do not directly address this second prediction, some other evidence is relevant here. Vigliocco, Vinson, Indefrey, Levelt and Hellwig (2004) investigated gender effects on semantic substitution errors in German, in a study similar to Experiment 1. Although German has grammatical gender, in contrast to Italian it has three genders (masculine, feminine and neuter) and a less transparent correspondence between the gender of nouns and sex of referents. No effect of grammatical gender was found on semantic substitution errors for animals (at least when speakers were asked to produce bare nouns). These different lines of investigation converge in suggesting that language-specific effects of gender do not arise as a consequence of a general mechanism sensitive to similarity.

Finally, the difference we observe in Experiment 2 between word and picture stimuli suggests that grammatical gender of Italian nouns (referring to animals) affects "thinking for speaking", but does not affect conceptual representations when language is not required (Slobin, 1996).

Acknowledgments

The work reported here has been supported by a Human Frontier Science Program grant (HFSP148/2000) and an Economic and Social Research Council grant (RES000230038) to Gabriella Vigliocco. We thank Jo Arciuli and Belen Lopez-Cutrin for help in preparing this manuscript.

References

Boroditsky, L., Schmidt, L., Phillips, W. (2003). Sex, syntax and semantics. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought*, Cambridge, MA: MIT Press.

Corbett, G.G. (1991). *Gender*. Cambridge: Cambridge University Press.

Dell, G.S., & Reich, P.A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20*, 611-629.

Fisher, C. (1994). Structure and meaning in the verb lexicon: Input for a syntax-aided verb learning procedure. *Cognitive Psychology, 5*, 473-517.

Foundalis, H.E. (2002). Evolution of gender in Indo-European languages. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, Fairfax, VA.

Garrard, P., Carroll, E., Vinson, D.P., & Vigliocco, G. (in press). Dissociating lexico-semantics and lexico-syntax in semantic dementia. *Neurocase*.

Garrett, M. F. (1984). The organization of processing structure for language production: Application to aphasic speech. In D. Caplan, A. R. Lecours and A. Smith (Eds.), *Biological perspectives on language.* (pp. 172-193). Cambridge: MA: MIT Press.

Jakobson, R., (1959) On linguistic aspects of translation, in R.A. Brower (ed.), *On translation*, (p.232-239). Cambridge, Mass: Harvard University Press.

Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211-240.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Sera, M.D., Elieff, C., Forbes, J., Burch, M.C., Rodriguez, W. & Dubois, D.P. (2002). When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General, 131*, 377-397.

Slobin, D. (1996). From "thought and language" to "thinking for speaking". In J. Gumperz & S. Levinson (Eds.), *Rethinking Linguistic Relativity*. (pp. 70-96). Cambridge, MA: Cambridge University Press.

Snodgrass, J. G. & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Learning, Memory and Cognition, 6*, 174-215.

Vigliocco, G., Vinson, D., Indefrey, P., Levelt, W., Hellwig, F. (2004). The interplay between meaning and syntax in language production. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*, 483-497.

Vigliocco, G., Vinson, D.P, Lewis, W. & Garrett, M.F. (in press). The meaning of object and action words. *Cognitive Psychology.*

Structural Bayesian Models of Conditionals

Momme von Sydow (momme.von-sydow@bio.uni-goettingen.de)
Department of Psychology, Georg-August-Universität Göttingen, Abt. 1, Goßlerstr. 14
D-37073 Göttingen, Germany

Abstract

In the past decade the traditional falsificationist view of hypothesis-testing tasks, such as Wason's selection task, has become criticized from a Bayesian perspective. In this report a normative extension of Oaksford's and Chater's (1994, 1998) influential Bayesian theory is proposed, that not only takes quantitative but also qualitative (structural) knowledge into account. In an experiment it is shown that humans appear to be sensitive to both the quantitative and the qualitative preconditions of the proposed normative models.

Introduction

According to falsificationism only tests of hypotheses that may lead to a falsification are normatively justified (Popper, 1934/2002). In the psychology of thinking Wason's (1966) selection task (WST) has become the most studied single task to investigate the testing of hypothesis, typically a indicative conditional in the form of "if *p* then (always) *q*." In this task, four cards are presented. The visible front sides of these cards represent the logical cases *p, non-p, q, non-q*. It is known that one side of each card shows either a *p*- or *non-p*-case and the other side either a *q* or *non-q*-case. In order to test whether the hypothesis is true or false, participants should turn over those cards that are needed to test the hypothesis. To falsificationists, who have been predominant in psychology of reasoning for long, only the selection of a *p*-card *and* a *non-q*-card is correct.

Since over three decades studies have shown that humans do not act in a falsificatory manner (e.g., Johnson-Laird & Wason, 1970): most participants selected the *p*-card and the *q*-card and only 4% selected the 'correct' combination of a *p*- and a *non-q*-card. Since 96% gave wrong answers in this very basic logical task, this finding casts doubt on the rationality of the so-called *animal rationale/zoon echon logon* (Aristotle).

In psychology, theories have been developed which kept a falsificatory core, but which explained the selections by additional mechanisms (e. g. mental model theory). Also other theories flourished, which completely broke with the concept of normative rationality altogether (Cheng & Holyoak, 1985; Cosmides, 1989; Gigerenzer & Hug, 1992). In the last decade, however, probabilistic and Bayesian approaches to the WST have been also proposed (early proposals were e. g. Kirby, 1994; Oaksford & Chater, 1994; Evans & Over, 1996). The optimal data selection model of Oaksford and Chater (1994, 1996, 1998; Oaksford, Chater & Grainger, 1999) represents the most refined approach and has received most attention (e. g.: Evans & Over, 1996; Laming, 1996; Klauer, 1999; Oberauer, 2000; Osman et al., 2001). Hence, I am here going to focus on this approach.

Models

The model of hypothesis testing by Oaksford and Chater (1994, 1998), shown in Table 1, distinguishes a dependence sub-model M_D and an independence sub-model M_I, which represent the truth or falsity of the conditional. As in logics $P(p \wedge \neg q | M_D)$ is set as zero. Different to logics the other cells in this model are quantified. By comparing of M_D and M_I it can be seen (without the further modeling steps) that in such a model not only the falsificatory selections *p-/non-q*-card selections, but also *q*-card-selections may provide a certain *information gain*. However the *non-p*-card never becomes informative, since Oaksford & Chater (1994) set $P(p)$ and $P(non\text{-}p|q)$ to be equal in both sub-models, which in turn cause a flexible q-marginal probability.[1]

This setting of parameters has been criticized by Laming (1996) as post hoc data model fitting, designed to preclude the prediction of (actually infrequent) *non-p*-selections. Laming argued that these assumptions could not be justified, since one may equally construct a model with different parameters that appears completely weird (Table 3). Oaksford & Chater (1996, p. 386) defended their setting of parameters: "Psychologically it reflects the finding that participants regard false antecedent instances (i.e., the not-p cases) as irrelevant to the truth or falsity of a conditional rule." (Cf. recently similar ideas by Over, in press, and by Evans, Headley and Over, 2003)

von Sydow (2002) argued at length against this view and in favor of a different approach. Oaksford's and Chater's above argument, for example, is against the spirit of their own approach, since by this argument also *non-q*-card selections could have been excluded a priori. Inspired by Laming's criticism, I discussed and empirically examined a model in which the resulting marginal probabilities $w(p_{res})$ and $w(q_{res})$ are set to be constant in both sub-models (Table 2). This model is actually long known from the philosophical literature on the raven paradox, but von Sydow has combined it with the further calculations of the refined model of Oaksford and Chater (1998), stressing that this model has necessary preconditions to be fulfilled by the empirical situation. At about the same time similar proposals were made (Hattori, 2002) and even Oaksford &

[1] Oaksford and Chater (1994, 1998) modified the probabilities used in Table 1 according to the following formula: $q:=[P(q)-P(p)\ P(M_D)]/[1-P(p)\ P(M_D)]$. Both models were analyzed, the pure model without the modification of $P(q)$ (Model 1) and the model with the modification. The predictions of both models are similar and in this paper I focus on the pure model (cf. table 4).

Table 1: Model of Oaksford and Chater (1994, 1998).
P(p) and P*(q|1-p)* are set to be the same in both sub-models (cf. footnote 1).
Notes for Table 1 to 3: The cells show probabilities for the Dependence Model M_D and the Independence Model M_I. Resulting marginal probabilities, $P(p_{res})$ and $P(q_{res})$, can differ from *P(p)* and *P(q)*. '*p*', '*q*' in italics abbreviates *P(p)*, P*(q)*.

M_D	q	non-q	
p	*p*	*0*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q+p-pq	*(1-p)(1-q)*	*1*

M_I	q	non-q	
p	*pq*	*p(1-q)*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q	*1-q*	*1*

Table 2: Model of von Sydow (2002), Oaksford and Wakefield (2003). *P(p)* and P*(q)* are set to be constant. (Cf. Table 1)

M_D	q	non-q	
p	*p*	*0*	*p*
non-p	*q-p*	*1-q*	*1-p*
	q	*1-q*	*1*

M_I	q	non-q	
p	*pq*	*p(1-q)*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q	*1-q*	*1*

Table 3: Model of Laming (1996). *P(q)* and P*(p|q)* are set to be constant. (Cf. Table 1)

M_D	q	non-q	
p	*pq*	*0*	*pq*
non-p	*(1-p)q*	*1-q*	*1-pq*
	q	*1-q*	*1*

M_I	q	non-q	
p	*pq*	*p(1-q)*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q	*1-q*	*1*

Wakefield (2003) in a revision turned to this model.[2] However, in these other proposals it was never stressed that in principle *all* models could be normatively justified (also the original one of Oaksford & Chater, 1994). In contrast, von Sydow argued that all could be justified, provided that their (implicit) preconditions hold in the experimental situation. In three experiments von Sydow (2002) ensured that the preconditions of the model from Table 2 were fulfilled, by ensuring fixed marginal probabilities. The results showed the predicted increase of *non-p*- and *non-q*-selections in high base rate conditions.

The aim of this present paper is to directly investigate whether humans are actually sensitive to these *different* structural preconditions. Therefore the original model of Oaksford and Chater (1994, 1998)[1], the model of von Sydow (2002) and also the model of Laming (1996) were modeled along the same lines. In regard to further steps of modeling (Bayes' Theorem, Wiener-Shannon-Information and the resulting *expected information gain* measure) I completely followed Oaksford and Chater (1998).[3]

Here only an extract of the modeling results can be presented (see Table 4). *Expected information gain* (*EIg*) values are shown for the different models for the parameter values used in the experiment (low base rate: *P(p)=.10, P(q)=.20, P(H$_D$)=.50*, high base rate *P(p)=.80, P(q)=.90, P(H$_D$)=.50*). Additionally also the normative predictions for the estimates of the marginal probabilities are shown. (The predictions are also mentioned in the results section.)

Table 4: Expected information gain and standardized expected information gain (with an error parameter) for card selections in different structural models (for low, *.10→.20,* and high base rates, *.80→90).* Resulting marginal probabilities $P(p_{res}|M_D)$, $P(q_{res}|M_D)$, $P(p_{res}|M_I)$, $P(q_{res}|M_I)$.[1]

EIg	von Sydow-Model				Oaksford-Chater-Model 1			
SEIg	p	¬p	q	¬q	p	¬p	q	¬q
low	.61	.01	.15	.05	.61	.00	.07	.05
	.58	.09	.20	.20	.63	.09	.15	.13
M_D	$P(p_{res})=.10$		$P(q_{res})=.20$		$P(p_{res})=.10$		$P(q_{res})=.28$	
M_I	$P(p_{res})=.10$		$P(q_{res})=.20$		$P(p_{res})=.10$		$P(q_{res})=.20$	
high	.05	.15	.01	.61	.05	.00	.00	.61
	.12	.20	.09	.58	.14	.09	.09	.67
M_D	$P(p_{res})=.80$		$P(q_{res})=.90$		$P(p_{res})=.80$		$P(q_{res})=.98$	
M_I	$P(p_{res})=.80$		$P(q_{res})=.90$		$P(p_{res})=.80$		$P(q_{res})=.90$	
EIg	Oaksford-Chater-Model2				Laming-Model			
SEIg	p	¬p	q	¬q	p	¬p	q	¬q
low	.67	.00	.10	.05	.61	.00	.00	.05
	.63	.08	.16	.12	.67	.09	.09	.16
M_D	$P(p_{res})=.10$		$P(q_{res})=.24$		$P(p_{res})=.02$		$P(q_{res})=.20$	
M_I	$P(p_{res})=.10$		$P(q_{res})=.16$		$P(p_{res})=.10$		$P(q_{res})=.20$	
high	.09	.00	.00	.61	.05	.07	.00	.61
	.17	.09	.09	.65	.13	.15	.09	.63
M_D	$P(p_{res})=.80$		$P(q_{res})=.97$		$P(p_{res})=.72$		$P(q_{res})=.90$	
M_I	$P(p_{res})=.80$		$P(q_{res})=.83$		$P(p_{res})=.80$		$P(q_{res})=.90$	

[2] Oaksford, Chater & Larkin (2000) had distinguished a similar model of reasoning from their model of hypothesis testing. The revision has been announced – without any reasons and without own data – in an overview article (Oaksford & Chater 2001, p. 353), which can not count as a full revision of their model.

[3] For alternative proposals cf. Laming (1996), Klauer (1999), and Chater & Oaksford (1999).

Method

Design and Participants The experiment had a 2 (low versus high base rate condition) × 3 (the three structural models) between-subjects design.

Seventy-two participants from the University of Göttingen took part in the experiment. The participants were randomly assigned to the six experimental conditions.

Materials and Procedure. Each participant was presented with what I call a 'Many Cards Selection Task' (MST) with many depicted cards (instead of four cards in a WST) in a paper and pencil version.

In all conditions the same cover story was used. Participants were asked to suppose that they were physicians at a university hospital. Their task was to find out whether the following hypothesis was true or false: "If a patient is infected by the Virus Adenophage (A), then he always shows the symptom Thoraxpneu (●)" This hypothesis was set in bold print. In order to set the parameter $P(M_D)$ in all models *to 0.5* the participants were told that it is equally likely, that the hypothesis is true or that there is no correlation between the virus and the symptom at all. The participants were told, that the head nurse is in charge of all the patient files, in the form of 100 patient cards. Each patient card on the front side provides information about tested viruses and on the backside information about symptoms.

The cards were then shown to the participants. First the head nurse laid out the front sides of the cards, showing whether a patient had the specific virus (A) or not (-). Then she quickly takes up the cards. Thereby the cards are completely mixed (bold print). Secondly she then laid out the backsides of the cards, showing whether a patient has shown the specific symptom (●) or not (O).

Depending on the experimental condition it varies which cards are shown. The proportion of cards *p*- versus *non-p*-cards and *q*- versus *non-q*-cards resulted from how the parameters were set (low base rate: $P(p)=0.1$, $P(q)=0.2$, high base rate condition $P(p)=0.8$, $P(q)=0.9$).

In the structural condition with constant marginal probabilities (von Sydow; 2002) $P(p|H_D)=P(p|H_I)$ and $P(q|H_D)=P(q|H_I)$ were induced by showing *all* fronts and backs of the cards (after mixing them in between). For the Oaksford and Chater (1994)-model, with $P(p|H_D)=P(p|H_I)$ and $P(q|non\text{-}p \mid H_D)=P(q|non\text{-}p|\ H_I)$, also all cards were first shown with the virus-side facing upwards ($P(p|H_x)$). But after mixing, the symptom-sides only of those patients were shown who had no virus ($P(q|non\text{-}p \mid H_x)$). Thereby I directly provided information on $P(\neg p \wedge q)$ and $P(\neg p \wedge \neg q)$, which should remain constant in this model. No direct information was provided of the *q-/non-p*-marginal probabilities, which are not constant in this model. Similarly, in the Laming (1996)-condition, with $P(q|H_D)=P(q|H_I)$ and $P(p|q \mid H_D)=P(p|q \mid H_I)$, all cards were first shown now with the symptom-side visible. After mixing, the virus-sides of the cards only of those patients were shown, who have had the specific symptom. Thereby I directly provided information on $P(p \wedge q)$ and $P(\neg p \wedge q)$, which are constant in that model and no direct information on the *p*- and *non-p*-marginal probabilities.

All participants were then instructed that the head nurse was not willing to turn over many cards separately. She would only allow *one* card to be turned over on its own. Participants were asked, what card they would select to test their hypothesis. Firstly the participants should suppose the head nurse had put two patient cards in front of them, one of a patient with the virus (A) and one card of a patient without the virus (-) (*p*-card, *non-p*-card).[4] Secondly they should instead suppose a situation in which two patient cards were placed before them, one of a patient with the symptom (●), one of a patient without the symptom (O) (*q*-card, *non-q*-card). In both cases they had to choose which card they would turn over.

Finally, four questions were used (in a frequency format), to survey the participant's estimation of the marginal probabilities resulting in each model, that is: $P(p_{res}|H_D)$, $P(q_{res}|H_D)$, $P(p_{res}|H_I)$, $P(q_{res}|H_I)$. The participants were asked how many of all 100 patients would have the Virus A and how many of all 100 patients would have Symptom T, assuming that the hypothesis is true or false.

Results and Discussion

First the card selections are described, then the estimations of the marginal probabilities.

Card selections

Table 5: Percentages and number of selections of the *p*- and *non-p*-cards and *q*- and *non-q*-cards. (N=72)

	Structural Models					
	Sydow		Oaksford		Laming	
	low	high	low	high	low	high
p	92%, 11	25% 3	83% 10	83% 10	83% 10	58% 7
non-p	8% 1	75% 9	17% 2	17% 2	17% 2	42% 5
q	75% 9	25% 3	58% 7	17% 2	42% 5	45% 5
non-q	25% 3	75% 9	42% 5	83% 10	58% 7	55% 6

[4] The original WST with its four cards may be interpreted as a sequential task, in which the first selection may influence the second, or in which even a planned second selection may influence the first. This would not be modeled by the general approach of Oaksford & Chater (1994). Such effects are minimized by this forced choice design. (Cf. also Klauer 1999.)

Moreover, this design is a severe test of the predicted increase of *non-q*- and *non-p*-card selections: not only the relevance of these cards, but their relative predominance is tested against the normally common *p*-card and *q*-card selections.

von Sydow (2002)-Model For this model a rise in the proportion of *non-q*-selections and *non-p*-selection was predicted for the high base rate condition. The descriptive results are shown in Table 5 and visualized in Figure 1.

Both differences were statistically significant, the *q-/non-q*-effect (Pearson: χ^2 *(1, n=24)=6.0*, one-tailed, *df=1; p<.01*) as well as the *p-/non-p*-effect (Pearson: χ^2*(1, n=24) =10.9*, one-tailed, *p<.001*). The parameters in this experiment were chosen that for this model *EIg(non-q|high)= EIg(q|low)* and *EIg(non-p|high)= EIg(p| low)*. The results of the *q-/non-q*-effect are indeed perfectly symmetrical, the *p-/non-p*-effect descriptively only shows a small *p*-bias. *Within* the high base rate condition more *non-p* than *p* and more *non-q* than *q*-selections were predicted. These cards even became predominant in a statistically significant way (both: χ^2 *(1, n=12)=3.0*, one-tailed, *p<.05*).

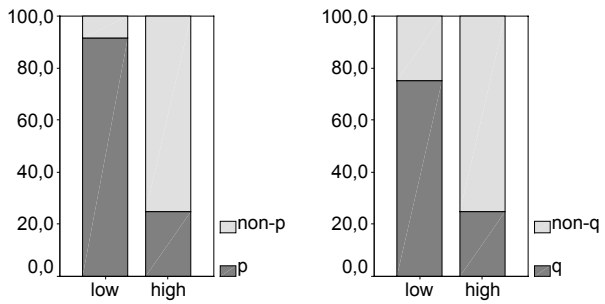


Figure 1: von Sydow-Model: (a) Proportion of *p-/non-p*-card selections and (b) proportion of *q-/non-q*-card selections in high and low base rate conditions.

Oaksford and Chater (1994)-Model This model similarly predicts an increase of *non-q-card* selections in the high base rate condition, but it does not predict an increase of *non-p*-card selections. The results are visualized in Figure 2.

For the *p-/non-p*-cards there was indeed no difference between the low and high base rate condition (Fisher-Yates test (*1, n=24*, one-tailed): *p=0.70*). As also hypothesized, the frequency of *non-q*-card selections was significantly higher in the high base rate condition than in the low base rate condition (Fisher-Yates test (*1, n=24,* one-tailed): *p<.05*). Even the perhaps surprising high rate of *non-q*-card selections in the low base rate condition appears reasonable with regard to the *EIG* and *SEIg* values (cf. Table 4).

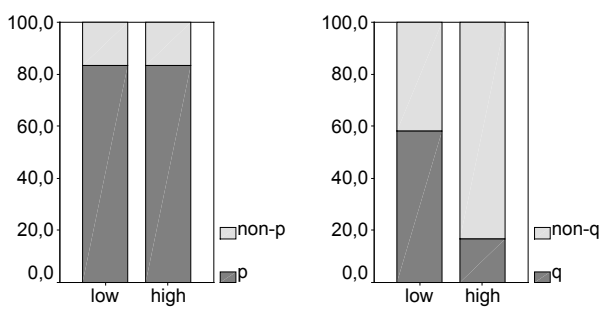


Figure 2: Oaksford-Chater-Model: Proportion of selections.

Laming (1996)-Model Although Laming's proposal was originally only thought as an absurd example, it was modeled and the prediction of a constantly high *non-q*-card selection and a *p-/non-p*-effect was derived.

As expected, no *q-/non-q*-effect was found (Fisher-Yates test (*1, n=23,* one-tailed): *p=.58*). But in difference to the predictions the *p-/non-p*-effect was not significant (Fisher-Yates test (*1, n=24,* one-tailed): *p=.18*). However, even here the results descriptively point in the predicted direction and in the high base rate condition over 40% preferred a *non-p*-card (figure 4).

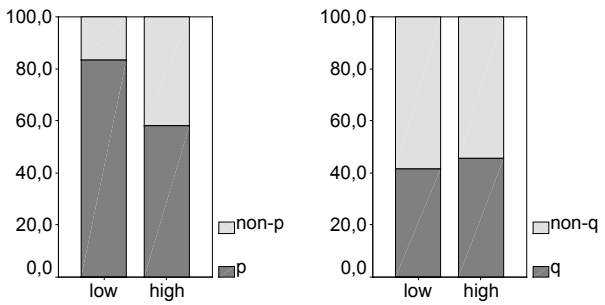


Figure 3: Laming-Model: Proportion of selections.

In summary, the card selections were clearly confirmative for both the von Sydow-model and the Oaksford-Chater-model, and they at least pointed in the predicted direction for the Laming-model.

Estimates of marginal probabilities Participants' estimates of the resulting marginal probabilities was a second depended variable to assess whether the participants fully understood the implications of the induced models.

Only an abridged analysis of these data can be given here. In Table 5 the means and modes of the subjective estimates of the marginal probabilities $P(p_{res})$ and $P(q_{res})$ are shown, if the participants had to assume the rule to be true or false.

An analysis of the data shows that the means are not the appropriate measures to assess the differences between the conditions, since in some cases *two* types of answers clearly predominated. Hence Table 5 also shows the modes (two modes are shown when both have the same frequency or when their frequency differed only by one). It was tested whether the number of cases represented by each mode of the estimations (or by the two modes) is predominant relatively to all other cases not matching that mode(s). This was tested for significance with a Chi^2-test *(df=1, one-tailed, 12≥N≥9).* (For Results cf. Table 5.)

In the von Sydow-Model the modes were all normative. Each mode had a frequency of over 70%. The χ^2-tests showed, that the number of estimations matching the modes was in all but one case significantly higher than all other estimations taken together (Table 5).

In the Oaksford-Chater-Model and in the Laming-Model a relevant number, but not all, of estimations confirmed the predictions. But it will be shown that the deviations also showed an interesting inner consistence.

Table 5: Estimates of the resulting marginal probabilities given the truth (M_D) or falsity (M_I) of the hypothesis. For each model the following values are shown: normative answers for $P(p_{res})$, then for $P(q_{res})$, means of these answers, modes. A mode (or two taken together) got an asterisk (*), if their predominance was also statistically significant. (They also always united over 75% of the answers.)

$P(p_{res})$ $P(q_{res})$		Sydow			Oaksford			Laming		
		Normative	Mean	Mode	Normative	Mean	Mode	Normative	Mean	Mode
Low base rate	M_D	10 20	10 18	10* 20*	10 28	10 19	10* 28;10*	2 20	13 17	2; 20* 20*
	M_I	10 20	10 19	10* 20*	10 20	10 18	10* 18*	10 20	5 19	2* 20*
High base rate	M_D	80 90	73 79	80* 90*	80 98	77 81	80* 98,80*	72 90	76 84	72;90* 90*
	M_I	80 90	76 85	80[5] 90*	80 90	73 58	80* 90; 50	80 90	67 77	72 90*

In both models the results clearly and significantly confirmed the predictions in regard to the constant marginal probabilities, that is in regard to $P(p_{res})$ in the Oaksford-Chater-Model and in regard to $P(q_{res})$ in the Laming-Model). In each of these four cases there was only one mode, which in number outweighed all other predictions significantly. Also as predicted, a change of modes (between M_D and M_I) was found in the Oaksford-Chater-Model in regard to $P(q_{res})$, and conversely in the Laming-Model in regard to $P(p_{res})$. But opposed to the predictions in both models two modes were found, given the hypothesis is assumed to be true. The two modes taken together significantly outweighed all other predictions in all four cases. In all these cases – independent of a high or a low base rate – one of the two modes exactly was the predicted one. The other mode in all cases was consistent with an equivalence interpretation of the hypothesis. In the Oaksford-Chater-Model this second mode of $P(q_{res}|M_D)$ matched the correct estimations of $P(p_{res}|M_D)$. Conversely in the Laming-Model the second mode of $P(p_{res}|M_D)$ exactly matched $P(q_{res}|M_D)$. Hence in both models one set of answers exactly shows the expected changes between M_D and M_I. Another set of answers is consistent with an interpretation of the hypothesis not as implication, but as equivalence. (Based on the low *N* no further analysis of this additional effect was possible.)

In summary, also the results for the estimations show that participants distinguished the tested models. The results for the von Sydow-model were unambiguously positive. In the Oaksford-Chater-Model and the Laming-model a substantial number of answers fully confirmed the predictions. In these models a second group, however, was consistent with an interpretation of the rule as equivalence. Interestingly, the ambiguity of the interpretation of the hypothesis appears to be a function of the induced model.

[5] Also here 70 % (of 11 answers) matched the mode.

General Discussion and Conclusion

The empirical results provide evidence that humans are sensitive both to the structural as well as to the quantitative aspects of the tested Bayesian models.

The card selections largely confirmed the predicted differential effects of structural models and of the card frequencies. Estimates of the resulting marginal probability provide evidence that at least a substantial part of the participants also understood these implications of the models.

Implications for Non-Bayesian Approaches

Approaches that are normatively based on basic formal logics (excluding e. g. fuzzy logics) and its falsificationist interpretation have clear normative predictions in all conditions of the experiment. In each and every case one equally ought to select the *p*- and the *non-q*-card, since these are the only cards by which a (conclusive) falsification could be achieved. The main traditional psychological theories of conditionals, the mental logics theory and the mental model theory (cf. Johnson-Laird & Byrne, 2002) are normatively still tightly linked to the falsificationist research program. But also with their additional psychological assumptions these theories cannot explain the particular pattern of probabilistic results found in this experiment.

Likewise the psychological theories which even break with any concept of normativity, such as the original pragmatic reasoning theory (Cheng & Holyoak, 1985) or the evolutionary social contract theory (Cosmides, 1989; Gigerenzer & Hug, 1992) cannot explain the fit of data to these normative models of reasoning.

The normative models as well as the empirical results of the experiment at least show the incompleteness of all these theories. This has to be said in such a cautious manner, since one has to concede that Bayesian models have not yet explained all the effects predicted by all these quite different theories either.

Implications for Bayesian Approaches

On the one hand the results of the present work show that the discussed Bayesian approaches of hypothesis testing (of single conditionals) need to be extended by a structural component, which determines what parameters are constant in that models.[6] On the other hand this extension (norma-

[6] The structural component proposed in this paper may be regarded as the *microstructure* of a conditional, which perhaps complements the effects of *macrostructure* already discussed in the context of causal Bayes-nets (cf. Waldmann & Hagmayer, 2001)

tively as well as empirically) strongly confirms the general approach of exactly these extended Bayesian models.

Working with a MST and by clearly fixing the preconditions, the results, not only of the model von Sydow (2002), Hattori (2002) and Oaksford & Wakefield (2003) *but also* the original model of Oaksford & Chater (1994, 1998, similarly now Over, in press, Evans et al., 2003) could be supported. (The evaluation of the model of Laming remained ambivalent.) Moreover the objection of Laming (1996) that the assumptions of the discussed basic models are licentious, which in principle affects all models, has been ruled out by introducing experimentally exactly the preconditions of these models.

But these largely confirmative results also show the necessity to extend Bayesian models discussed by the structural aspect examined. From this it results that it is false, both normatively and empirically, to assume that only *one* universal Bayesian model could and should fit all data. Also those authors who have adopted a probabilistic or Bayesian account, mostly still seek a universal model for hypothesis testing or reasoning with conditionals (e. g. Oaksford & Chater, 1994, 1998; also Oaksford & Wakefield, 2003; and even Evans et al., 2003 and Over, in press). Instead the results of my experiment show that additional hidden preconditions need to be taken into account. In this regard I do follow early writings of Evans & Over (1996), which stressed that there is no universal technical measure of uncertainty reduction. On the other hand, in my opinion, only the more sophisticated models in the tradition of Oaksford & Chater do allow a detailed investigation of the phenomena in question. This paper could be seen as contribution towards a synthesis of these positions.

On the larger scale such a synthesis would sustain normative necessity, as the logicistic research program also has done. Nevertheless it allows for a plurality of preconditions, which has been stressed by domain specific accounts. Whether *domain-specific normative Bayesian models* may serve as a more general research program can only be found out by further theoretical analysis and empirical investigation.

Acknowledgments

I am grateful to York Hagmayer, Michael Waldmann and Björn Meder for useful comments on the experiment and on earlier versions of this paper. I also would like to thank Nick Chater and two anonymous reviewers for their helpful comments.

References

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391-416.

Cosmides, L. (1989). The logic of social exchange, *Cognition*, 31, 187-276.

Evans, J. St. B.T., Handley, S.H., & Over, D.E. (2003). *Conditionals and conditional probability*. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, *29*, 321-335.

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*, 356-363.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127-171.

Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *The Quarterly Journal of Experimental Psychology, 55 A (4)*, 1241-1272.

Johnson-Laird, P.- N. & Byrne, R. (2002). Conditionals : A theory of meaning, pragmatics, and inference. *Psychological Review, 109 (4)*, 646-678.

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51*, 1-28.

Klauer, K. Ch. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review, 106, 1*, 215-222.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608-631.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103*, 381-391.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world.* Psychology Press, Hove (GB).

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Science, 5 (8)*, 349-357.

Oaksford, M., & Chater, N., Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5 (3),* 193-243.

Oaksford, M., & Chater, N., Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology. Leaning, Memory, and Cognition, 26, 4*, 883-899.

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling, Probabilities Do Matter. *Memory and Cognition, 31,* 143-153.

Oberauer, K., Wilhelm, O., & Diaz, R.-R. (1999). Bayesian rationality for the Wason selection task? *Thinking and Reasoning. 5 (2),* 115-144.

Over, D. E. (in press). Naïve probability and its model theory. In V. Girotto & P.-N. Johnson-Laird (Eds.) *The Shape of Reason.* Hove: Psychology Press.

Osman, M., & Laming, D. (2001). Misinterpretation of conditional statements in Wason's selection task. *Psychological Research, 65,* 128-144.

Popper, K. R. (1934/2002) *Logik der Forschung.* Mohr Siebeck: Tübingen.

Sydow, Momme von (2002). *Probabilistisches Prüfen von wenn-dann-Hypothesen.* Diplomarbeit (MA-thesis), Department of Psychology, Universität Bonn.

Wason, P. C. (1966). Reasoning, 135-151. In B. M. Foss (Ed..), *New Horizons in Psychology*. Harmondsworth: Pinguin.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82,* 27-58.

A Theoretical Framework to Understand and Engineer Persuasive Interruptions

Muhammad Walji (Muhammad.F.Walji@uth.tmc.edu)
University of Texas Health Science Center at Houston, School of Health Information Sciences,
7000 Fannin, Houston, TX 77030 USA

Juliana Brixey (Juliana.J.Brixey@uth.tmc.edu)
University of Texas Health Science Center at Houston, School of Health Information Sciences,
7000 Fannin, Houston, TX 77030 USA

Kathy Johnson-Throop (Kathy.A.Johnson@jsc.nasa.gov)
NASA Johnson Space Center,
Houston, TX 77058 USA

Jiajie Zhang (Jiajie.Zhang@uth.tmc.edu)
University of Texas Health Science Center at Houston, School of Health Information Sciences,
7000 Fannin, Houston, TX 77030 USA

Abstract

Interruptions are often seen as distracting or sometimes devastating elements that need to be minimized or eliminated. However, interruptions are also used to increase efficiency, productivity, prevent errors, and even influence behavior. Existing theories and taxonomies of interruptions fail to account for the helpful aspects of interruptions. Therefore we propose a theoretical framework to help explain the positive aspects of interruptions. Warnings & alerts, reminders, suggestions and notifications are examples of interruptions that have beneficial outcomes by changing and influencing behavior. We propose a cognitive theory of interruptions based on the properties of the users, their tasks, and best presentations depending on the desired effectiveness of the interruption. Norman's 7-stage action model serves to explain how and why an interruption is accepted, and potential mismatches between the goal of the interruption and the user. Potential applications of this model include better understanding the effects of interruptions, and guidance to design effective and persuasive warnings and alerts, reminders, suggestions and notifications.

Introduction

Interruption has been an active area in human-computer interaction research for some time. A comprehensive review was provided by McFarlane and Latorella (2002). Interruptions are typically defined as a change or disturbance in a process or in people's activities.(Cooper & Franks, 1993; McFarlane & Latorella, 2002) Interruptions are categorized along different dimensions by different researchers, such as source, effect, content, applicability, and duration by Cooper & Franks (1993) and individual properties, methods, meaning, source, channel, change, and effect by McFarlane and Latorella (2002).

Significant research has been expelled in determining how to classify, prevent, minimize, and provide tools to help users deal with interruptions. However, there is little understanding how interruptions can be exploited for positive outcomes, while at the same time minimizing some of their most disruptive properties. After all, interruptions are constantly used to help manage and complete important everyday tasks. Such interruptions also have the ability to influence and change behavior. In order to better understand and explain how interruptions can be engineered to be positive and persuasive we propose a theoretical framework and conceptualization. The theoretical framework may also guide designers on discovering factors to help develop appropriate interruptions.

Effects of Interruptions

Detrimental Effects of Interruptions

The effects of interruptions are generally described as negative Users perceive an interrupted task as being more difficult to complete than an uninterrupted task (Bailey, Konstan, & Carlis, 2000). An interruption is also thought to take longer to process and return back to task when it is unrelated to the task at hand (Cutrell, Czerwinski, & Horvitz, 2001). The added memory load seems to make it difficult for a task to be resumed. It also becomes difficult to remember what task was being processed before the interruption. (Burmistrov & Leonova, 1996; Dix, Ramduny, & Wilkinson, 1995). Further, the complexity of the task being interrupted effects the disruptiveness of an interruption. Interrupting complex tasks inhibits performance, and has no effect on simpler tasks (Burmistrov & Leonova, 1996). Interestingly, people can recall details about interrupted tasks better than uninterrupted tasks.(McFarlane & Latorella, 2002)

People also have individual differences in their ability to respond and manage interruptions (McFarlane & Latorella, 2002). Interruptions also affect performance. Users are thought in general to perform slower on interrupted tasks (Bailey et al., 2000), although some evidence exist that an interruption may actually speed up task completion (Zijlstra,

Roe, Leonara, & Krediet, 1999). However, the actual effect of an interruption will likely depend on the actual tasks being performed, and the interruption itself. There is conflicting evidence if similarity between the interrupted and interrupting tasks has any effect on performance (Bailey et al., 2000; Gillie & Broadbent, 1989).

Timing of interruptions may also have an effect. Interruptions coming early during a search task are described as likely to result in the user forgetting the primary task goal than an interruption arriving later on (Cutrell et al., 2001). The presentation of the interruption are also important. For example aurally presented interruptions are thought to be acknowledged more quickly than visual stimuli. Auditory ongoing tasks are more resistant to interruptions than visual ones (Latorella, 1996) Thermal interruptions have larger detrimental effect than light on disruptiveness and performance (Arroyo, Selker, & Stouffs, 2002). Motion as a notification system is effective compared with static items (Bartram, Ware, & Calvert, 2001). Traveling motions as a visual stimuli are more disruptive than anchored motions (Bartram et al., 2001)

Therefore much effort has been expended to determine the negative effects of various interruptions and their modalities. However there are also different perspectives from which the effects of interruptions may be viewed. Indeed an interruption may be devastating to the task in progress. But when looking at the individual performing various tasks, the interruption may not have a detrimental impact on the whole. Most research has focused on the task level, which may be an inappropriate level of analysis in some cases.

Beneficial Effects of Interruptions

Types of interruptions that may serve beneficial purposes include warnings and alerts, reminders, notifications and suggestions. Of course warnings and alerts etc., may not always be interruptions. We define a warning and alert etc. as an interruption when it causes a change or disturbance in a person's activity or behavior. Table 1 summarizes the characteristics of these interruptions. We provide examples in a healthcare context, although these types of interruptions would also exist in other domains. Our examples are also technologically focused and include persuasive interruptions embedded into computer systems, mobile devices, and medical equipment which are increasingly being used in healthcare.

Warnings & Alerts are usually a sign or signal of something negative occurring, or a notice to be careful. They are intended to make people aware of an impending danger or difficulty. For example, drug interaction warnings embedded into drug prescribing systems warn doctors and pharmacists about dangerous drug-drug interactions when prescribing or filling a prescription. These warnings are designed to interrupt the current task, and alert the clinician to a potential adverse event. Although such warnings may be critical in preventing errors, it is found that in practice such warnings are often ignored or overridden (Wilson, 2003), suggesting the need for better designed warnings. Hospitals are increasingly 'buzzing' with auditory alerts from a variety of medical equipment (Meredith & Edworthy, 1995). The purpose of such devices are to monitor patients and alert physicians or nurses when they need to take action. However, there is rarely any synchronization or awareness between the large number of standalone medical equipment emanating various alerts and tones; resulting in many ignored warnings.

Warnings and alerts are often urgent and need to be handled quickly. Warnings and alerts may either have an explicit or implicit action associated with them. For example a drug interaction warning may indicate explicitly that there is a potential interaction with a drug and provide a list of medications that may be suitable replacements. An audible alert may be more implicit, simply indicating an off nominal state, without providing any explicit instructions or actions.

Reminders are a form of interruption that cause an individual to remember or recall an event. Clinical decision support systems often remind physicians of standard tests or procedures that conform to clinical practice guidelines. (Bates et al., 2003) Such reminders are deemed important as they provide a mechanism to foster uniformity in treatment and to assist in managing the burgeoning costs of healthcare. These reminders often occur while the physicians is documenting or ordering the tests and procedures. Medication reminders may also assist patients in adhering and complying with their medication regimens (Bennett & Glasziou, 2003). Although the urgency or importance of reminders may vary, many will include an explicit associated action. For example a medication reminder may announce the time, dose and route for the drug.

Suggestions are ideas or proposals that are propagated to individuals. Patients often receive suggestions and recommendations from their care-givers. For example diabetics are urged to exercise more and eat healthier. Physicians may be informed that their patient may be eligible for a particular clinical trial. Pharmaceutical companies also engage in suggestive practices to prescribers when they promote their particular brand of medication. Such suggestive interruptions can be from face-to-face encounters with a pharmaceutical sales representative or through the use of sponsored drug reference databases. Suggestions are unlikely to be of high urgency or importance. But effective suggestions may explicitly state associated actions that are recommended.

Notifications are usually described as the process of informing. Notifications are defined as the most generic type of interruption, with the least degree of importance or urgency. A notification may purely be informational in purpose with no explicit instruction for action. For example

a notice stating the availability of a patient's lab results informs a physician that their requested order is ready. However, notifications may lead to actions implicitly without specific instructions. For example the lab test may indicate that a particular patient needs an immediate surgical procedure. Therefore notifications may lead to implicit actions.

Table 1: Beneficial Interruptions and their Characteristics

Interruption Type	Importance / Urgency	Action
Warnings & Alerts	High	Implicit or Explicit
Reminders	High-Low	Explicit
Suggestions	Medium – Low	Explicit
Notifications	Low	Implicit

Persuasive Interruptions

Fogg (1998) suggests computers and technology can be persuasive (change attitudes or behavior) as tools, social actors and/or media. We suggest that technology-based interruptions can be designed so they too can influence behavior and attitudes in order to achieve positive outcomes. In fact beneficial interruptions described earlier as warnings and alerts, reminders and suggestions disrupt a person's current task, and may cause them to change their behavior. Of course not all positive interruptions need to change or influence behavior. The persuasiveness of interruptions may be directly linked to the interruption type and their corresponding importance to deliver a particular message. For example warnings may be high in importance, and need to influence a change in behavior immediately and therefore very persuasive. While a notification, which is just informational in content, may not influence behavior and therefore may not be particularly persuasive.

Theoretical Framework

We propose a theoretical framework for interruptions (figure 1) to help explain the different dimensions that are involved in making an interruption persuasive and beneficial. Table 2 shows the details of the framework in a form of taxonomy.

User Properties

Individuals or users that are affected by interruptions are likely to have unique characteristics and properties. Therefore it is important to identify key features that may impact the effectiveness of interruptions and how they respond and deal with them. For example a physician has different characteristics than a nurse. A challenge in producing effective interruption are to deliver them when most opportune and least detrimental. Therefore a users location, environment, time of day (or week or year), or schedule (in Outlook for example) may be exploited to establish if they can be interrupted. Horvitz et al have explored the use of subtle clues in design of attentive user interfaces to discover the attention of users combined with user preferences in design of notification platform to intelligently route messages (Horvitz, 1999).

Task Properties

In addition to determining user characteristics, it is also important to determine properties of the interrupted and interrupting tasks. Certain tasks may be particularly susceptible to the detrimental effects of interruptions. However, determining a user's current task is challenging. Computer based tasks may be more amenable to discovering current task or workload. But in more complex, dynamic or distributed domains, it is likely that the users will interact with a multitude of (unlinked) devices including phones, pagers, PDA's, among others.

Various methods to determine user interrupt-ability have been explored. Instant messaging applications allow users to indicate their current availability. Alternatively, task complexity may be automatically measured. The number and type of applications the user has open, or number of key strokes, or mouse clicks within a certain time period may indicate the user's workload. The user's contextual information may also be exploited, such as time of day or week. A user may conduct certain tasks at certain times of a week. However, many users do not follow a rigid schedule and may elect to make changes. Another approach has been to discover "activity awareness" between groups which take into account situational, group, task and tool factors and subsequently provide a notification system to indicate availability. (Carroll, Neale, Isenhour, Rosson, & McCrickard, 2003) However, further work is needed to discover how best to determine current task properties in order to present an interruption at the most optimal time.

Presentation

In addition to user and task properties, the presentation of an interruption may be critical. The presentation of an interruption involves two stages. First, the interruption must alert the user of its presence. Heat, light, sound, vibration, and motion may capture attention differently with different efficiencies. Second, a message representation must be delivered. Analysis of the user, task and priority of the interruption context will help determine the appropriate mode of interruption. The presentation may also differ depending on type of interruption and on the device used to interrupt. A visual pop-up may effectively capture a users attention while using a computer, but may be ineffective on a cell phone stowed in coat pocket.

In addition to being effective and minimally disruptive, the message of the interruption can also be engineered to be persuasive. In a multi-tasked environment, users are presented with a multitude of interruptions and are constantly deciding whether to act upon the interruption. If

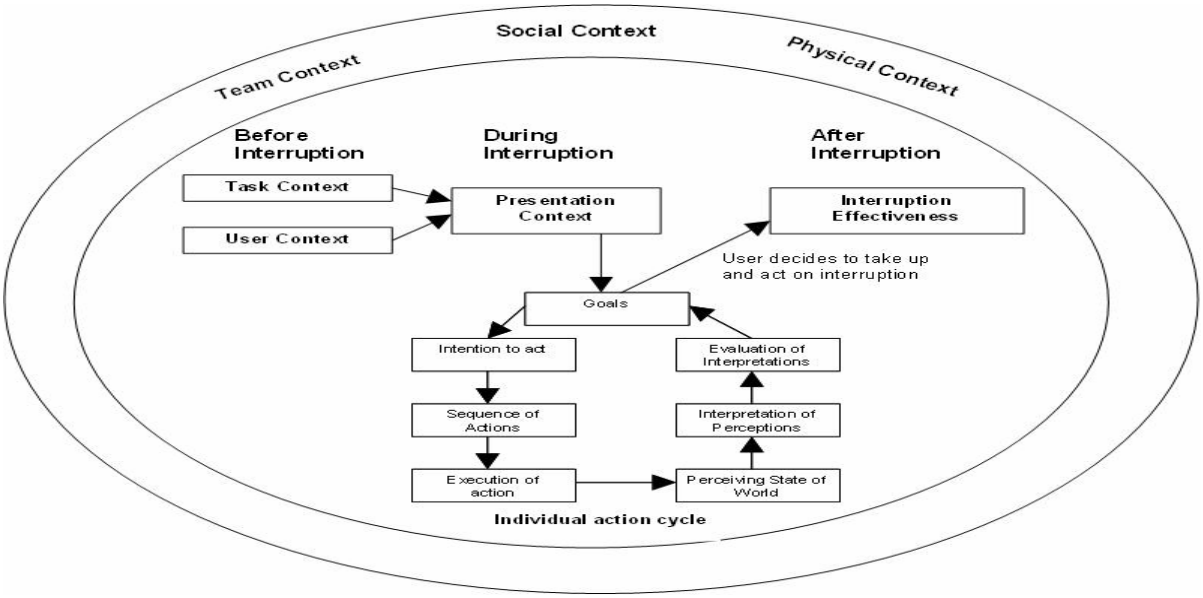


Figure 1: Cognitive theory of persuasive interruptions

different interruptions (such as a warning or notification) are presented with the same degree of persuasiveness they may be handled in the same manner. However, persuasive elements such as positive reinforcement, personalization, and social cues amongst others can also be used to enhance the persuasiveness of an interruption when appropriate. Currently there is little research on how modifying the persuasiveness of a message of an interruption effects its acceptance.

Individual Action Cycle

Norman's 7-stage action model has been incorporated into the cognitive theory in order to help explain at the individual action level why an interruption is accepted and acted upon (Norman, 1988). The seven stages are divided into three categories, one for goals, 3 for stages of executions and 3 for evaluation. The goal stage may be particularly important because an individual's perceptions or intentions may need to be related to intention of the interruption itself as personified by its presentation. The stages of execution are also useful in determining if the suggested interruption can be acted upon. The evaluation stage where the individual perceives the state of the world after executing an action may also assist in determining the success of an interruption. Therefore the 7-stage action model provides a useful perspective in helping to explain how modifying the presentation of an interruption impacts discrete stages of individual action.

Interruption Effectiveness

The effectiveness of an interruption will largely depend upon the original goal and perspective used. A drug interaction warning that interrupts a physician while inputting order entry may be effective if it results in a change of medication; as it may avoid a hospitalization for the patient. However, it may also cause the physician to lose focus and forget the original task. Therefore it is important to clarify the perspective from which effectiveness is judged. In our model we propose cognitive, perceived value and performance based measures to evaluate and engineer interruptions once the perspective has been defined. Cognitive factors may include loss of memory or disruptiveness of interruptions. Perceived value factors such as annoyance and anxiety are often associated with interruptions. Interruptions affect performance, by changing time to complete tasks, providing opportunities for errors, and forgetting to resume previous tasks.

Similarly they may effect financial performance or result in a more favorable outcome (such as prevention of hospitalization) In our model, information from the context of the user, tasks and presentation can be exploited in order to find an optimal balance between cognitive, perceived value and performance measures depending on the perspective and desired outcomes.

Assessing Context

The surrounding conditions or circumstances that make up the environment around an individual may provide important information in order to successfully deliver an interruption. The dynamics of interruptions in team environments are different than those of single individuals. In team environments, a team member can intercept an interrupting activity for another team member who is already engaged in a previous task. An audible interruption targeted to one team member may interrupt the work of colleagues nearby. Or an interruption for one individual may result in a cascade of interruptions for others.

Table 2: Taxonomy of Persuasive Interruptions

		Examples
User Properties	Individual characteristics of users, their contextual situations and preferences	
Context	Where is the individual? Where can an individual be interrupted? Is an interruption more appropriate at a certain location or time	Hospital, Emergency Room (ER), Attending to critical patient, etc.
Characteristics	What are the individual characteristics of users? What are their strengths and limitations?	Expertise, skills, knowledgebase, age, education, cognitive capacities and limitation
Task Properties	The properties of the interruption itself and the task it will interrupt	
Interruption Type	What is the intent of the interruption?	Warning, alert, reminder, suggestion or notification
Interrupted Task Type	What task will be interrupted?	Work related (computer based, meeting etc.), Social (lunch, sleep etc.)
Task Interrupt Scale	How important is the task to be interrupted?	Low, Medium, High
Stage of Interruption	What is the stage of the current task?	Goals, Intention to Act, Sequence of actions, Execution of action sequence, Perceiving state of the world, Interpreting the perception, Evaluation of interpretations
Broadcast or Single Interruption	Is the interruption in the context of team or collaborative environment, or individual environment?	Individual, small team, large team, etc.
Presentation	Factors addressing how the interruption can be presented to the user	
Customization	To what degree is the presentation customized?	Generic, Personalized, Targeted or Tailored
Mode of Interruption	How will the user be alerted of the presence of the interruption?	Heat, lights, sound, vibration, and motion
Display type	How will the message be communicated?	Prompt, pop-up, voice alert
Device	What device will be used to convey the message of the interruption	Personal Computer (PC), Personal Digital Assistant (PDA), Telephone, Cell phone, Pager
Persuasive elements	What types of persuasive techniques are incorporated into the interruption?	Media, Tool, Social Actor, Positive reinforcement, personalization, credibility etc.
Interruption frequency	How often will the interruption be presented?	Once only, more than once, every hour etc.
Resumption method	How will the individual be assisted to resume their original task?	Log of previous tasks, reminder of previous task, screenshot of previous state etc.
Interruption Effectiveness	Assessing the effectiveness of the interruption	
Perspective	Who is the intended beneficiary of the interruption? What is the net benefit?	Physician being interrupted, Patient, Healthcare system
Cognitive	What is the cognitive impact of the interruption on the individual?	Loss of memory, disruptiveness, number of errors
Perceived value	What are the individual perceptions of the interruption?	Annoyance, anxiety, interest, boredom, curiosity
Performance	How does the interruption effect the performance of the interrupted task? To what degree is the task associated with the interruption completed?	Completion of tasks, time to complete task, number of errors, dollars saved

Benefits of Persuasive Interruption Model

Other models of interruptions have been developed in an attempt to eliminate, minimize or manage the detrimental effects of interruptions. However, these models fail to describe the positive effects of interruptions. Latorella's (1996) Stage Model of interruption management is a detailed description of how people may manage an interruption and how it effects a current task in terms of detection, distraction, disturbance and disruption. The model of persuasive interruptions is more concerned with dimensions of the user, task and presentation properties and how that influences the effectiveness of the interruption. We suggest Norman's 7-stage action model can explain how and why an individual receives an interruption. McFarlane

(2002) has also proposed a taxonomy of human interruptions that includes elements such as source, individual characteristics, method of coordination etc. Our model incorporates features of McFarlane's taxonomy but is more operationalized and detailed. For example McFarlane suggests looking at the individual characteristics of users, while we expand this view to also consider other relevant contextual features, such as time, location and environment.

Conclusion

In this work we identify and discuss four types of beneficial interruptions: warnings and alerts, reminders, suggestions and notifications. We then propose a theoretical framework and taxonomy in order lay the foundation to develop guidelines for persuasive interruption design.

Future work will improve the framework by experimentally testing and validating the model of persuasive interruptions. We are particular interested in discovering the effects of various persuasive techniques when applied to the message of an interruption. Potential applications of this model include better understanding the effects of interruptions, and guidance to better design effective and persuasive interruptions.

Acknowledgments

Supported by a training fellowship from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NLM Grant No. 5T15LM07093), and supported in part by grant No. NCC 2-1234 from NASA's Human-Centered Computing and Intelligent Systems Program.

References

Arroyo, E., Selker, T., & Stouffs, A. (2002). *Interruptions as multimodal outputs: Which are the less disruptive?* Paper presented at the 4th IEEE International Conference on Multimodal Interfaces (ICMI'02). Institute of Electrical and Electronics Engineers.

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2000). *Measuring the Effects of Interruptions on Task Performance in the User Interface.* Paper presented at the IEEE Conference on Systems, Man and Cybernetics, Nashville, TN.

Bartram, L., Ware, C., & Calvert, T. (2001). *Moving icons, detection and distraction.* Paper presented at the Human-Computer Interaction – INTERACT 2001.

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., et al. (2003). Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc, 10*(6), 523-530.

Bennett, J., & Glasziou, P. (2003). Computerised reminders and feedback in medication management: a systematic review of randomised controlled trials. *Med J Aust., 178*(5), 217-222.

Burmistrov, I., & Leonova, A. (1996). *Interruption in the Computer Aided Office Work: Implications to User Interface Design.* Paper presented at the Human-Computer Interaction: Human Aspects of Business Computing. Proceedings of EWHCI'96, Moscow.

Carroll, J. M., Neale, D. C., Isenhour, P. L., Rosson, M. B., & McCrickard, D. S. (2003). Notification and awareness: synchronizing task-oriented collaborative activity. *International Journal of Human-Computer Studies, 58*(5), 605-632.

Cooper, R., & Franks, B. (1993). Interruptibility as a Constraint on Hybrid Systems. *Mind and Machines, 3*, 73-96.

Cutrell, E., Czerwinski, M., & Horvitz, E. (2001). *Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance.* Paper presented at the Human-Computer Interaction--Interact '01.

Dix, A., Ramduny, D., & Wilkinson, J. (1995). *Deadlines and Reminders: Investigations into the Flow of Cooperative Work.* (No. Technical Report RR9509): University of Huddersfield.

Fogg, B. (1998). *Persuasive computers: perspectives and research directions.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similiarity, and complexity. *Psychological Research, 50*(4), 243-250.

Horvitz, E., Jacobs, A., & Hovel, D. (1999, July 1999). *Attention-Sensitive Alerting.* Paper presented at the Proceedings of UAI '99, Conference on Uncertainty and Artificial Intelligence, Stockholm, Sweden.

Latorella, K. A. (1996). *Investigating Interruptions: Implications for Flightdeck Performance.* Unpublished Doctoral Dissertation, State University of New York at Buffalo, Buffalo.

McFarlane, D. C., & Latorella, K. A. (2002). The Scope and Importance of Human Interruption in HCI Design. *Human-Computer Interraction, 17*(3), 1-62.

Meredith, C., & Edworthy, J. (1995). Are there too many alarms in the intensive care unit? An overview of the problems. *J Adv Nurs., 21*(1), 15-20.

Norman, D. A. (1988). *The Design of Everyday Things.* New York: Doubleday.

Wilson, J. (2003). Crying wolf! Why computerised drug interaction alerts need an overhaul. *The Pharmaceutical Journal, 271*(7276), 708.

Zijlstra, F. R. H., Roe, R. A., Leonara, A. B., & Krediet, I. (1999). Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology, 72*, 163-185.

Revising Causal Beliefs

Clare R. Walsh (Clare_Walsh@brown.edu)
Department of Cognitive & Linguistic Sciences, Box 1978,
Providence, RI 02912 USA

Steven A. Sloman (Steven_Sloman@brown.edu)
Department of Cognitive & Linguistic Sciences, Box 1978,
Providence, RI 02912 USA

Abstract

The aim of our studies is to examine how contradictions affect causal beliefs. For example, the discovery that your colleague Mark who has been following diet A suffers from iron deficiency may lead you to revise your belief that diet A provides a sufficient supply of iron. Would you also revise your belief that it causes you to lose weight? Experiment 1 shows that our belief that Mark will lose weight is reduced after encountering the contradiction. Experiment 2 shows that people are also less likely to believe that others will lose weight. The results suggest that people resolve contradictions by generating explanations that revise their causal model.

Belief Revision

As we go through life, we are constantly changing our beliefs. We give up old attitudes and we add new ones. When we discover credible information that contradicts our existing beliefs, then rationally we must revise our beliefs in order to restore consistency. Our aim is to examine how this is done.

For example, imagine you believe the following:

If the drink contains sugar, then it tastes sweet

and you believe that in fact:

The drink on the table contains sugar.

But when you taste the drink you discover that it is not sweet leading you to withdraw your earlier inference. Much of our everyday reasoning is non-monotonic. People frequently overturn old conclusions in the light of new evidence. They readily suppress valid deductive inferences when they are presented with new information (Byrne, 1989; Byrne, Espino & Santamaria, 1999).

The discovery that the drink does not taste sweet may also lead us to revise our initial beliefs. Perhaps the drink does not contain sugar. Or it may be that a drink containing sugar doesn't necessarily cause it to taste sweet. For example, perhaps it contains a lot of lemon which suppresses the sweetness. Both possibilities are sufficient to resolve the inconsistency so one question is how to choose from among these possibilities. Logic provides no guidance (Revlin, Cate, & Rouss, 2001). The problem has been studied in philosophy (e.g., Harman, 1986) and in artificial intelligence (e.g. Gärdenfors, 1988). The focus there has been to develop formal principles to guide rational belief change (e.g., Alchourrón, Gärdenfors, & Makinson, 1985). The major principle underlying all existing theories of belief revision is that we should minimize the amount of information that is lost when we revise our beliefs (e.g., Gärdenfors, 1988; Harman, 1986; James, 1907).

Despite the extensive research in developing formal models of belief revision, evidence on how people revise their beliefs is sparse. The way they do so may be very different from the formal systems developed in artificial intelligence (Legrenzi, Girotto, & Johnson-Laird, 2003). Work on attitude change does suggest that, in the face of new evidence, people will treat all contextually relevant beliefs as modifiable in order to increase consistency (Festinger, 1957; Simon & Holyoak, 2002; Thagard, 1989).

What should we do when we discover information that contradicts a causal belief? To the extent that causal relations describe law-like generalizations, the minimal change may be to retain the causal belief and to give up some of the factual information that led to the contradiction. Alternatively, when causal beliefs describe a theory, then evidence to the contrary is reason to dispense with the theory (Popper, 1959).

Studies show that people frequently focus on conditionals more than categorical facts when they revise their beliefs (e.g., Elio & Pelletier, 1997) although less so when they describe familiar causal than unfamiliar relations (Byrne & Walsh, in press, Walsh & Byrne, 2004) and the tendency to do so will depend on the initial degree of belief in the conditional (Diuessaert, Schaeken, De Neys & d'Ydewalle, 2000). Furthermore, when people revise a causal statement, they rarely reject it outright (Byrne & Walsh, 2002). Instead, they may revise their interpretation of the relation (Johnson-Laird & Byrne, 2002). They frequently modify their causal belief by stating that the contradictory example is an exception to the rule or by imagining that possible disabling conditions are present (i.e., factors that prevent a cause from producing its usual effect; Byrne & Walsh, 2002). And people retain a higher degree of belief in a causal conditional when there are few available disabling conditions (Elio, 1997).

We address three questions which examine how a contradiction to a causal belief impacts on our belief system. The questions provide clues to the processes underlying belief change. In Experiment 1, we examine whether resolving a contradiction to a causal belief leads people to revise their judgment about that single belief or whether it has implications for other causal judgments. In attempting

to minimize the changes they make, people may alter a causal belief in a way that leaves other causal beliefs unchanged. Alternatively, if people introduce disabling conditions to modify a causal belief, this may lead to changes which resonate through the belief system. Support for this latter view comes from the finding that when people discover that a cause does not produce an expected effect, they may doubt whether other expected effects will follow (Walsh & Johnson-Laird, 2004).

People may mentally construct a causal model to represent the causal relations between events (e.g., Sloman & Lagnado, 2004). In Experiment 2, we examine whether people use their existing causal model to generate explanations about the situation in which the contradiction occurred or whether their explanations involve a change to the causal model itself. In addition, we examine whether people generate just one or several alternative hypotheses to explain the contradiction.

Experiment 1

We propose that when people encounter a contradiction to a causal belief they generate an explanation for why the cause occurred without the effect (Walsh & Johnson-Laird, 2004). Rather than giving up their belief in the causal relation, they tend to modify it (Byrne & Walsh, 2002) and they may do so by specifying certain conditions that will disable the relation. For example, imagine that despite your belief that exercise causes weight loss, you discover that Anne has exercised but didn't lose weight. Rather than inferring that exercise is not effective you may decide that it is only effective if it is cardiovascular or if you do not at the same time consume more calories. If our hypothesis is correct, then the explanation may influence other causal judgments, for example, whether the exercise increased Anne's fitness level. Our experiment was designed to test this hypothesis.

Method

We constructed six experimental problems. Each problem began by stating a pair of causal statements with a common antecedent, for example:

Jogging regularly causes a person to increase their fitness level.

Jogging regularly causes a person to lose weight.

To measure initial belief in the first statement, we asked the following question:

Tim jogged regularly. What is the probability that his fitness level increased?

We then introduced a contradiction to the second causal statement and we examined whether this influenced their belief in the first statement:

Sam jogged regularly but he did not lose weight. What is the probability that his fitness level increased?

The six problems were of the same form but with different causal materials.

We also constructed two control problems, which did not involve any contradiction and again we measured whether there was any belief change. For example:

Sam jogged regularly and he did lose weight. What is the probability that his fitness level increased?

Participants responded to the questions by giving a number between 0 and 100 (where 0 = definitely not and 100 = definitely).

The participants were 23 undergraduates of Brown University who took part in return for payment or course credit.

Results

Table 1 shows the mean probability ratings before and after the contradiction. The probability of the second consequent was rated as significantly lower after reading the contradiction (mean = 59%) than before reading the contradiction (mean = 77%; t (df = 22) = 6.07, $p < .001$). The pattern occurred for 21 out of 23 participants and the remaining two were tied. The pattern also occurred for each of the six types of semantic content and there was no significant difference in the amount of belief change between the different contents. In the control problems, there was no significant change in the probability of the second consequent when no contradiction was presented ($p > .5$).

Table 1: Mean probability ratings for experimental and control problems in Experiment 1

Problem format:	If A then B If A then C
Experimental Problems	
Given:	Probability of B (0-100)
A	77
A and not C	59
Control Problems	
A	81
A and C	82

The results show that when people receive information that contradicts one causal statement, they will be less confident that other expected consequents will follow from the same cause. One explanation for our finding is that people resolve the contradiction by introducing conditions which would disable the relation. These disabling conditions may also reduce the probability that other consequents will follow from the same cause. An alternative explanation is that people are generally less confident about what they've been told, perhaps because they consider the source less credible, so they reduce their judgments. In our second experiment, we compare these hypotheses.

Experiment 2

The aim of our second experiment was twofold. First, we wanted to know whether people's causal judgments depend on the explanation that was generated for the contradiction

and on that explanation alone. If their probability judgments depend on their explanations, then their judgments should be predictable from their explanation regardless of the contradicted fact. In contrast, if a contradiction just reduces confidence, then their probability judgments should vary with contradiction, and not with the explanation. We tested this by explicitly asking participants to generate an explanation for the contradiction before making a causal judgment, e.g.,

(1) Anne jogged regularly but she didn't lose weight.
Why?
What is the probability that her fitness level increased?

We then asked participants to use this explanation to make another causal judgment. For example, if a participant gave the explanation that Anne's appetite increased, then we asked them the following question:

(2) John jogged regularly and his appetite increased.
What is the probability that his fitness level increased?

If people use their stated explanation (and not the contradicted fact) to make the causal judgment in (1) and they don't consider any other hypotheses, then we expect the probability judgments in (1) and (2) to be equal. Previous research has shown that people frequently neglect to consider alternative hypotheses (e.g., Klayman & Ha, 1987). However, if reasoners do consider other explanations or if their causal judgments are reduced merely because they have less confidence in what they have been told, then we expect these judgments to differ.

The second question that we address in this study is whether people draw on information that already exists in their causal model to generate an explanation for a contradiction or whether resolving a contradiction leads people to revise the causal model itself. We did this by asking participants two further questions. Before reading the contradiction we asked them the following:

(3) Tom jogged regularly.
What is the probability that his fitness level increased?

And after reading the contradiction and generating the explanation, we asked them the following:

(4) Mary jogged regularly and you don't know if her appetite increased.
What is the probability that her fitness level increased?

If a reasoner's causal model already contains information about the relation between appetite and fitness level and they use this information in answering (3), then we expect their responses to questions (3) and (4) to be equal. But if they change their causal model when resolving the contradiction, we expect their answer to these two questions to be different. This study examines these two questions.

Method

We used the same six pairs of causal beliefs as used in the experimental problems in Experiment 1. Each pair was followed by five questions as presented in Table 2. The questions were presented orally to the participants and the experimenter recorded their responses. The first question again measured participants' initial belief in the probability of the first conditional. Question 2 introduced a contradiction to the second conditional. This time we explicitly asked participants to generate an explanation for why the contradiction might have occurred before asking them to rate the probability that the consequent of the first conditional occurred.

The following three questions measured the probability of the consequent of the first conditional under different conditions, namely, when the explanation given in question 2 was either, unknown, absent, or present. For example, take the problem described in Table 2. If participants answered question 2a by saying that Kevin was taking sleeping pills, then in question 3 we told participants that Frank was worried but it is not known if he is taking sleeping pills and we asked for the probability that he had difficulty concentrating. In question 4, we asked for the same probability judgment given that Helen was not taking sleeping pills. And finally, in question 5, we asked for the probability given that Evelyn was taking sleeping pills.

Table 2: The format of the problems used in Experiment 2

Worrying causes difficulty in concentrating. Worrying causes insomnia.
1. Mark was worried. What is the probability that he had difficulty concentrating?
2. a. Kevin was worried but he didn't have insomnia. Why? b. What is the probability that he had difficulty concentrating?
3. Frank was worried but *you don't know if the explanation holds.* What is the probability that he had difficulty concentrating?
4. Helen was worried and *you know that the explanation does not hold.* What is the probability that she had difficulty concentrating?
5. Evelyn was worried and *you know that the explanation does hold.* What is the probability that she had difficulty concentrating?

The experiment allows us to examine whether the probability ratings depend on the single explanation that was generated for the contradiction. The experiment also allows us to test whether participants change their causal model before and after the contradiction.

The participants were 20 undergraduates of Brown University who took part in return for payment.

Results

The mean responses for each question are presented in Table 3. The results replicate the finding of Experiment 1. The probability of the second consequent was rated as significantly higher in question 1 before reading the contradiction (mean = 85%) than in question 2 after reading the contradiction (mean = 63%; $t = 5.03$, $p < .001$).

Table 3: Mean probability ratings for each of the five questions in Experiment 2

Problem format:	If A then B If A then C
Given:	Probability of B (0-100)
1. A	85
2. A and not C	63
3. A and explanation unknown	71
4. A and explanation absent	85
5. A and explanation present	62

Our second finding was that responses to question 2 and question 5 did not differ significantly ($t = 0.60$, $p > .5$) and this pattern occurred for all six types of problem content. For problems in which the contradiction reduced the judged probability of B (response to question 2 was lower than to question 1), participants gave the same answer to question 2 and 5 for 53% of problems. We would expect greater variety if participants were considering multiple hypotheses. Hence the results are consistent with the view that in many cases, people consider just the one hypothesis given in their explanation and they fail to consider other possibilities. They allow this hypothesis to mediate their later causal judgments without considering the possibility that they are wrong (see also Shaklee & Fischhoff, 1982).

Finally, our results suggest that people resolve contradictions by making a change to their causal model. Ratings for question 1 were significantly higher than for question 3 when the explanation was unknown (mean = 71; $t = 5.37$, $p < .001$). People do not merely change their causal judgments about the specific case in which the contradiction occurred. They extend these changes to new situations. Responses to question 1 did not differ significantly from responses to question 4 (mean = 85; $t = 0.1$, $p > .8$). People do not generally resolve contradictions by drawing on events that they have already represented in their causal model.

We also examined the nature of explanations given for the inconsistency. The most common explanation was to introduce a disabling condition which would prevent the cause from producing its usual effect. 74% of responses were of this type. In many cases, the conditions disabled the cause from both consequences. For example, the fact that worry did not lead to insomnia may be explained by the fact that the person did relaxation exercises. This in turn may reduce the probability that worry will lead to a difficulty in concentrating. The next most common type of response was to suggest that the level or amount of the cause was not sufficient to produce the effect, for example, there was not enough sugar in the drink or the person was not very worried. 18% of responses were of this type. In both cases, the pattern of responses and significance ratings for the probability of B were the same as for the overall ratings.

Discussion

The results of our experiments confirm previous findings that people prefer to modify than to give up a causal belief when they encounter a contradiction. The results also give us insight into how those modifications alter other causal judgments. In Experiment 1 we showed that when we discover a situation where a cause does not produce an expected consequence, we become less certain whether the cause will lead to other expected consequences in this instance. The results of Experiment 2 show that we also become uncertain about whether other expected consequences will follow in a situation involving a different agent.

The findings suggest that discovering a contradiction can lead us to change the information that we use to make causal judgments. The contradiction makes salient or forces us to imagine conditions that may impact on these judgments. Hence the basis for making our judgments has changed.

The results reaffirm the view that monotonic logic systems are inadequate for understanding how people reason. In most cases, when people reason from cause to effect the conclusion is indeterminate. It is rarely possible to state all of the conditions in which the cause will necessarily produce an effect (e.g., Johnson-Laird & Byrne, 2002). One approach used by artificial intelligence researchers is to make the default assumption that all of the necessary conditions are present unless there is information to the contrary (Minsky, 1975). Similarly, people may mentally construct models that do not represent all of the information explicitly (Johnson-Laird & Byrne, 1991) although they may consider additional factors if they come to mind easily (Cummins, Lubart, Alksnis & Rist, 1991).

An alternative way to approach these problems is to assume that judgments are probabilistic. Probabilistic judgment does not require specification of all of the conditions that prevent a cause from having its usual effect; the judgment merely reflects the likelihood that this occurs.

Our results suggest that when people encounter a contradiction they generate explanations. The most common type of explanation is to describe a condition that disables the cause from its effect. These disabling conditions may often be ones that people haven't previously considered and as a result they introduce these new conditions into their causal model. These new conditions may have the effect of disabling the cause from other possible consequences.

Introducing a new disabling condition into a causal model could have two possible results. One is that it could explain

why the effect does not always follow from the cause but the probability judgment may remain unchanged. A second possibility is to decide that this new condition should lower the probability that the effect will follow from the cause. Our results suggest that people tend to use the second approach. When our participants considered new conditions, they used these to reduce their probability judgment further.

Acknowledgments

We thank Phil Johnson-Laird for discussions on this topic. This research is supported by NASA grant NCC2-1217 to Steven Sloman.

References

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the Logic of Theory Change. *Journal of Symbolic Logic, 50*, 510-530.

Byrne, R.M.J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.

Byrne, R.M.J., Espino, O. and Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40*, 347-373.

Byrne, R. M. J. & Walsh C. R. (2002). Contradictions and Counterfactuals: Generating Belief Revisions in Conditional Inference. In W. Gray & C. Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates (pp. 160 - 165).

Byrne, R. M. J. & Walsh C. R. (in press). Resolving Contradictions. To appear in: Johnson-Laird, P.N. & Girotto, V. (Eds.). *The shape of reason: essays in honour of Paolo Legrenzi.*

Cummins, D.D., Lubart, T., Alksnis, O. & Rist. (1991). Conditional reasoning and causation. *Memory and Cognition, 19*, 274-282.

Diuessaert, K, Schaeken, W., De Neys, W. & d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. *Current Psychology of Cognition*, 19, 277-288.

Elio R. (1997). What to believe when inferences are contradicted. In M. Shafto & P.Langley (Eds). *Proceedings of the 19th Conference of the Cognitive Science Society.* Hillsdale: Erlbaum (pp. 211-216).

Elio, R. & Pelletier, F.J. (1997). Belief change as propositional update. *Cognitive Science, 21*, 419-460.

Festinger, L. (1957). A theory of cognitive dissonance. Oxford: Row, Peterson.

Gardenfors, P. (1988). *Knowledge in flux.* Cambridge, MA: MIT Press.

Harman, G. (1986). *Change in view.* Cambridge, MA: MIT Press.

James, W. (1907). *Pragmatism – A New Name for Some Old Ways of Thinking*. New York: Longmans, Green & Co.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Erlbaum.

Johnson-Laird, P.N. and Byrne, R.M.J. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review, 109*, 646-678.

Klayman, J. & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Legrenzi, P., Girotto V., & Johnson-Laird, P.N. (2003). Models of consistency. *Psychological Science, 14*, 131-137.

Minsky, M.L. (1975). Frame-system theory. In Schank, R.C., and Nash-Webber, B.L. (Eds.) *Theoretical Issues in Natural Language Processing*. Preprints of a Conference at MIT, Cambridge, MA.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson and Co.

Revlin, R., Cate, C.L., & Rouss, T.S. (2001). Reasoning counterfactually: combining and rending. *Memory and Cognition, 29*, 1196-1208.

Shaklee, H. & Fischhoff, B. (1982). Strategies in information search in causal analysis. *Memory & Cognition, 10*, 520-530.

Simon, D., & Holyoak, K.J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and Social Psychology Review, 6*, 283-294.

Sloman, S.A. & Lagnado, D.A. (2004). Do We "do"? *Manuscript in submission.*

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-502.

Walsh, C.R. & Byrne, R.M.J. (2004). Belief revision, the inference contradiction effect and counterfactual conditionals. *Manuscript in preparation.*

Walsh, C.R. & Johnson-Laird, P.N. (2004). Changing your mind. *Manuscript in submission.*

Toward A Multilevel Analysis of Human Attentional Networks

Hongbin Wang (Hongbin.Wang@uth.tmc.edu)*
Jin Fan (Jin.Fan@mssm.edu)**
Yingrui Yang (yangyri@rpi.edu)***
*School of Health Information Sciences, University of Texas Health Science Center at Houston
**Department of Psychiatry, Mount Sinai School of Medicine
***Department of Cognitive Science, Rensselaer Polytechnic Institute

Abstract

Attention is a complex multilevel system subserved by at least three interacting attentional networks in the brain. This paper describes a multilevel computational model of attentional networks, developed in both the symbolic architecture of ACT-R and the connectionist framework of leabra. We evaluated the model using the Attentional Networks Test and the simulation results fitted the empirical data well. We argue that developing multilevel computational models helps to link findings at different levels.

Introduction

Suppose a student S was asked to solve the equation "2x + 3 = 9" (Figure 1A), and he used 2 seconds to produce the answer "x = 3". Both cognitive scientists X and Y were interested in understanding how S did it. Scientist X recorded S's detailed verbal protocol (Figure 1B), based on which, and other relevant behavioral measures, X hypothesized the possible knowledge structures underlying S's problem solving and developed a symbolic computational model that simulated the process (Figure 1C). On the other hand, scientist Y adopted sophisticated brain imaging techniques such as electroencephalograph (EEG) and functional Magnetic Resonance Imaging (fMRI) and acquired a high-resolution recording of S's brain dynamics during problem solving (Figure 1D). Based on some well-established neural computing principles, Y then developed a biologically realistic connectionist model to simulate the brain activities underlying S's performance (Figure 1E). Though both models fitted the data well, the two models are clearly different. While the symbolic model offers a description of the process with psychological plausibility and high behavioral relevance, the connectionist model emphasizes the process' biological realism and brain foundations. One question is, do we, cognitive scientists who endeavor to discover unified theories of cognition, have justifiable reasons to prefer one to another?

This question and similar others have led to a long debate in the rather brief history of cognitive science (e.g., Churchland & Sejnowski, 1992; Newell, 1990; Rumelhart & McClelland, 1986). Recently a BBS (Behavioral and Brain Sciences) target article was dedicated to this issue (Anderson & Lebiere, 2003). The authors adopted a set of 12 criteria, which they called "The Newell Test", to systematically compared and contrasted ACT-R, a rule-based cognitive architecture (Anderson & Lebiere, 1998), and the connectionist modeling framework. Their conclusion was that both frameworks had great strengths as well as serious limitations as candidates of the unified theory of cognition.

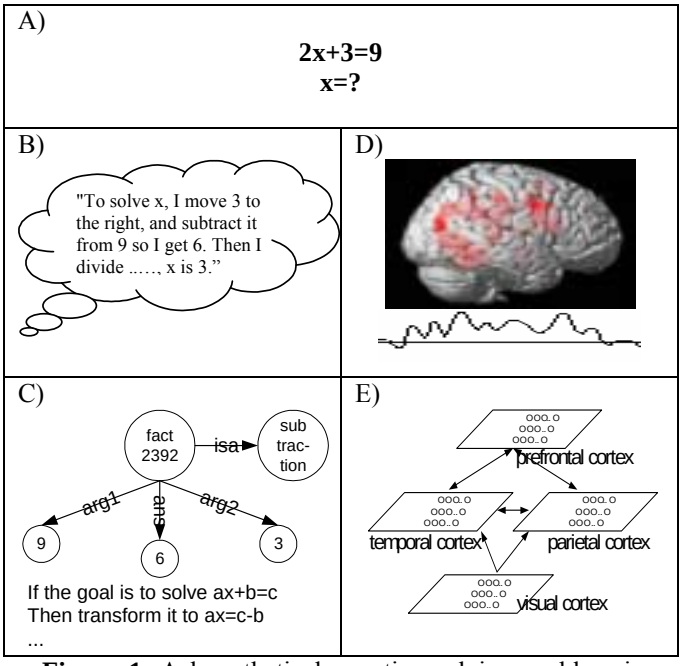


Figure 1. A hypothetical equation-solving problem is presented in A. Verbal protocol and brain imaging data are presented in B and D. Sketches of a symbolic model and a connectionist model of task are presented in C and E.

This is hardly surprising given the inherent complexity of the human mind itself. It has long been recognized that the mind is a multilevel construct and can be analyzed at different levels. Marr, for example, distinguished and separated among computational theory, representation and algorithm, and hardware implementation (Marr, 1982). Similar distinctions were made by Newell among different bands of cognitive functions (Newell, 1990). Newell argued that different bands utilize different basic operators, which have different time scales. More importantly, different bands form a hierarchy. Multiple lower lever basic operators can be combined to form higher level basic operators. In other words, lower level operators can be summarized up at higher level though this summarization may not be linear.

Single level analyses have been the dominant methodology in cognitive science. Experimental psychology and symbolic modeling, for example, largely depend on controlled experiments and behavioral observation. Recent advances in cognitive neuroscience allow us to directly observe, with high temporal-spatial resolutions, how an active brain functions during cognitive performance (Posner & Raichle, 1994). As a result, biologically realistic neural networks modeling has flourished (O'Reilly & Munakata, 2000). Efforts have also been made to probe the function of mind at lower molecular levels (e.g., Bellugi & George, 2001; Squire & Kandel, 2000). While all these levels of analyses tell us important aspects of the mind, neither of them alone is adequate to describe the whole picture. The human mind is a complex entity and may leave shadows at different levels when it works (Penrose, 1996). However, in order to achieve a unified theory all of the pieces have to be somehow linked together.

One approach would be to develop so called "hybrid systems", which typically combine symbolic and subsymbolic components together (e.g., Sun & Alexandre, 1997). We, for example, have developed a hybrid model of human abductive reasoning by combining a Soar component (a symbolic architecture) for hypothesis generation and a connectionist component for hypothesis evaluation (Johnson, Zhang, & Wang, 1997). Although hybrid systems take advantage of both types of components and can become quite powerful, they often bear little true psychological and neurophysiological significance due to the fact they are artificially assembled systems. While it is well agreed that human cognition involves mechanisms and operations at, among others, both psychological and neuronal networks levels, simply piecing them together is *ad hoc* and trivializes the problem (see also Wang, Johnson, & Zhang, 2003)

In this paper we argue that we need a multilevel modeling approach. That is, we need to develop well-fitted computational models at multiple levels for any given cognitive phenomenon. Because the mind manifests itself at multiple levels, each level is real and tells a unique story of the mind on its own. When we develop models for a specific phenomenon at multiple levels, we would be able to compare them, contrast them, and more importantly, mutually justify them. By doing so, we expect that a more complete picture of the mind might emerge.

This paper is organized as follows. We first briefly review findings on human attentional networks and introduce the Attentional Network Test (ANT) (Fan, MaCandliss, Sommer, Raz, & Posner, 2002). We then demonstrate the multilevel modeling approach by reporting and cross-validating two computational models for the same ANT task, one developed in ACT-R, and the other in leabra, a biologically realistic connectionist modeling framework (O'Reilly & Munakata, 2000). While both models fitted data well they emphasized different levels of explanations. Finally the implications of this practice are discussed.

Human Attentional Networks

Although "everyone knows what attention is" (James, 1890), how attention works remains one of the most challenging questions in science (Parasuraman, 2000; Pashler, 1998). Recent advances in cognitive psychology and cognitive neuroscience have suggested that there exist multiple attentional networks in the brain, each of which subserves different types of attention (Fan et al., 2002; Posner & Dehaene, 2000; Posner & Petersen, 1990). At least three attentional networks, for alerting, orienting, and executive control, have been distinguished at both cognitive and neuroanatomical levels (see Figure 2A). Specifically, alerting involves a change in the internal state to become ready for any incoming task-related events. Neuroimaging evidence has revealed that the alerting network consists of some frontal and parietal areas particularly of the right hemisphere. Orienting, closely related to the conventional selective visuo-spatial attention, involves selectively focusing on one or a few items out of many candidate inputs. Evidence has shown that the orienting network includes parts of the superior and inferior parietal lobe, frontal eye fields and such subcortical areas as the superior colliculus of the midbrain and the pulvinar and reticular nucleus of the thalamus. Finally, executive control of attention is related to monitoring and resolving conflicts. Executive control is often needed in higher level mental operations including planning, decision making, error detection, novel or not well-learned responses, and overcoming habitual actions. Converging evidence from neuroimaging and neuropathology studies has suggested that the executive control network consists of the midline frontal areas (anterior cingulate cortex), lateral prefrontal cortex, and the basal ganglia.

The ANT paradigm was recently developed to simultaneously measure the performance of the three attentional networks and evaluate their interrelationships (Fan et al., 2002). It is essentially a combination of a spatial cueing task (Posner, 1980) and a flanker task (Eriksen & Eriksen, 1974), as illustrated in Figure 2B. The stimulus consists of a row of 5 horizontal arrows and the participants' task is to report the pointing direction (left or right) of the center arrow (the target) by pressing a key. The four arrows surrounding the target, with two on each side, are called the flankers. These flanker arrows point either in the same direction as that of the target (the congruent condition), or in the opposite direction (the incongruent condition). An additional condition (the neutral condition) is also included in which the flankers are four straight lines with no arrowheads. To introduce an orienting component, the row can be presented at two locations, either above a fixation point (top) or below it (bottom). To introduce an alerting component, the row may be preceded by a cue (the cued condition) or may not (the no-cue condition). In addition, when there is a cue, this cue may be presented at the center fixation location (the center-cue condition), at the top or bottom location where the stimulus row is to appear (the

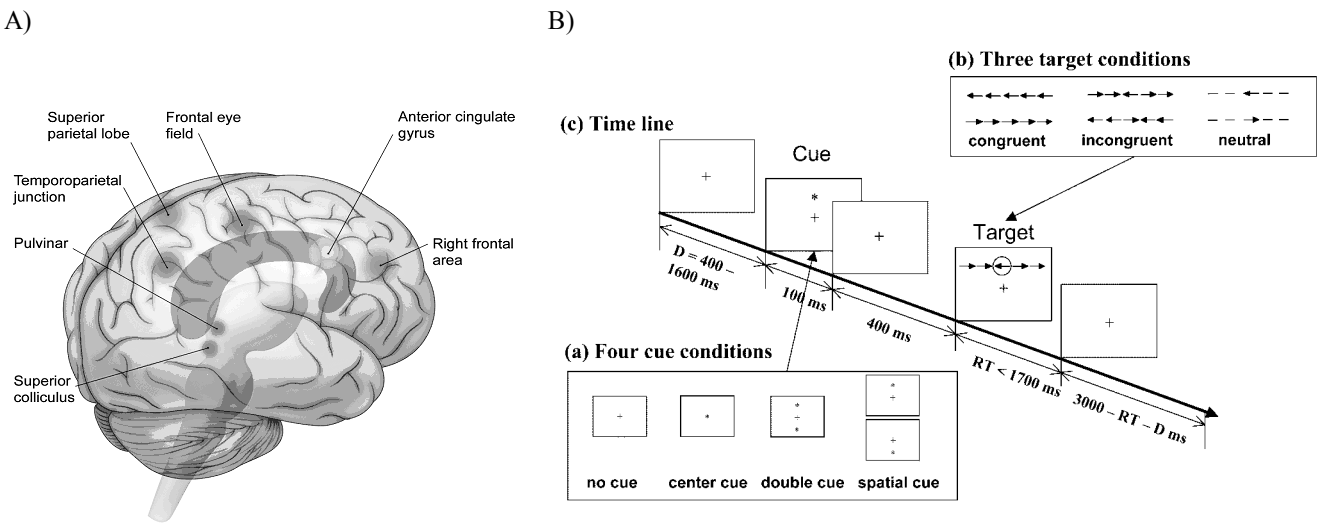


Figure 2. Human attentional networks (A) and the ANT task (B)

spatial-cue condition), or at both top and bottom locations (the double-cue condition). Note that while a spatial-cue precisely predicts where the stimulus is to appear, in both the center-cue condition and the double-cue condition the participant cannot infer that information from the cue.

Fan et al. (2002) tested 40 normal adult participants using the ANT paradigm. Their reaction time (RT) results are shown in Figure 3A. They then proposed the following formula as a measure of the efficiency of each of the three attentional networks:

· Alerting efficiency = RT(no-cue) – RT(double-cue),
· Orienting efficiency = RT(center-cue) – RT(spatial-cue),
· Conflict efficiency = RT(incongruent) – RT(congruent),

which resulted in the efficiency measures of 47 ± 18 ms, 51 ± 21 ms, 84 ± 25 ms, for alerting, orienting, and executive control, respectively.

Fan et al. (2001) also reported an fMRI study using the ANT paradigm. Their results were consistent with the general findings shown in Figure 2A.

Multilevel Computational Modeling of Human Attentional Networks

While both the behavioral and neuroimaging studies using the ANT paradigm revealed important psychological and neurophysiological characteristics of human attentional networks, there exists a gap between these two levels of analyses. In particular, how do these different attentional neural networks work together to generate psychologically meaningful behavior? It has been well agreed that the link between neural activities and psychological performance is nontrivial and must be taken into account seriously to avoid "neo-phrenology". Developing well-principled and constrained computational models help in the regard (Cohen & Tong, 2001).

Traditional computational modeling approaches to human attention have typically adopted various connectionist modeling techniques (e.g., Cohen, Dunbar, & McClelland, 1990). While it has been fruitful, this practice fails to account for the manifestations of attention at symbolic/cognitive levels. As we illustrated earlier, attention, as an essential aspect of human cognition, is a complex multilevel construct. In order to understand the computational mechanisms of attention at different levels and the links among them, we need multilevel models.

We have developed a multilevel model for the ANT task. One sub-model was developed in the symbolic modeling framework of ACT-R and focused on the psychological aspects of the task. The other was developed in the connectionist modeling framework of leabra and emphasized the neurophysiological aspects of the task. A preliminary cross-validation of two models is discussed.

ANT on ACT-R

ACT-R is a production rule based cognitive modeling architecture developed by John Anderson and colleagues over a period of nearly two decades (see Anderson & Lebiere, 1998). In essence, ACT-R explains human cognition by proposing a model of the knowledge structures and knowledge deployment that underlie cognition. Although ACT-R consists of a nontrivial subsymbolic component for computations involving activation and association, it is fundamentally a symbolic modeling framework in that it relies extensively on various symbolic structures for knowledge representation. For example, ACT-R makes a fundamental distinction between declarative and procedural knowledge. Declarative knowledge corresponds to things people are aware of and can usually describe to others and is represented in ACT-R by chunks. Procedural knowledge is knowledge that people display in behavior but are not conscious of and is represented by production rules (condition-actions pairs). Both chunks and production rules are fundamental symbolic structures in ACT-R and are regarded as the atomic components of thought in the sense

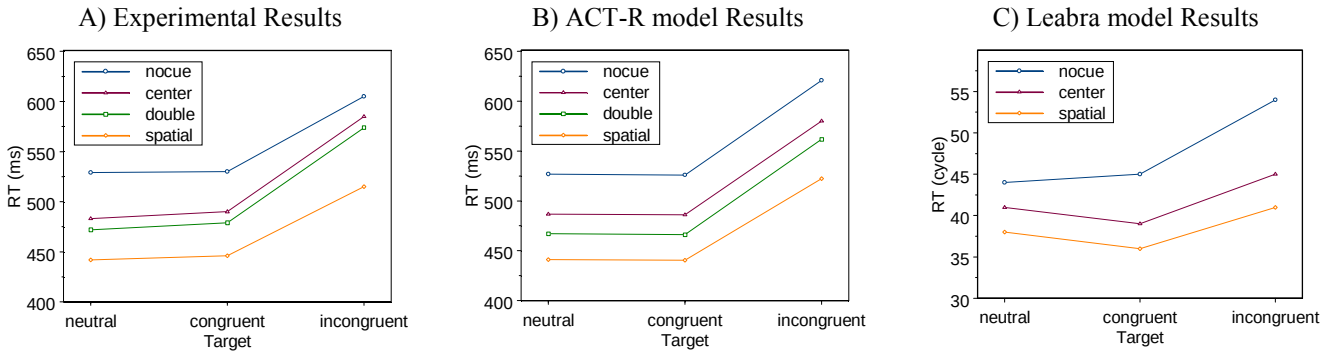


Figure 3. Experimental (A, based on Fan et al. (2002)) and modeling results (B and C).

that they are as far down as one can go in the symbolic decomposition of cognition. In ACT-R, on average every fifty (50) milliseconds, one production rule is chosen to fire, a few declarative chunks are processed, and cognition advances one step. Therefore, it is claimed that ACT-R captures the symbolic grain size of cognition.

We developed a computational model for the ANT task in the framework of ACT-R (Wang, Fan, & Johnson, 2004). Our purpose is two-fold. First, we want to explore how different types of attention work together in a single framework to produce the cognitive performance. Second, such a model offers a solid testbed for us to cross-validate those models based on various connectionist modeling results and neuroimaging data.

We started by analyzing the major functional components in the ANT task. We distinguished six major stages in a typical ANT trial: fixation and cue expectation; cue or stimulus judgment; cue processing; stimulus expectation; stimulus processing; and response. We then mapped these functional components onto 36 ACT-R production rules. With these rules our model could perform the ANT task and interact with the same experimental environment that human participants interact.

We evaluated the performance of the model by using the model as a "simulated subject" to perform the ANT experiment. The RT results of 100 "simulated subjects" are presented in Figure 3B. A correlation analysis shows very high correlations (0.99 for RTs and 0.97 for error rates) between the simulation and experimental results. We then followed the same procedure discussed early to estimate the effects of the three attentional networks based on the simulated RT data, resulting in the efficiency measures of 55 ± 7.4 ms, 45 ± 7.0 ms, 86 ± 7.4 ms, for alerting, orienting, and executive control, respectively. A close match between the two sets of data is apparent, with a notable exception that the simulated standard deviations are consistently smaller than the empirical ones. The reason is that we did not add any between-subject variance in our model. As a result, these simulated variances actually reflected those within-subject variations in performing the ANT task. Overall these results suggest that the model captured well the various attentional effects that the ANT task was designed to measure.

The concept of production rule is fundamental to our model of attention. One of the key features of the model is that it mapped the effects of attentional networks to production rules. Rules fire in sequence and operate at a rate of about 40-50 ms per production rule. As argued by ACT-R, production rules define the atomic components of thought at the symbolic level. When we examined the efficiency measures of attentional networks reported in Fan et al (2002) it seemed that they (51 ms, 47 ms, and 84 ms, for alerting, orienting, and executive control, respectively) fell well into the range of a few rule firings time period. Perhaps all we need is about one (for alerting and orienting) or two (for executive control) additional production rules to explain symbolically the work of attentional networks. This is indeed what our model demonstrated.

ANT on Leabra

Leabra (*l*ocal, *e*rror-driven and *a*ssociative, *b*iologically *r*ealistic *a*lgorithm) is a connectionist modeling framework proposed recently by O'Reilly and Munakata (2000). There are at least three features that distinguish it from other connectionist modeling frameworks. First, it has sound neurological foundations. It is biologically realistic in multiple aspects. Its neurons compute based on membrane potentials and ion channels. Its neuronal connections are often bi-directional and cannot change signs (i.e., changing from an excitatory link to an inhibitory link, and vice versa). It uses biologically inspired learning rules such as Hebbian learning for unsupervised learning and the generalized recirculation algorithm (but not the biologically unrealistic backpropagation) for error-driven learning. Second, leabra is a coherently integrated framework. Many distinctions in traditional neural network modeling, including supervised vs unsurprised learning, feedforward vs recurrent networks, and pattern recognition vs self-organization maps, are all unified in a single coherent framework, based on well-supported biological principles. Third, partly due to its biological realism, it is now possible, for example, to designate a specific neural network to simulate a specific area of brain, and flexibly connect the multiple such networks, each of which can have its own properties such as the average activation level and the connection density, to simulate various brain pathways. As a result, it offers great

flexibility to build a hierarchy of neural networks and link network activities to higher-level symbols.

A connection model of the ANT task was developed in the framework of leabra. The structure of the model is shown in Figure 4. This model contains modules for all the three attentional networks. In addition, it contains modules for perception (visual input and primary visual cortex), object recognition (object pathway), and response (output). The networks are connected in such a way that they conform to the known functional an anatomical constraints as much as possible (Farah, 2000; O'Reilly & Munakata, 2000).

The model works as follows. When a cue comes on, the primary visual cortex module is activated, which in turn triggers the alerting network. This cue-induced alerting affects later stimulus processing because the alerting network will remain excited for a while which will activate the orienting network in general causing it to become ready for the incoming stimulus. In addition, when the cue is a spatial one (i.e., a cue that indicates where the target stimulus is to appear), it will further make the corresponding sub-region of the orienting network even more excited. This occurs because the orienting network adopts a retinotopy-based spatial representation of the environment. This extra excitation in the sub-region of the orienting network will facilitate the corresponding stimulus processing in the object pathway network, due to the connections between them. This accounts for the orienting effect. Finally, note that it is the object pathway network that is responsible for the arrow direction detection. When the incongruent stimulus (e.g., a left arrow flanked by four right arrows) is presented, the object pathway network may propose different responses, which compete for the final expression in the output network. The executive control network then activates making the center arrow defeating the flankers. This is where the executive control attention plays a role.

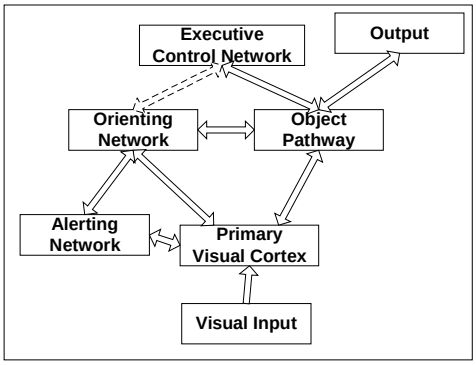


Figure 4. A leabra model of ANT.

The performance of the model was evaluated by using it to perform the ANT task. Stimuli are presented to the model in a similar way as to a human. Depending on the conditions, a cue, which can be either a center cue or a spatial cue, may be presented for a fixed time period before the stimulus presentation (note that the double cue condition was not simulated here since the current version of model were not equipped with enough neurons). The number of cycles the output module takes to produce a stable response after the stimulus presentation serves as a measure of the reaction time. The simulation results are shown in Figure 3C. A regression analysis showed that

RT(ms) = 12 * RT(cycle)

with a R-square of 0.99. It is clear that the model fits the behavioral data reasonably well.

Discussion

Human attention is a multi-component multilevel construct. Both behavioral and neuroimging studies using the ANT paradigm revealed important aspects of the function of human attentional networks. Multilevel computational modeling helps to probe how these multiple components work together and manifest themselves at multiple levels.

The multilevel model we reported in this paper consisted of a sub-model developed in the framework of ACT-R and the other in the framework of leabra. While the former sub-model focused on the symbolic knowledge structure of cognitive performance and psychological plausibility, the latter focused on the subsymbolic neural information processing and biological realism. However, since both models simulated the same ANT task and fitted the empirical data well, the combined multilevel model offered a real possibility to cross-validate the models and probe the computational link among different levels.

First of all, the model illustrated interesting relationships between production rules and underlying neural computation. As demonstrated in the ACT-R model, rules are fundamental units of psychological reality and typically proceed serially. However, the underlying neural networks process information in parallel. The parallelism of neural computation and the serial nature of rule firing can be mapped against each other along the time line. Since both types of models decompose the cognitive performance into sub-units that occur at tens of millisecond scales, the mapping may be able to tell how rules are implemented in neural level computation. Based on the models, for example, we can map one ACT-R rule (40 ms in the current model) to about three leabra cycles (about 12 ms per cycle). Though such a simple and linear mapping should not be taken literally, it does provide a vivid footnote about how parallel neural computing is summarized psychologically by serial rule firings. It illustrates that we may not be able to find a "rule center" in the brain. Instead, rules can be implemented anywhere in the brain – they are simply pattern matching. For example, there is a symbolic rule that summarizes the conflict monitoring and detection operation typically subserved by the anterior cingulate cortex. The general neural priming underlying alerting in the alerting networks is summarized by another task switch rule.

Our model also demonstrates how functionally identical operations can be implemented by different mechanisms at different levels. One interesting finding from Fan et al. (2002) is the small but reliable difference in RT (about 11 ms) between the center-cue and the double-cue conditions.

A convenient explanation is that in the double-cue condition due to *diffused attention* both stimulus locations had been primed a little, which saved a little time when the stimulus appeared later. While it is easy to model priming and diffused attention in a connectionist model (e.g., our leabra model), how it is implemented at a symbolic rule level raises a challenge. Our ACT-R model adopted a mechanism in which several symbolic and psychologically meaningful move-attention operations were carried out sequentially. The simulated RT difference was 19 ± 8 ms.

The multilevel model for human attentional networks we reported in this paper has allowed us to compare/contrast the computational mechanisms at different levels and to probe the important computational links between psychologically meaningful mental operations and neural activities. It also enjoys potentially significant prediction power in that the model at one level can lead to nontrivial predictions about the operations at another level. However, we recognize that for this approach to work models at each level have to be independently and/or mutually validated. Further analyses and more detailed alignments of our current model remain to be done.

Acknowledgments

This work is partially supported by a grant from the Office of Naval Research (Grant No. N00014-01-1-0074). We thank Drs. Todd R. Johnson and Jiajie Zhang for their help in this work.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Press.

Anderson, J. R., & Lebiere, C. (2003). The Newell Test for a theory of Mind. *Behavioral and Brain Science, 26*, 587-639.

Bellugi, U., & George, M. S. (Eds.). (2001). *Journey from cognition to brain to gene: Perspectives from Williams Syndrome*. Cambridge, MA: MIT Press.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological Review, 97*(3), 332-361.

Cohen, J. D., & Tong, F. (2001). The face of controversy. *Science, 293*, 2405-2407.

Fan, J., MaCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14*(3), 340-347.

Fan, J., McCandliss, B. D., Flombaum, J. I., & Posner, M. I. (2001). *Imaging attentional networks.* Paper presented at the Society for Neuroscience 2001 Annual Meeting, San Diego, CA.

Farah, M. J. (2000). *The cognitive neuroscience of vision*. Malden, MA: Blackwell Publishers.

James, W. (1890). *Principles of psychology*. New York: Holt.

Johnson, T. R., Zhang, J., & Wang, H. (1997). A hybrid learning model of abductive reasoning. In R. Sun & F. Alexandre (Eds.), *Connectionist Symbolic Integration*. Hillsdale, NJ: Lawrence Erlbaum.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

Parasuraman, R. (Ed.). (2000). *The attentive brain*. Cambridge, MA: MIT Press.

Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Penrose, R. (1996). *Shadows of the mind: A search for the missing science of consciousness*. New York: Oxford University Press.

Posner, M. I., & Dehaene, S. (2000). Attentional networks. In M. S. Gazzaniga (Ed.), *Cognitive neuroscience: A reader*. Malden, MA: Blackwell Publishers.

Posner, M. I., & Petersen, S. E. (1990). The attention systems of the human brain. *Annual Review of Neuroscience, 13*, 25-42.

Posner, M. I., & Raichle, M. E. (1994). *Images of mind*. New York: Scientific American Library.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol 1: Foundations*. Cambridge, MA: MIT Press.

Squire, L. R., & Kandel, E. R. (2000). *Memory: From mind to molecules*. New York: Scientific American Library.

Sun, R., & Alexandre, F. (Eds.). (1997). *Connectionist-symbolic interaction: From unified to hybrid approaches*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, H., Fan, J., & Johnson, T. R. (2004). A Symbolic Model of Human Attentional Networks. *Cognitive Systems Research, 5,* 119-134.

Wang, H., Johnson, T. R., & Zhang, J. (2003). [Commentary] A multilevel approach to modeling human cognition. *Behaviroal and Brain Science, 26,* 626-627.

Alignment of Reference Frames in Dialogue

Matthew E. Watson (s0129081@sms.ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ, UK

Martin J. Pickering (Martin.Pickering@ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ, UK

Holly P. Branigan (holly.branigan@ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ, UK

Abstract

Previous research has shown that interlocutors in a dialogue align their utterances at several levels of representation. This paper reports two experiments that use a confederate-priming paradigm to examine whether interlocutors also align their spatial representations during dialogue. Experiment 1 showed a significant reference frame priming effect: Speakers tended to use the same reference frame to locate an object in a scene as the frame that they had just heard their interlocutor use. Experiment 2 demonstrated the same pattern even when the speaker's description and their partner's previous description involved different prepositions. Hence the effect cannot be explained in terms of lexical priming of a particular preposition. Our results are strong evidence that interlocutors in a dialogue align non-linguistic as well as linguistic representations.

Research on dialogue has suggested that the traditional methods employed in psycholinguistics may not give a true, or at least complete, account of human language. The traditional approach focuses largely on monologue and involves investigating single word utterances in isolated controlled circumstances, e.g. the picture naming paradigm, or the lexical decision task. However, Clark (1996) pointed out that the natural setting for language is dialogue, and that language does not normally occur in these isolated circumstances, thus questioning the ecological validity of traditional methods. The realization of this has led to a research program into how language is used in dialogue (e.g., Clark & Wilkes-Gibbs, 1986; Horton & Keysar, 1996; Garrod & Anderson, 1987). Research in this framework has shown that interlocutors in a dialogue tend to align their utterances: Over the course of a conversation participants will come to communicate in a similar fashion to each other. This occurs at several levels of communication, including the conceptual (Garrod & Anderson, 1987), lexical (Clarke & Wilkes-Gibbs, 1986) and syntactic (Branigan, Pickering, & Cleland, 2000) levels. In these experiments, participants usually achieved alignment without resorting to overt negotiation. In the case of syntactic alignment at least, many subjects were not aware that they were aligning.

Pickering and Garrod (in press) proposed a mechanism for how alignment is achieved between interlocutors. According to this theory, alignment is the basis for successful dialogue; misunderstanding occurs when alignment is not achieved. Alignment occurs when the two interlocutors employ equivalent representations at different levels, and arises from an automatic priming mechanism. This allows alignment to be achieved quickly and efficiently without reliance upon time-consuming strategies of open negotiation. Indeed, such strategies are only employed when the primitive mechanisms fail. To prevent unnecessary negotiation Pickering and Garrod suggest a second primitive mechanism that allows repair of representations when misalignment occurs; see Garrod and Pickering (2004) for a summary.

Dialogue research has shown alignment of linguistic representations, but alignment is hypothesized also to occur for conceptual representations, such as those associated with object location. A speaker's conceptual representation of where objects are located is reliant upon an overall spatial representation, which underpins the use of spatial language. In order to describe object locations effectively it is important that both interlocutors take the same perspective (Levelt, 1989) concerning the objects they are locating. For example, an addressee must understand whose left a speaker is talking about. In the same way that interlocutors align on which lexical terms should be used to describe a scene, it would be advantageous for interlocutors to align on which perspective a scene should be described from.

The perspective that is taken depends upon the reference frame that is applied to a spatial representation of a scene. A reference frame is an axial co-ordinate system that defines regions extending from the origin, whose axes are labelled with directional terms. The object to be located (figure object) can then be located in relation to another object (reference object) based upon the directional axes of the reference frame. However, there are three different types of reference frame (at least in English; other languages use

only two or even one; Levinson 2003) that a speaker can employ in order to locate an object: absolute, relative, and intrinsic. It is important that the addressee knows which of these the speaker is using in order to successfully understand an utterance.

The absolute reference frame locates an object based upon salient, stable features of the environment, for example, the cardinal directions. The dot in Figure 1 can be described as *west of the chair* if the page is held horizontally with the top of the page facing north.

The intrinsic reference frame locates an object based upon the directional features of the reference object. The dot in Figure 1 can be described as *above the chair* because it is in alignment with the top of the chair.

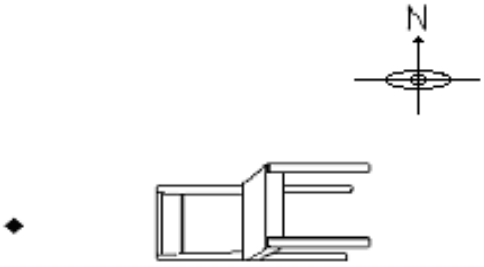


Figure 1. The dot can be described as west using an absolute reference frame, above using an intrinsic reference frame or left using a relative reference frame.

The relative reference frame locates an object in relation to the viewpoint of an observer. The axes of the reference frame are labelled based upon the features of the person upon whose viewpoint the location is based. In Figure 1 the dot would be described as *left of the chair* using a relative reference frame. (In many cases the relative reference frame is used from the viewpoint of a speaker or an addressee, but it can also be from a third person perspective.)

The above tripartite classification of reference frames follows that proposed by Levinson (1996, 2003), and is distinct from the classification traditionally employed in the psycholinguistic literature, which identified absolute, deictic, and intrinsic reference frames, all defined on the basis of their origin. (Deictic reference frames are all reference frames with an egocentric origin.) Levinson pointed out that this traditional system is not an appropriate way to categorize reference systems because it is possible to have a non-deictic relative reference frame, such as *The ball is to the right of the tree as you look at it*, and a deictic intrinsic reference frame such as *The ball is in front of me*.

When describing an object's location, an individual has to select one of these reference frames to use in preference to either of the other two reference frames. Carlson-Radvansky and Jiang (1998) showed that reference-frame selection is achieved via inhibition of non-selected reference frames. When participants used a relative reference frame to identify an object's location, they were slower to describe an object's location using an intrinsic reference frame immediately afterwards. Inhibition operates not only on the endpoint of an axis, but on at least the entire axis, e.g. if *left* (intrinsic) is inhibited then using *right* or *left* (intrinsic) in the subsequent description will take longer than using a relative reference frame.

The findings of Carlson-Radvansky and Jiang (1998) suggest that reference frames are influenced by low-level priming. However, the results do not establish whether or not this occurs during dialogue: Reaction time was used as a measure of cognitive effort in trials whereas in dialogue any effect of priming must manifest itself by a change in the person's linguistic behaviour. Furthermore, Carlson-Radvansky and Jiang's (1998) experiment only investigated inhibition of the endpoint of an axis and the inhibition of the axis itself. If interlocutors align reference frames, we would expect them to align the entire reference frame rather than just part of it. Therefore it is unclear whether this kind of priming is enough to cause the alignment of reference frames between interlocutors in the manner described by Pickering and Garrod (in press).

In two series of experiments Schober (1993, 1995) showed that the reference frame which an individual selects is affected by their partner in a conversation. Individuals who described the location of an object to a partner who viewed the scene from a different perspective were more likely to describe the location from their partner's perspective. When the partner queried such descriptions, they used their own perspective to describe object location. Schober concluded that interlocutors use conscious strategies to collaborate in ways of describing object location.

Schober's results suggest that interlocutors may align reference frames. However, it is not clear that this is necessarily the case. In his experiments, two participants interacted freely, allowing little control over what was said by each pair. This means that pairs of participants may be reverting to default reference frames. Furthermore, in a large proportion of trials participants located objects using terms that required no reference frames (e.g. *next to*, *between* and so on).

The present work is an experimental investigation to discover whether or not interlocutors align reference frames. The investigation uses a confederate-priming paradigm (e.g. Branigan et al., 2000) where a naïve participant and a participant who is - unknown to the naïve participant - a confederate of the experimenter and who is following a script, communicate during the experiment. If interlocutors do align reference frames then they will use a reference frame significantly more when they have just heard an utterance using that reference frame than when they have just heard an utterance using an alternative reference frame. Alternatively interlocutors may select a reference frame based solely upon the perceptual properties of the spatial

array, in which case they should be unaffected by the reference frame just used by their partner. Our experiments also set out to separate priming for reference frames from lexical priming. If priming of reference frames exists separately from lexical priming, we can expect subjects to use a reference frame significantly more if they have just heard an utterance using that reference frame even if the same spatial term is not applicable to both utterances.

Experiment 1

Method

Participants 12 students of the University of Edinburgh were paid volunteers in the experiment, which lasted 20 minutes. All were native English speakers.

Materials The experiment was run on two computers positioned back to back, using E-prime software. One program was created for the confederate and consisted of sentences positioned in the centre of the screen of the form "The dot above the chair". This formed the script for the experiment. The second program was for the participant and displayed pictures for the match and describe phases of the experiment.

12 monochrome objects were used as reference objects, all fitting into a rectangle 93 pixels high and 121 pixels wide. Two versions of each object were used, one rotated 90° clockwise and one rotated 90°anti-clockwise.

The figure object was an 11x11 pixel square rotated so that its vertices were the top, bottom, leftmost, and rightmost points. The figure object was located above, below, left, or right (in a relative reference frame) of the reference object. The centre of the figure object was positioned between 125 and 130 pixels from the centre of the reference object.

Design There were 3 within-participants and within-items factors: Prime Reference Frame (Relative vs. Intrinsic); Preposition (Same Preposition vs. Different Preposition); and Target Plane (Vertical vs. Horizontal). These are exemplified in Figure 2. The prime scene in Figure 2 can either be described as *The dot above the camera* (relative reference frame) or *The dot right of the camera* (intrinsic reference frame). In the top diagram of Figure 2, alignment requires using the same preposition (either *above* or *right of*); in the bottom diagram, alignment require using a different preposition (either *left* or *below*). Finally, the top target scene is aligned vertically whilst the bottom target scene is aligned horizontally.

Two lists of 96 trials were constructed, with each trial consisting of a match phase and a describe phase. The reference objects in each list were rotated clockwise and anti-clockwise on half of the trials each. Reference frame was counterbalanced across list and rotation. Preposition overlap was counterbalanced across rotation in each list. Participants saw 12 trials in each of the 8 conditions formed by crossing the three factors. The trials were presented in a fully randomized order, which was different for each participant.

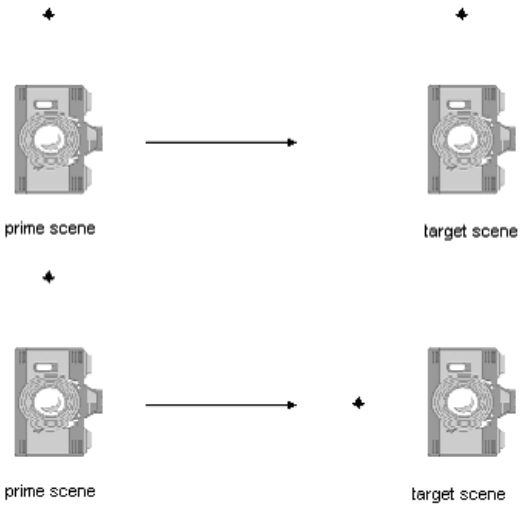


Figure 2: The top diagram shows the 'same preposition' condition. The bottom diagram shows the different preposition condition.

Procedure The two participants were introduced to each other (throughout the experiment, the experimenter treated the confederate as if she was a naïve participant).

The participant and confederate each sat at a computer each. The computers were situated back to back so that neither could see each other, or the other's screen. After hearing instructions, participants pressed the space bar to begin a practice session of 8 trials, one trial corresponding to each of the 8 conditions. Instructions then appeared on the screen signalling the end of the practice session and the start of the experiment. Each trial proceeded as follows: After participants pressed <space> to begin, the match screen appeared. The match screen contained two examples of a reference object (both the same, with one on the left and one on the right) and a dot located above, below, left or right of each one. The confederate gave a description of the location of the dot in relation to the object. The participant then chose which of the two examples on the screen matched the confederate's description of the dot location accurately, pressing the M key if it was the right-hand example and the Z key if it was the left-hand example. Participants were told that if they were not sure which picture matched their partner's description to pick the one they thought matched most closely.

After selection the match scenes disappeared (no feedback was given) and a fixation cross appeared in the centre of the screen. This remained on screen for 1000ms. The fixation cross was then replaced by a reference object in the centre of the screen with a dot above, below, left, or right of it. Participants then described the location of the dot in relation to the object. After describing this they pressed space and the scene disappeared. It was replaced by a fixation cross in

the centre of the screen for 500ms. This then disappeared and the next trial began with a match task.

Results

For the analysis participants' first responses were used. The percentage of intrinsic responses were then analyzed using two 2x2x2 repeated measures ANOVAs (by participants (F1) and by items (F2)), with Prime reference frame (intrinsic or relative), preposition (same or different), and Target plane (horizontal or vertical) as factors.

Table 1 shows the mean number of intrinsic responses used by subjects in each of the 8 conditions. There was a significant main effect of Prime reference frame (62.9% vs. 53.3%; $F1(1,11) = 26.86$, $p < .01$; $F2(1,11) = 9.35$, $p < .05$). That is, participants were significantly more likely to use an intrinsic reference frame after the confederate had used an intrinsic reference frame, compared to when the confederate had used a relative reference frame.

Table 1:
Mean percentage of intrinsic responses in Experiment 1.

	Relative Prime		Intrinsic Prime	
	Same	Diff	Same	Diff
Vertical	32.3	52.8	54.9	52.3
Horizontal	63.6	64.4	73.3	70.9

When the figure and reference objects were aligned vertically, participants used an intrinsic reference frame 48% of the time compared to 68% when the alignment was horizontal. This difference was significant ($F1(1,11) = 8.07$; $p < .05$; $F2(1,11) = 101.17$; $p < .01$), showing that participants were significantly more likely to use an intrinsic reference frame when the objects were aligned horizontally than when they were aligned vertically.

As expected there was no effect of preposition ($p > .05$): Participants used an intrinsic reference frame as much when the prepositions were the same as when they were different. This is regardless of which reference frame the confederate used.

There was a significant two-way interaction between Prime Reference Frame and Preposition ($F1(1,11) = 13.07$; $p < .01$; $F2(1,11) = 6.19$; $p < .05$). All other two-way interactions were non-significant ($p > .05$). Post-hoc analyses showed that these interactions occurred because of a difference between two of the eight conditions, relative, same, vertical and relative, different, vertical: The former yielded 32.3% intrinsic responses whereas the latter yielded 52.8% intrinsic responses ($t(23) = -2.91$; $p = .01$).This means that participants were more likely to use a relative reference frame when the reference and figure object were aligned vertically (i.e. they would use *above* and *below* to describe the dots' location) following the confederate using a relative reference frame when there was preposition overlap (i.e. the confederate used above or below) than when there was no preposition overlap (i.e. the confederate used left or right).

Discussion

The results of Experiment 1 show an effect of alignment of reference frames. Participants were more likely to use an intrinsic reference frame after the confederate had used an intrinsic reference frame.

The significant effect of Target Plane indicates that participants preferred to use the lexical terms *above* or *below* to *left* or *right,* regardless of reference frame. This was expected because the top/bottom axis is easier to identify (than the left/right axis) due to asymmetries of the reference objects along this axis (Bryant & Wright, 1999).

One of the important goals of the experiment was to distinguish lexical priming effects from reference-frame priming effects. A sole effect of reference frame priming would have meant that participants aligned reference frames as much when the prime and target scenes were the same (represented in the upper portion of Figure 2) as when the prime and target scenes were different (represented in the lower portion of Figure 2). However, the presence of a significant interaction between Prime reference frame and preposition condition meant that this was not the case. This interaction was caused by two of the conditions; the other three pairs of same/different conditions yielded no significant differences between them. This indicates that the apparent lexical priming effect was evident only when the relative reference frame was used and the figure and reference objects were aligned vertically. Such a situation would seem unusual, because it should be the case that lexical priming is evident for all same/different pairs of conditions.

However, there is an alternative explanation for this pattern of data that does not rely upon lexical priming. We noted that participants used intrinsic left and right differently (in fact, inversely) to the confederate. Thus, participants would describe the prime scenes in Figure 2 as *the dot left of the camera*, whereas the confederate described them as *the dot right of the camera.* Therefore for half of the match tasks in the relative, vertical, different condition, the non-matching scene also provided a match to the confederate's description if an intrinsic reference frame was applied (according to the participant's interpretation). This would be the only condition in which potential confusion could arise. Therefore, for this condition, if participants chose the non-matching scene in the match task they would effectively be primed to use the intrinsic reference frame rather than the intended relative reference frame.

In Experiment 2, we therefore made the confederate describe intrinsic left and right in the way that participants had done in Experiment 1, in order to see whether the observed interaction was due to lexical priming, or was instead an artefact of the participants' misinterpretation of what the confederate was describing as intrinsically left and right.

Experiment 2

Experiment 2 was a replication of Experiment 1, with the exception that what was described as left and right intrinsic was reversed in accordance with participants' interpretations from Experiment 1. 16 further students from the University of Edinburgh were paid volunteers in the experiment, which lasted 20 minutes.

Results

The analyses were conducted in the same fashion as Experiment 1. Table 2 shows the mean number of intrinsic responses used by subjects in each of the 8 conditions. As in Experiment 1, there was a significant main effect of Prime Reference Frame ($F1(1,15) = 6.79$; $p < .05$; $F2(1,11) = 24.36$; $p < .01$): Participants used an intrinsic reference frame more often following an intrinsic description by the confederate than following a relative description by the confederate.

Table 2:
Mean percentage intrinsic responses in Experiment 2.

	Relative Prime		Intrinsic Prime	
	Same	Diff	Same	Diff
Vertical	32.1	26.6	41.2	39.1
Horizontal	39.6	38.1	52.1	48.9

However, the interaction between Prime Reference Frame and Preposition did not reach significance, indicating that there was no effect of using the same lexical item for the prime and target ($F1(1,15) = .018$; $p > .05$; $F2(1,11) = 3.02$; $p > .05$). All other interactions were non-significant (all $p > .05$).

General Discussion

The results of this study show that interlocutors align reference frames when describing objects' locations. Importantly, the results indicate that alignment is not due to lexical priming caused by the experimental participant repeating the preposition just used by the confederate.

The apparent lexical priming effect shown in Experiment 1 was due to the participants interpreting left and right intrinsic differently to what was intended by the confederate. When the source of this difficulty was addressed in Experiment 2, this effect was not evident. The results showed no difference in the proportion of reference-frame alignment when the naïve participant used the same preposition as the confederate, as when a different preposition was used.

Our results support the hypothesis that interlocutors align at many levels of representation when conversing (Pickering & Garrod, in press). Furthermore, it extends this alignment beyond linguistic representations and into an aspect of conceptual representation, i.e., the spatial domain. These results, however, do not precisely determine the mechanism by which alignment is achieved. In particular it is not clear whether participants make some use of a deliberate strategy to make the task easier for their partner. For example, it is possible that participants may be partly aware of the importance of aligning without realizing exactly what they are aligning on.

What is surprising about these results is that there was no cumulative effect of lexical priming and reference frame priming. Other studies have shown a larger alignment effect when more factors are common between the prime and the target (e.g. Branigan et al., 2000; Cleland & Pickering, 2003). The lack of a cumulative effect may be due to the nature of the lexical items used in this experiment. The prepositions were used to refer to both their intrinsic relation and relative relation, and so held little meaning independent of the reference frame.

The results also support the work of Carlson-Radvansky and Jiang (1998) who showed that reference frames were subject to negative priming. Their investigation only focused upon inhibition along a single axis of a representation. The results of this study extend these findings and show that activation of one axis of a reference frame activates the whole reference frame (at least in 2 dimensions), indicating that reference frames are a holistic representation.

Previous work (Schober, 1993, 1995) has shown that interlocutors will co-ordinate the reference object and origin of a reference frame to the matcher in a match-and-describe task. However, this did not show that interlocutors were aligning reference frames; as Levinson (2003) has argued, it is possible to have a non-egocentric relative reference frame and an egocentric intrinsic reference frame. The results presented here provide strong evidence that interlocutors do align reference frames. Ongoing work is investigating the predictions made by Levinson's definitions of reference frames that an egocentric/intrinsic description (e.g. *the ball in front of me*) can prime the use of an allocentric/intrinsic description (e.g. *the ball in front of the car*).

Previous work (Branigan et al., 2000; Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987) has shown that interlocutors align representations during dialogue. The results of these experiments extend this body of evidence to show that independent of lexical priming, alignment extends beyond the language faculty and that interlocutors also align reference frames to describe objects' locations in a scene.

References

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000) Syntactic co-ordination in dialogue. *Cognition,* 75, B13-B25.

Bryant, D. J., & Wright, W. G. (1999). How bodily asymmetries determine accessibility in spatial frameworks. *Quarterly Journal of Experimental Psychology,* 52A, 487-508.

Carlson-Radvansky, L. A., & Jiang, Y. (1998). Inhibition accompanies reference frame selection. *Psychological Science*, 9, 386-391.

Clark, H. H. (1996). *Using language.* Cambridge: Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition,* 22, 1-39.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun phrase structure. *Journal of Memory and Language, 49*, 214-230.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition,* 27, 181-218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*, 8-11.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition,* 59, 91-117.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge MA: MIT Press.

Levinson, S. C. (1996). Frames of reference and Molyneux's question: Cross-linguistic evidence. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space,* pp. 109-169. Cambridge MA: MIT Press.

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity.* Cambridge: Cambridge University Press.

Pickering, M. J., & Garrod, S. (in press). Towards a mechanistic theory of dialogue. *Behavioural and Brain Sciences.*

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition,* 47, 1-24.

Schober, M. F. (1995). Speakers, Addressees, and frames of reference: Whose effort is minimized in conversations about locations? *Discourse Processes*, 20, 219-247

Modeling Individual Differences in Category Learning

Michael R. Webb (michael.webb@dsto.defence.gov.au)
Command and Control Division, Defence Science and Technology Organisation
Edinburgh, South Australia, 5111, AUSTRALIA

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

Many evaluations of cognitive models rely on data that have been averaged or aggregated across all experimental subjects, and so fail to consider the possibility that there are important individual differences between subjects. Other evaluations are done at the single-subject level, and so fail to benefit from the reduction of noise that data averaging or aggregation potentially provides. To overcome these weaknesses, we develop a general approach to modeling individual differences using *families* of cognitive models, where different groups of subjects are identified as having different psychological behavior. Separate models with separate parameterizations are applied to each group of subjects, and Bayesian model selection is used to determine the appropriate number of groups. We demonstrate the general approach in a concrete and detailed way using the ALCOVE model of category learning and data from four previously analysed category learning experiments. Meaningful individual differences are found for three of the four experiments, and ALCOVE is able to account for this variation through psychologically interpretable differences in parameterization. The results highlight the potential of extending cognitive models to consider individual differences.

Introduction

Much of cognitive psychology, as with other empirical sciences, involves the development and evaluation of models. Models provide formal accounts of the explanations proposed by theories, and have been developed to address diverse cognitive phenomena ranging from stimulus representation (e.g., Shepard 1980), to memory retention (e.g., Anderson & Schooler 1991; Estes 1997), to category learning (e.g., Ashby & Perrin 1988; Berretty, Todd, & Martignon 1999; Kruschke 1992; Tenenbaum 1999). One recurrent shortcoming of these models, however, is that (whether intentionally, or as an unintended consequence of methodology) humans are usually modeled as 'invariants', and not as 'individuals'. This occurs because, most often, models are evaluated against data that have been averaged or aggregated across subjects, and so the modeling assumes that there are no individual differences between subjects.

The potential benefit of averaging data is that, if the performance of subjects really is the same except for 'noise' (i.e., variation the model is not attempting to explain), the averaging process will tend to remove the noise, and the resultant data will more accurately reflect the underlying psychological phenomenon. When the performance of subjects has genuine differences, however, it is well known (e.g., Estes 1956; Myung, Kim, & Pitt 2000) that averaging produces data that do not accurately represent the behavior of individuals, and provide a misleading basis for modeling.

Even more fundamentally, the practice of averaging data restricts the focus of cognitive modeling to issues of how people are the same. While modeling invariants is fundamental, it is also important to ask how people are different. Experimental data reveal individual differences in cognitive processes, and in the psychological variables that control those processes, that also need to be modeled.

Cognitive modeling that attempts to accommodate individual differences usually assumes that each subject behaves in accordance with a different parameterization of the same basic model, and so the model is evaluated against the data from each subject separately (e.g, Ashby, Maddox, & Lee 1994; Nosofsky 1986; Wixted & Ebbesen 1997). Although this avoids the problem of corrupting the underlying pattern of the data, it also foregoes the potential benefits of averaging, and guarantees that models are fit to all of the noise in the data.

Another problem with individual subject analysis, from a model theoretic perspective, is that fitting each additional subject requires an extra set of free parameters, and so leads to a progressively more complicated accounts of the data as a whole. As has been pointed out repeatedly in the psychological literature recently (e.g., Myung & Pitt 1997; Pitt, Myung, & Zhang 2002), it is important both to maximize goodness-of-fit and minimize model complexity to achieve the basic goals of modeling. Unnecessarily complicated models that "over-fit" data often do not provide any insight or explanation of the cognitive processes they address, and are less capable of making accurate predictions when generalizing to new or different situations.

A better approach, therefore, is to partition subjects according to their individual differences, and model the averaged or aggregated data from each group. Under this approach, data are addressed by a set of models, called a model *family*, where a different parameterization is applied to each group of subjects. Where averaging is appropriate, within groups of subjects, it is applied. Where averaging is not appropriate, between groups of subjects, it is not applied.

In this paper, we apply these ideas to model individual differences in category learning, using Kruschke's (1992) well known, empirically successful, and widely used ALCOVE model. Our basic approach, however, is applicable to any model of category learning or, indeed, models of other cognitive phenomena.

Modeling Individual Differences in Category Learning

Formally, a model family $\mathcal{M}$ partitions the subjects S into G groups $S \rightarrow \{S_1, \ldots, S_G\}$, and so partitions the complete data D into G averaged data sets $D \rightarrow \{D_1, \ldots, D_G\}$. For the ith data set, a model family also specifies a model parameterization θ_i. Any possible partitioning of subjects can be considered, including the possibility that all subjects are in the same partition (corresponding to aggregating across subjects), or that each has their own partition (corresponding to a complete individual analysis). Differences in the category learning processes between groups are revealed by differences in the parameter values they use.

Because of the enormous flexibility allowed by model families, they can be made almost arbitrarily complicated, and could potentially fit any data set perfectly by adding new models, with extra parameters, to account for any remaining unexplained variation in data. It is necessary, therefore, for model fitting methods to use model selection criteria that balance goodness-of-fit and model complexity. The application of Bayesian model selection criteria (e.g., Pitt *et al.* 2002) is most easily pursued by specifying a probabilistic account, in the form of a likelihood function, of the relationship between a parameterized model family and empirical data.

To develop a likelihood function for category learning, suppose, under a proposed partitioning of subjects, the ith partition has k_i subjects, and that the n category learning trials are divided into blocks, with the jth block having b_j trials. Choosing one block with $b_1 = n$ corresponds to an analysis of the average response probabilities over all trials. Choosing n blocks with all $b_j = 1$ corresponds to a trial-by-trial analysis.

In a two category learning experiment, the data take the form of counts, d_{ij}, of the number of correct responses made by all of the subjects in the ith partition on the jth block of learning trials. Suppose also that a category learning model M, with its parameterization θ_i, predicts a correct response probability of γ_{ij} at the ith group of subjects on the jth block. Then the likelihood of the data arising under the model is given by the binomial distribution: $p(d_{ij} \mid M_i, \theta_i) = \binom{b_j k_i}{d_{ij}} \gamma_{ij}^{d_{ij}} (1-\gamma_{ij})^{b_j k_i - d_{ij}}$. The likelihood of a model family simply extends this result to consider every block of trials and every partition, so that

$$p(D \mid \mathcal{M}) = \prod_i \prod_j \binom{b_j k_i}{d_{ij}} \gamma_{ij}^{d_{ij}} (1-\gamma_{ij})^{b_j k_i - d_{ij}}. \quad (1)$$

The extension of this likelihood function to more general category learning experiments with more than two possible category responses, using a multinomial distribution, is straightforward.

Having defined the likelihood function, the Bayesian Information Criterion (BIC: Schwarz 1978) can be applied to balance goodness-of-fit with the complexity of a model family. The BIC is given by:

$$\text{BIC} = -2 \ln p(D \mid \theta^*) + P \ln N, \quad (2)$$

where P is the number of parameters in the model family (i.e., the sum of all the parameters used by the models for each group), N is the total number of data, and θ^* is the maximum likelihood parameterization over all the models. Different possible model families, corresponding to different groupings of subjects, can be compared in terms of their BIC values, with the minimum BIC corresponding to the most likely account of the data.

Demonstration Using ALCOVE

Kruschke's (1993) Study

ALCOVE is a model of category learning that uses an exemplar-based stimulus representation, similarity-based generalization that is mediated by selective attention, and error-based learning from external feedback. The standard ALCOVE model Kruschke (1992) uses four free parameters. These control the rate of learning for attention weights (λ_a), the rate of learning for the associations between stimulus representations and category responses (λ_w), the gradient of the generalization function that measures stimulus similarity (c), and the way in which different levels of evidence for category alternatives are mapped onto response probabilities (ϕ).

Kruschke (1993) considered the ability of ALCOVE to model human category learning for filtration and condensation Categorization tasks (Garner 1974). The results of four separate experiments were reported, covering two filtration tasks (called position-relevant and height-relevant, due to the nature of the stimuli) and two condensation tasks (called condensation A and

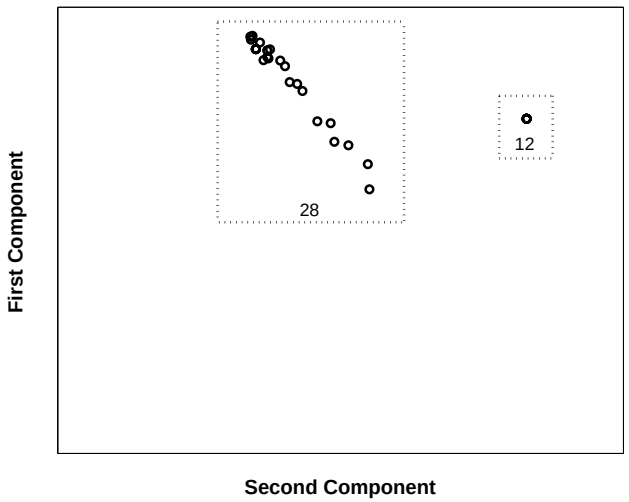


Figure 1: The application of the heuristic for partitioning subjects to find two groups for the position-relevant filtration data.

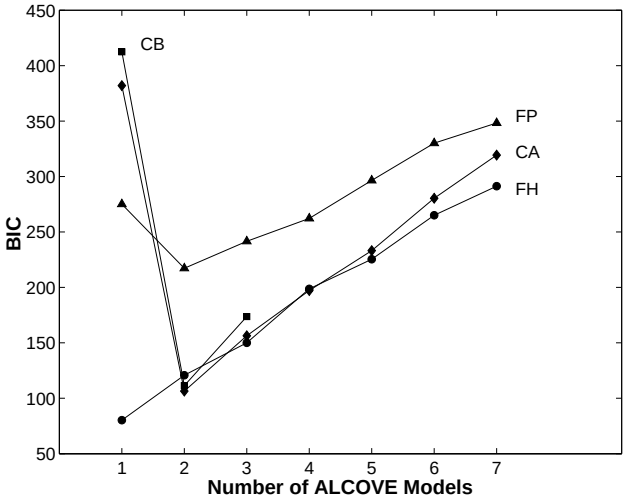


Figure 2: The pattern of change in BIC values for each clustering of the position-relevant filtration (FP), heigh-relevant filtration (FH), condensation A (CA) and condensation B (CB) category learning data.

condensation B). The data involved a total of 160 subjects, with 40 completing each task. Kruschke (1993) fit ALCOVE to all four sets of experimental results simultaneously, using trial-by-trial data formed by averaging across all 40 subjects. An examination of the individual learning curves in the raw data, however, reveals a large degree of variation between subjects within each experiment, and raises the possibility that there are psychologically meaningful individual differences in category learning.

Heuristic for Partitioning Subjects

In classification and clustering, an essential requirement for the determination of homogenous classes is a calculable similarity or distance measure between objects being compared (Gordon 1999). For category learning, the objects are the individual experimental observations for each subject, (i.e., each subject's learning curve). A candidate measure for describing the similarities between these curves is the correlation coefficient, which we used in a two-stage heuristic. In the first stage, singular value decomposition is applied to produce an ordered eigenvector-based representation of the similarities between the learning curves of subjects. In the second stage, a simple k-means clustering algorithm is applied to this representation to find clusters of subjects.

For each of Kruschke's (1993) four category learning tasks, this heuristic was applied to produce a range of partitions of the data, from a single group with all 40 subjects, to seven groups with differing numbers of subjects in each group. As a concrete example of this process, the clusters found when the subjects were divided into two groups for the position-relevant filtration task are shown in Figure 1. Each circle represents the learning curve of a subject, represented according to their values along the first two component eigenvectors. The two groups of subjects identified by k-mean clustering are superimposed using broken lines. One cluster on the left encompasses 28 of the subjects, while a much tighter cluster on the right encompasses the remaining 12 subjects.

Model Fitting and Evaluation

For each of the clusterings for each task, maximum likelihood fits of ALCOVE were found using a different parameterization for each group according to Eq. (1). BIC values were then calculated for each model family using Eq. (2), giving the results[1] shown in Figure 2. It is clear that the minimum BIC for three of the four tasks (position-relevant filtration, condensation A and condensation B) is achieved when two separate groups of subjects are considered, while the height-relevant filtration data are best modeled by considering all of the subjects as learning in the same way.

Figures 3 and 4 give more detailed results for, respectively, the position-relevant filtration and condensation

[1]The full range of BIC values for the CB task is not shown because, when four or more groups are considered, at least one of the groups contains only subjects who become less accurate as learning blocks progress. ALCOVE is qualitatively unable to accommodate the decrease in the averaged learning curve for this type of group, leading to very poor fit, and very large BIC values. We have omitted these values.

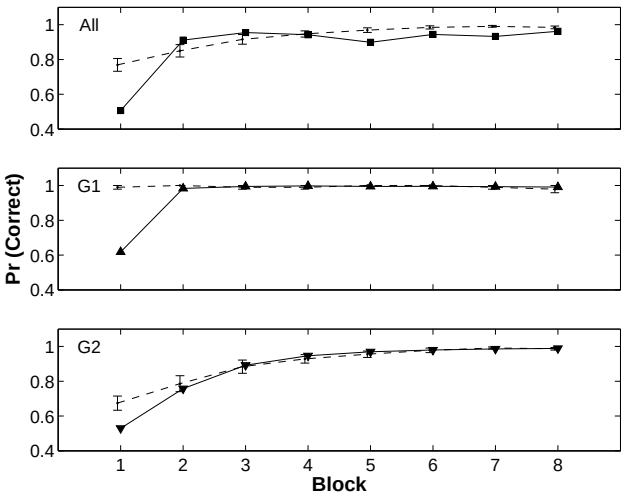


Figure 3: The change in accuracy across learning blocks for the subjects (broken lines) and ALCOVE (solid lines), for the one group ("All") and two group ("G1" and "G2") model families on the position-relevant filtration task.

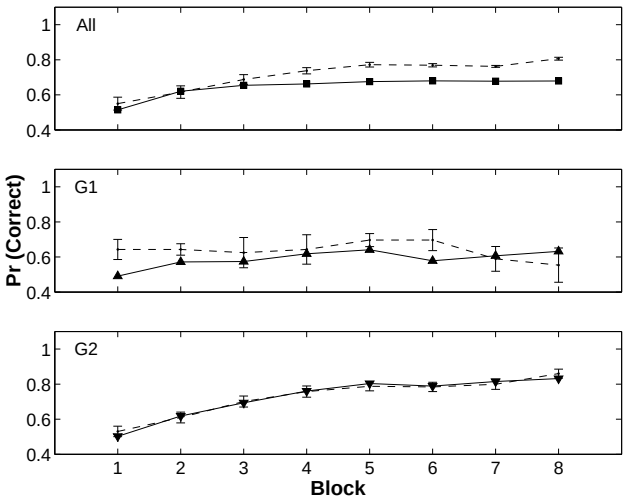


Figure 4: The change in accuracy across learning blocks for the subjects (broken lines) and ALCOVE (solid lines), for the one group ("All") and two group ("G1" and "G2") model families on the condensation A task.

A tasks. In both of these figures, the top panel, labeled "All", shows the average accuracy of all subjects across the eight learning blocks, and the maximum likelihood fit of ALCOVE to these data. The middle and bottom panels show the first (G1) and second (G2) groups of subjects proposed by the two-group model family that is prefered by the complexity analysis. These panels show the average accuracy for both groups of subjects separately, together with the maximum likelihood ALCOVE learning curve.

Figure 3 shows that the moderate learning evident when treating the subjects as having no individual differences is better modeled as coming from two distinct groups of subjects. Some subjects, in the first group, maintain near-perfect accuracy throughout the category learning task. Other subjects, in the second group, learn more gradually, only achieving near-perfect accuracy in the last few learning blocks. Figure 3 shows that, with the exception of the rapid achievement of accuracy in the first block for the first group of subjects, ALCOVE is able to model both of these patterns of learning[2].

In a similar way, Figure 4 shows that the gradual increase in accuracy, evident when treating the subjects as having no individual differences, is better modeled as coming from two distinct groups of subjects. The first group exhibits almost no learning, while the second learns at a moderate rate. Once again, ALCOVE is able to model both of these patterns of learning. In fact, ALCOVE has more difficulty accommodating the learning data resulting from averaging across all of the subjects. What the individual differences analysis developed here suggests is that this inability may not indicate a fundamental weakness in ALCOVE, but rather that the averaging process involved in summarizing human performance has masked important individual differences, and corrupted the underlying learning patterns in the original data.

Table 1 shows the maximum likelihood parameter values for each group of subjects in the model family with the lowest BIC value, for all four learning tasks. These parameter values are generally interpretable in terms of the different learning behavior revealed by the individual differences analysis. For example, for the position-relevant filtration task, the first group of subjects have a greater λ_w value than the second group, consistent with their more rapid learning. For this task, both groups have high ϕ values, consistent with their decisiveness (or 'confidence') in mapping evidence into response probabilities. Both groups of subjects in the condensation A task, however, have much lower ϕ values, consistent with their inferior learning performance, and the first group in this task, who basically fail to learn, have a very low ϕ value. Other comparisons of this type, both within and across tasks, generally have meaningful and useful interpretations,

[2]It is possible the application of one of ALCOVE's descendents, such as RASHNL (Kruschke & Johansen 1999) or the unified mixture of experts model (Kruschke 2001), which emphasize rule-oriented learning and incorporate a rapid attention shifting capability (Kruschke 1996), could overcome the deficiency.

Table 1: Maximum likelihood parameter values for each group of subjects in the model family with the lowest BIC value, for all four learning tasks. FP=position-relevant filtration, FH=height-relevant filtration, CA=condensation A, CB=condensation B.

Task	Group	λ_w	λ_a	c	ϕ
FP	G1	0.38	0.49	1.68	3.20
	G2	0.06	27.0	6.83	2.66
FH	All	0.23	0.58	1.56	1.00
CA	G1	0.47	1.14	2.53	0.27
	G2	0.24	0.38	7.52	0.93
CB	G1	0.41	0.32	0.79	0.31
	G2	0.17	0 .02	3.37	1.09

and highlight the ability of ALCOVE to represent psychologically important variations in category learning through its free parameters.

Discussion

There are at least two conclusions that can be drawn from modeling individual differences in Kruschke's (1993) category learning data using ALCOVE. The first is that there is strong evidence for large and meaningful differences in the learning behavior of groups of subjects for three out of the four tasks. Previous analyses, adopting the standard cognitive modeling practice of considering all of the subjects as a single group, are insensitive to these potentially important patterns of variation. The second conclusion is that, for these data, the basic ALCOVE model is generally able to capture the individual differences in learning, when asked to model appropriate groups of subjects. It does this by applying different psychologically meaningful parameterizations to accommodate variations in learning behavior. In this sense, what the results presented here demonstrate is that accounting for individual differences using model families has the potential to extend and increase the usefulness of existing cognitive models significantly.

From this promising start, there are a number of directions in which the basic approach described here can be refined and extended. Most generally, the extension to other cognitive phenomena provides a rich set of opportunities for future research. As with category learning, there is evidence of individual differences in the similarity data used to model stimulus representations (e.g., Ashby *et al.* 1994), and in the curves of forgetting used to model memory retention (e.g., Anderson & Tweney 1997; Heathcote, Brown, & Mewhort 2000; Myung, Kim, & Pitt 2000; Wixted & Ebbesen 1997), and in a range of other data from which cognitive models have been developed.

Considering a broader range of cognitive phenomena highlights the possibility of extending individual difference accounts to incorporate fundamentally different models to capture between-subject variation, rather than relying solely on parametric variation within the same basic model. In memory retention, for example, one group of subjects could be modeled using a power function while another group is modeled using an exponential decay function. For stimulus representation, some groups of subject could be modeled using a featural representation while others use a dimensional representation. In the category learning context considered here, it may make sense to model some subject groups using ALCOVE or its descendants, but apply a very different category learning model to others, such as the fast and frugal account provided by Categorization-By-Elimination (Berrety *et al.* 1999).

One of the weaknesses of the demonstration presented here is the reliance on the BIC to compare different competing individual differences models. While the BIC is conceptually and computationally straightforward, it is insensitive to the complexity effects arising from the functional form of parametric interaction within the individual models (Myung & Pitt 1997). This is a potentially important shortcoming, especially if fundamentally different models are used to explain performance for different subject groups. There are, for example, many competing models of retention that use two parameters (Rubin & Wenzel 1996), with different complexities that the BIC is unable to distinguish. The obvious remedy for this problem is to use more sophisticated model selection criteria that are sensitive to all of the components of model complexity. These include measures such as the Stochastic Complexity Criterion (SCC: Rissanen 1996) and Normalized Maximum Likelihood (NML: Rissanen 2001). For cognitive models that resist the formal analysis needed to derive these measures, an alternative is to use numerical methods, such Markov Chain Monte Carlo (e.g., Gilks, Richards, & Spiegelhalter 1996) to approximate the Bayesian posterior distributions that compare model families.

A final possibility for refining the approach demonstrated here is to use a more principled optimization approach to determine the groupings of subjects. The method used here, based on k-means clustering of correlations, is a sensible heuristic one. It is particularly well suited to a model like ALCOVE that requires considerable computation effort when finding maximum likelihood parameter values. The clustering heuristic is designed to identify good partitions of the subjects into groups, and only requires parameter fitting to be done once for each possible number of subject groups. For other models, however, such as analytic models of memory retention, finding maximum likelihood parameterizations is straightforward. In these cases, a more explicit optimization approach to finding partitions could be adopted, because repeated pa-

rameter fitting is possible. For example, a stochastic hill-climbing procedure could be used to find subject groups that minimize the BIC, SCC or NML of the model family.

Collectively, these possibilities describe a principled and general approach for building and evaluating cognitive models, using a variety of basic models and numbers of parameterizations, to accommodate individual differences. It is a more general approach to cognitive modeling than one that averages data, assuming there are no individual differences. It is a more powerful and succinct approach than one that uses subject-by-subject analysis. While much of the work to realize this potential remains to be done, the demonstration presented here, using multiple ALCOVE models to capture differences in category learning, provides a good concrete example of its potential. It shows how using model families, and relying on principled model selection criteria, can be used to develop detailed and interpretable accounts of both how people are cognitively the same, *and* how they are different.

Acknowledgments

This research was supported by Australian Research Council Grant DP0451793.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science 2*(6), 396–408.

Anderson, J. R., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition 25*, 724–730.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science 5*(3), 144–151.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review 95*(1), 124–150.

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination. In G. Gigerenzer & P. M. Todd (Eds.), *Simple Heuristics That Make Us Smart*, pp. 235–254. New York: Oxford University Press.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin 53*(2), 134–140.

Estes, W. K. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review 104*(1), 148–169.

Garner, W. R. (1974). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gordon, A. D. (1999). *Classification* (Second ed.). London: Chapman & Hall/CRC Press.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review 7*(2), 185–207.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review 99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science 5*, 3–36.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition 22*, 3–26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology 45*, 812–863.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition 25*(5), 1083–1119.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition 28*(5), 832–840.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review 4*(1), 79–95.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*(1), 39–57.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*(5), 1712–1717.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review 103*(4), 734–760.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science 210*, 390–398.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11*, Cambridge, MA, pp. 59–65. MIT Press.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition 25*(5), 731–739.

The Origin of the Linguistic Gender Effect in Spoken-Word Recognition: Evidence from Non-Native Listening

Andrea Weber (aweber@coli.uni-sb.de)
Dept. of Computational Psycholinguistics, Saarland University, 66041 Saarbrücken, Germany

Garance Paris (gparis@coli.uni-sb.de)
Dept. of Computational Psycholinguistics, Saarland University, 66041 Saarbrücken, Germany

Abstract

Two eye-tracking experiments examined linguistic gender effects in non-native spoken-word recognition. French participants, who knew German well, followed spoken instructions in German to click on pictures on a computer screen (e.g., *Wo befindet sich die Perle*, "where is the pearl") while their eye movements were monitored. The name of the target picture was preceded by a gender-marked article in the instructions. When a target and a competitor picture (with phonologically similar names) were of the same gender in both German and French, French participants fixated competitor pictures more than unrelated pictures. However, when target and competitor were of the same gender in German but of different gender in French, early fixations to the competitor picture were reduced. Competitor activation in the non-native language was seemingly constrained by native gender information. German listeners showed no such viewing time difference. The results speak against a form-based account of the linguistic gender effect. They rather support the notion that the effect originates from the grammatical level of language processing.

Introduction

Gender is a grammatical category that varies largely across the languages of the world. The range goes from elaborate gender systems in some languages to the absence of gender in others. Both German and French are languages with grammatical gender. The form of definite articles, for example, marks gender in both languages. German definite articles are $der_{(masc.)}$, $die_{(fem.)}$, and $das_{(neut.)}$; French definite articles are $le_{(masc.)}$ and $la_{(fem.)}$ respectively. Grammatical gender usually becomes most noticeable when we learn a second language with gender. Is it *der Berg* ("mountain") or *die Berg* in German? Do the French say *le citron* ("lemon") or *la citron*? The present study investigated how gender marking influences the recognition of spoken-words in a non-native language. Results help clarify the origin of the linguistic gender effect.

It is generally accepted in the psycholinguistic community that during the recognition of spoken words, multiple word candidates get simultaneously activated and compete against each other (e.g., Marslen-Wilson & Welsh, 1978; McQueen, Norris, & Cutler, 1994). When a native speaker hears, for example, the German word *Perle* ("pearl"), lexical representations of words with similar onsets, such as *Perücke* ("wig"), will initially be activated along with *Perle*. Activated word candidates compete for recognition until they no longer match incoming segmental information. Thus, *Perücke* will drop out of the competitor set as the /l/ in *Perle* is being heard. It has also been shown that non-native listeners consider candidate words in both the non-native and their native language simultaneously (e.g., Marian & Spivey, 2003; Weber & Cutler, 2004). Thus, for French listeners the beginning of German *Perle* may additionally activate French words like *perruque* and *persil*.

Eye-tracking is a methodology that has been found to be eminently suited for the investigation of competitor activation (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). It makes use of the fact that participants make saccadic eye movements to pictures of objects on a computer screen as the names of the objects are mentioned in spoken sentences. Locations and latencies of eye movements on pictures are recorded using a camera mounted on a headband and can be used to examine lexical competition in spoken-word recognition. While participants hear the name of a target picture, they look more often to pictures with names that are similar in onset with the target name than to pictures with phonologically unrelated names. It has been shown that such competition effects, defined as fixation proportions to pictures, closely map to activation levels of word candidates as simulated in computational models of spoken-word recognition such as TRACE (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001).

By now, numerous eye-tracking studies have successfully confirmed phonological competitor activation in spoken-word recognition. Dahan, Swingley, Tanenhaus, and Magnuson (2000) took the subject one step further by testing whether morphosyntactic context can affect competitor activation. In particular, they tested whether gender marking on definite articles influences the recognition of subsequent nouns. A number of studies had already looked at lexical gender effects in word recognition using experimental paradigms other than eye-tracking (e.g., Bates, Devescovi, Hernandez, & Pizzamiglio, 1996; Colé & Segui, 1994; Grosjean, Dommergues, Cornu, Guillelmon, & Besson, 1994). These studies found that the presence of gender-congruent articles or adjectives enhances the recognition of target nouns whereas gender-incongruent forms slow recognition down. Dahan et al. (2000), investigated the role of gender information on spoken-word recognition more directly: They tested the activation of competitors that matched the initial sounds of a target noun

but mismatched the gender marking on the article. They found that the presence of a gender-marked definite article could prevent early activation of competitors inconsistent with that gender: Upon hearing *cliquez sur le bouton* ("click on the$_{(masc.)}$ button"), French listeners did not fixate the picture of a *bouteille* ("bottle$_{(fem.)}$") more often than pictures with unrelated names.[1]

The study by Dahan et al. (2000), however, could not assess the origin of the lexical gender effect. Are listeners really sensitive to grammatical gender information in the preceding context or is it simply listeners' sensitivity to the co-occurrence of the form of the article with the form of the noun that restricts lexical access? In order to reduce the high co-occurrence of definite articles and nouns, Dahan and colleagues interposed a gender-marked adjective in a follow-up study. In their preliminary results, activation of gender-mismatching competitors was no longer reduced when low frequency gender-marked adjectives preceded target nouns. This was seen as evidence for a form-based origin of the gender effect. In a Russian eye-tracking study, however, Sekerina (2003) found that gender-marked color adjectives do restrict referential sets to gender-matching nouns. She interpreted the results as evidence for a grammar-based effect of gender in spoken-word recognition.

Spoken-word recognition in a non-native language offers the possibility to distinguish between a form-based and a grammar-based account of the linguistic gender effect. The gender of a noun can differ across languages: *Canon* is, for instance, feminine in German but masculine in French. The present study tested whether French listeners, who are highly proficient in German, use native French gender information during the recognition of spoken words in German. Since the form of the article differs in German and French, presentation of the German article should not give rise to co-occurrence information for the French form of the article and a given noun. Thus, if the gender of words in French exerts an effect on the recognition of spoken words in German (even though French is not presented), this would strongly suggest that the locus of the gender effect is not form-based.[2]

Recent eye-tracking studies have shown that listeners cannot deactivate the lexicon of the native language even in a monolingual non-native situation where the native vocabulary is irrelevant (Marian & Spivey, 2003; Spivey & Marian, 1999; Weber & Cutler, 2004). Native language competitors that were phonologically related to the non-native target were activated more than phonologically unrelated words: Upon hearing the English target *desk*, Dutch listeners, who knew English well, fixated the picture of a lid more than unrelated pictures because the Dutch name for lid (*deksel*) was phonologically related to *desk* (Weber & Cutler, 2004). Similarly, grammatical information from the native language might interfere with non-native listening. Imagine native French speakers listening to German in an eye-tracking study. Spoken instructions in German tell them to click on target pictures on a screen. The name of the target picture is preceded by the definite article in the instructions, and target and competitor names overlap in onset in both languages. In the non-native presentation language German, target and competitor names share gender, so the gender marking on the article cannot exclude the competitor as a lexical candidate. In the native language French, however, target and competitor differ in gender. If we find no competitor activation for French listeners, this would suggest that they use native French gender information to disambiguate between target and competitor.

Experiment 1

Method

Participants Eighteen native speakers of French, mostly students (mean age of 22), took part in the experiment for monetary compensation. They had normal or corrected-to-normal vision and normal hearing. On average, they had studied German as a foreign language for 10 years in secondary education, starting at a mean age of 12 (ranging from 10 to 16). To confirm their high proficiency in the non-native language, they underwent a vocabulary test in German after completing the eye-tracking experiment. For each target and competitor noun in the experiment plus a number of filler nouns with neuter gender, they had to name the correct gender. The average score was 78% correct.

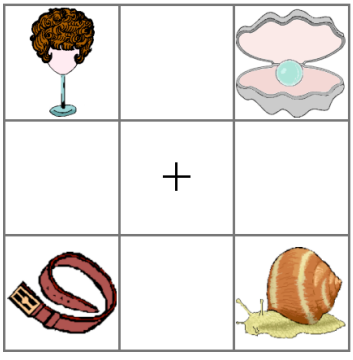


Figure 1: Example of visual display presented to participants.

[1] Dahan et al. (2000) also showed that when no phonological overlap between picture names was given, gender-marked articles were not sufficient to restrict participants' attention to pictures with gender matching names.

[2] Only very few studies have looked at gender marking effects in non-native spoken-word recognition. Guillelmon and Grosjean (2001), for example, found in an auditory naming study no effects of congruency for late English-French bilinguals. It is not established yet whether gender marking influences competitor activation in a second language.

Table 1: Examples of German (G) target-competitor pairs and their French (F) translations.

		target	competitor
same-gender pair	G	Perle$_{(fem.)}$	Perücke$_{(fem.)}$
	F	perle$_{(fem.)}$	perruque$_{(fem.)}$
		"pearl"	"wig"
different-gender pair	G	Kassette$_{(fem.)}$	Kanone$_{(fem.)}$
	F	cassette$_{(fem.)}$	canon$_{(masc.)}$
		"tape"	"canon"

Materials Thirty German nouns referring to picturable objects were chosen as targets.[3] Each target was paired with a competitor. The onset of the competitor overlapped phonemically with the onset of the target in both German and French (e.g., German target *Perle* /pɛrlə/ was *perle* /pɛrl/ in French; German competitor *Perücke* /pɛrykə/ was *perruque* /pɛryk/ in French). The target was always of the same gender in German and French, but the gender of the competitor divided the pairs into two groups (see Table 1). In 15 "same-gender" pairs, target and competitor shared gender in both languages. The target *Perle* ("pearl"), feminine in both German and French, was for example paired with the competitor *Perücke* ("wig"), also feminine in both languages. In these pairs, neither German nor French gender information could constrain initial competitor activation. In 15 "different-gender" pairs, target and competitor still shared gender in German, but were of different genders in French. The target *Kassette* ("tape"), feminine in both languages, was for instance paired with the competitor *Kanone* ("canon"), which is feminine in German but masculine in French. Whereas German gender information could not exclude the competitor as a potential lexical candidate in these pairs, French gender information could.

Two phonologically unrelated distractors, with random gender, were added for each target (e.g., *snail* and *belt*). Neither the German nor the French names of the unrelated distractors overlapped with the German target nouns. The target was heard in the experiment, whereas competitor and unrelated distractors were not heard. The overall lexical frequency of targets and competitors did not differ significantly in either of our target-competitor pairs.

Thirty filler trials were added. Great care was taken in the fillers to dispel expectations that pictures with phonologically similar names or matching gender were likely targets. Three more representative trials were constructed as practice trials.

All pictures were colored line drawings, taken from the IMSI MasterClips Image Collection (1990). In pre-tests, we asked participants to name and rate target and competitor pictures. The agreement between participants' responses and intended names was 88% correct, and the goodness of the pictures was rated with a mean of 5 on a scale from 0 to 6.

German target nouns, preceded by their definite article with nominative case marking, were embedded in a carrier sentence (e.g., *Wo befindet sich die Perle*, "Where is the pearl"). Spoken instructions were recorded. The duration of putative overlap between target and competitor (e.g., the duration of /pɛr/ in *Perle*) was on average 200 ms for same-gender pairs and 174 ms for different-gender pairs.

Procedure Participants were tested individually. At the beginning of a session, they received instructions in German, telling them to click on the object on the screen that was mentioned in a sentence. Sentences were presented auditorily over headphones and started 550 ms after the appearance of the pictures on the screen. The set of pictures was not shown to the participants before the experiment.

While they were listening, participants' eye movements were monitored using an SMI EyeLink head-mounted eye-tracker. A camera on the participants' dominant eye provided the input to the tracker. Onset and offset times and the spatial coordinates of the participants' fixations were recorded (250 Hz sampling rate). All pictures were presented in color on a 3 x 3 gray grid (see Figure 1). Each cell measured 7.5 x 7.5 cm, corresponding to a visual angle of approximately 7°, which is well within the resolution of the eye-tracker (better than 1°). The pictures of a target item, its competitor, and two unrelated distractors were displayed together in one trial. Positions of target and competitor objects were randomized across trials. Each experimental trial was preceded by at least one filler trial. Along with the eye movements, the position of the mouse click was recorded.

For the analysis, graphical software was used to display the locations of the participants' fixations as dots superimposed on the four pictures for each trial and each participant. Fixations were coded as pertaining to the cell of the target object, the competitor object, or one of the two unrelated distractors. Fixations that lay clearly outside the cell of an object were not used for the computation of the fixation probabilities. Saccade times were not added to fixation times.

Results and Discussion

Seventeen trials were removed from the analysis because participants clicked on an object other than the target or no fixation on the target object was found (3.2% of all trials). The low percentage of errors suggests that French participants had no difficulties performing the task in German. Fixation proportions, at successive 10 ms time frames, were averaged over participants and items for separate analyses.

[3] Since the French gender system is limited to feminine and masculine, selected German target nouns were either of feminine or masculine gender, but never neuter.

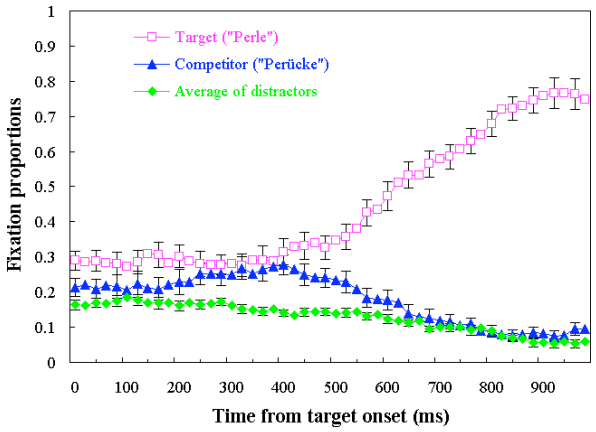


Figure 2: Same-gender pairs. Fixation proportions of French listeners over time for German targets, competitors, and averaged distractors.

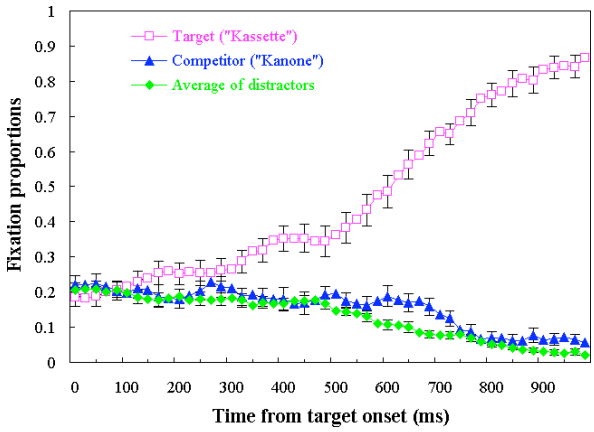


Figure 3: Different-gender pairs. Fixation proportions of French listeners over time for German targets, competitors, and averaged distractors.

Figure 2 presents the averaged proportions of fixations after target noun onset for trials with same-gender pairs. Fixation proportions for the two unrelated distractors were averaged. It takes typically about 150 to 200 ms before a programmed eye movement is launched (e.g., Matin, Shao, & Buff, 1993). Thus, fixations on the target object that are triggered by acoustic information are observable starting around 200 ms after target noun onset.

In same-gender pairs, French listeners fixated competitor objects more than distractor objects. Between 200 and 600 ms, the proportion of fixations was on average 23.9% for the competitor and 14.9% for the unrelated distractors. A one-factor ANOVA on the mean proportion of fixations between 200 and 600 ms, with picture (with the two levels 'competitor' and 'unrelated distractors') as the within-participants factor, showed that the competitor was fixated significantly more than the average of the unrelated distractors ($F_1[1, 17] = 11.41$, $p < .005$; $F_2[1, 14] = 13.92$, $p < .005$). Neither gender information from the non-native presentation language, nor gender information from their native language could narrow the lexical candidates down to the target. In consequence, the competitor was activated during the presentation of the target due to their phonological similarity.

Prior to the point that fixations could be driven by acoustic information from the target noun, no variation between fixation proportions was found. Analyses in the 0-200 ms time window showed no reliable difference in initial fixations between competitor and unrelated distractors ($F_1[1, 17] = 2.15$, $p > .1$; $F_2 < 1$). Thus, the difference between fixations to the competitor and the unrelated distractors in the 200-600 ms time window cannot be attributed to a general bias toward the picture of the competitor.

The pattern of results changed for different-gender pairs. French participants no longer fixated competitor objects more than distractor objects (see Figure 3). Over the 200-600 ms time window, 17.9% of the fixations were on average to the competitor and 15.6% to the distractors. A one-factor ANOVA confirmed the lack of a difference in viewing times (F_1 & $F_2 < 1$). As before, no reliable differences were found for different-gender pairs in initial fixation proportions between 0 and 200 ms after target noun onset (F_1 & $F_2 < 1$).

In different-gender pairs, gender information carried by the article in German could not constrain competitor activation, but French gender could. Despite its phonological similarity with the target noun, the competitor was not activated when the article of the target noun did not match in gender with the competitor in French. Evidently, French listeners used native French gender information to constrain competitor activation in German.[4] The experiment was conducted in German, and the linguistic form of the article did not exclude the competitor as a potential lexical candidate. In other words, the probability of the target noun being *Perle* or *Perücke* was equally high after hearing the phoneme sequence /di:pɐr/, *die Per*. Nevertheless, competitor activation was eliminated for French listeners. This suggests that the high form-based co-occurrence of article and target did not constrain lexical access in our experiments, but rather grammatical gender carried by the article did.

Experiment 2

As a control, we presented the same stimuli to listeners whose native language was German. If native gender information of the pictures had restricted eye movements in Experiment 1, both same-gender pairs and different-gender pairs should now offer competition for German listeners in Experiment 2, since in German target and competitor share gender in both pairs.

[4] The same analyses were run again after removing trials for which the French native speakers made a mistake in the vocabulary test. The results were comparable.

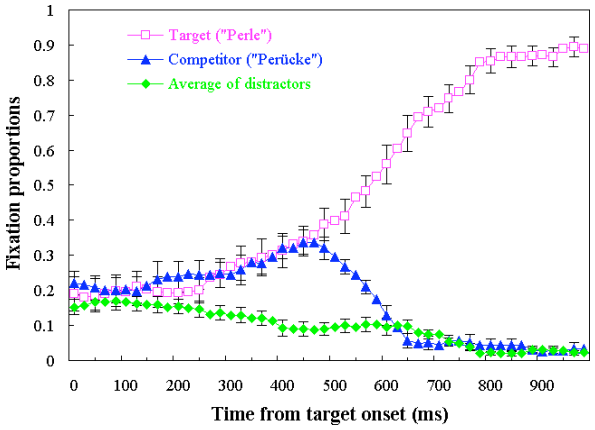


Figure 4: Same-gender pairs. Fixation proportions of German listeners over time for German targets, competitors, and averaged distractors.

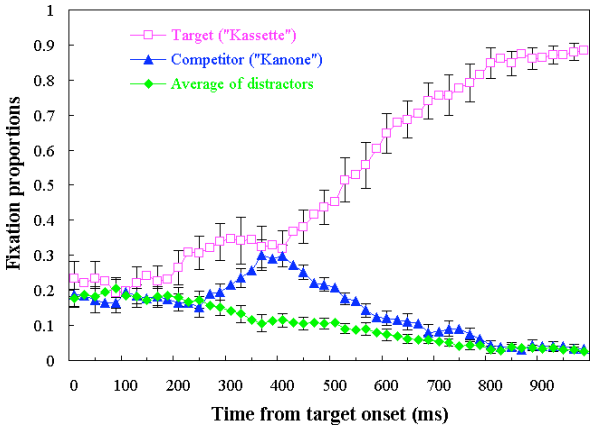


Figure 5: Different-gender pairs. Fixation proportions of German listeners over time for German targets, competitors, and averaged distractors.

Method

Participants Twelve native speakers of German participated, in return for a small payment. They were all students (mean age of 21), and had normal or corrected-to-normal vision and normal hearing. They had all learned French as a second language in school, but were not required to exercise their proficiency here.

Materials The materials were as in Experiment 1.

Procedure The procedure was as in Experiment 1. Participants were not made aware of potential cross-language competition in the experiment.

Results and Discussion

Participants never clicked on an object other than the target. Figure 4 shows the averaged proportions of fixations after target noun onset for trials with same-gender pairs. As is immediately apparent, higher fixation probabilities were observed for the competitor than for the unrelated distractors. Over the 200-600 ms time window, the proportion of fixations was on average 26.74% for the competitor and 11.16% for the unrelated distractors. This difference was significant in a one-way ANOVA ($F_1[1, 11] = 22.97$, $p < .002$; $F_2[1, 14] = 19.47$, $p < .01$). Just as the French listeners in Experiment 1, German listeners activated competitors when gender-marked articles could not exclude them as potential lexical candidates. No reliable difference in viewing times was observed in the first 200 ms after target noun onset ($F_1[1, 11] = 1.30$, $p > .2$; $F_2[1, 14] = 1.97$, $p > .1$).

In contrast to the French listeners, however, German listeners also looked more often at the competitor than at the unrelated distractors in different-gender pairs. Between 200-600 ms after target noun onset the proportion of fixations was on average 20.98% for the competitor and 11.84% for the unrelated distractors. An ANOVA showed a significant effect of type of picture ($F_1[1, 11] = 10.68$, $p < .01$; $F_2[1, 14] = 8.34$, $p < .02$). Again, viewing times for competitor and unrelated distractors did not differ in the first 200 ms after target noun onset (F_1 & $F_2 < 1$).

The results of Experiment 2 showed that during the presentation of the target noun, German listeners activated the competitor in both same-gender and different-gender pairs.

Summary

A recent eye-tracking study by Dahan et al. (2000) has shown that grammatical context can constrain lexical access. In their study, French participants followed spoken instructions in French to click on pictures on a screen while their eye movements were monitored. Eye movements to pictures were interpreted as evidence for the activation of the words corresponding to those pictures. We know from previous eye-tracking studies that competitor pictures with names that overlap in onset with the name of a target picture are fixated more than pictures with unrelated names (see e.g., Tanenhaus et al., 1995). In the spoken instructions in Dahan et al's study, the names of the target pictures were immediately preceded by articles. In the absence of gender marking on the article (i.e., French plural article *les*), competitor activation was found for phonologically related nouns. However, when competitors matched in initial sounds with a target noun but mismatched in gender marking on the preceding article, early competitor activation was eliminated.

The present eye-tracking studies investigated the role of linguistic gender for the process of listening to a non-native language. An interesting aspect of non-native listening is that the gender of words can vary between the native and the non-native language. Thus, gender information as conveyed

by the presentation language, i.e. the non-native language, can be opposed to gender information from the listeners' native language. In Experiment 1, French participants followed spoken instructions in German to click on pictures on a computer screen (e.g., *Wo befindet sich die Perle*, "Where is the pearl"). When target and competitor noun shared gender in both German and French, French participants fixated competitor pictures more than unrelated pictures. However, when target and competitor were of the same gender in German but of different gender in French, early fixations to the competitor picture were eliminated. This result was interpreted as evidence that competitor activation in the non-native language was constrained by native gender information. In Experiment 2, German listeners were presented with the same materials and showed no such difference in viewing time.

In general, our results support Dahan et al's (2000) findings that gender information influences lexical access, but also crucially offer new insights with respect to the origin of the gender effect. On one account, listeners compute distributional regularities between the co-occurrence of the form of the article and the form of the noun and use these form-based regularities to restrict lexical access. On another account, distributional regularities would be computed using grammatical categories. On the form-based account, probabilities would express the likelihood of the target being *Perle* upon hearing the segmental sequence /diːpɛr/; on the grammar-based account, probabilities would express the likelihood of the target being *Perle* upon hearing /pɛr/ plus having feminine gender information from the context. Within one language, these two accounts are difficult to tease apart. However, non-native listening offered the possibility to separate them, because linguistic gender effects of the non-presentation language are unlikely to be caused by form-based regularities of that language.

The fact that, for French listeners in Experiment 1, competitor activation in German was eliminated when French gender information mismatched the gender of the target speaks against an form-based account of the linguistic gender effect. Our results rather support the notion that the linguistic gender effect originates from the higher, grammatical level of language processing.

Acknowledgments

We thank Delphine Dahan for inspiring discussions on the topic. Further thanks go to Alissa Melinger for helpful comments on an earlier version of this paper. This research was funded by SFB 378 "ALPHA" to the first author, awarded by the German Research Council.

References

Allopenna, P., Magnuson, J., & Tanenhaus,. M. (1998). Tracking the time course of spoken-word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419-439.

Bates, E., Devescovi, A., Hernandez, A., & Pizzamiglio, L. (1996). Gender priming in Italian. *Perception & Psychophysics*, *85*, 992-1004.

Blair, I., Urland, G., & Ma, J. (2002). Using internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers*, *34*, 286-290.

Colé, P., & Segui, J. (1994). Grammatical incongruency and vocabulary types. *Memory & Cognition*, *22*, 387-394.

Corbett, G. (1991). *Gender*. Cambridge: Cambridge University Press.

Dahan, D., Magnuson, J., & Tanenhaus, M. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye-movements. *Cognitive Psychology*, *42*, 317-367.

Dahan, D., Swingley, D., Tanenhaus, M., & Magnuson, J. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language*, *42*, 465-480.

Grosjean, F., Dommergues, J., Cornu, E., Guillelmon, D., & Besson, C. (1994). The gender-marking effect in spoken-word recognition. *Perception & Psychophysics*, *56*, 590-598.

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory and Cognition*, *29*, 503-511.

IMSI Master Clips (1990). *Premium image collection 303,000*. (http://www.imsisoft.com)

Marian, V., & Spivey, M. (2003). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics*, *24*, 173-193.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology*, *10*, 29-63.

Matin, E., Shao, K., & Buff, K. (1993). Saccadic overhead: information processing time with and without saccades. *Perception & Psychophysics*, *53*, 372-380.

McQueen, J., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 621-638.

Sekerina, I. (2003). Grammatical gender and mapping of referential expressions in Russian. Talk presented at the *9th Annual Conference on Architectures and Mechanisms for Language Processing*, Glasgow, Scotland.

Spivey, M., & Marian, V. (1999). Crosstalk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, *10*, 281-284.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken-language comprehension. *Science*, *268*, 1632-1634.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*, 1-25.

Structural Differences in Abstract and Concrete Item Categories

Katja Wiemer-Hastings (Katja@niu.edu)
Kimberly K. Barnard (Barnardkk@earthlink.net)
Jon Faelnar (JDF1078@aol.com)
Department of Psychology, Northern Illinois University
DeKalb, IL 60115 USA

Abstract

We explored reasons why categories of abstract items, such as *cognitive processes* or *communication*, are weak determinants of abstract item similarity. Three experiments, using exemplar listing tasks and similarity ratings on the exemplars, compared the structure of taxonomic categories of abstract versus concrete entities. In comparison to concrete item categories, we found that there were consistently fewer typical items for abstract item categories, but that nonetheless item pairs within abstract item categories were rated about as similar to each other as items matched in typicality in concrete categories. Abstract item pairs from different categories were more similar than concrete item pairs from different categories, indicating lower semantic distance of the categories. Taken together, these data suggest that taxonomic categories are less informative for abstract than concrete items. We discuss alternative factors in abstract concept organization.

Taxonomies for Abstract Items

When asked to sort the items *apple*, *cabbage*, *squirrel* and *duck* into stacks, most people will presumably sort them into types of *produce* and *animals*. Concrete item categories have a strong, graded structure organized around prototypical items, and are relatively distinct from other categories. As a consequence, concrete item categories offer salient dimensions that are readily used in reasoning involving their members. Because of their family resemblance structure, in which items share groups of features with other category members, two typical items from the same category, such as squirrel and duck, are likely rated as similar. Categories are important organizational structures that enable us to use our knowledge in reasoning processes such as classification, similarity judgments, and to effectively acquire new knowledge through analogies and inferences. As such, categories touch upon and inform many areas of research that are central to cognitive science.

Categories for concrete things are only a subset of our knowledge. We also have representations of abstract and complex entities, such as processes, events, mental experiences, stories, and relations. Abstract concepts constitute a large and important part of our daily experiences and actions. Even so, very little is known about categories of abstract concepts. A complete understanding of the organization of our knowledge cannot be achieved unless we understand how other kinds of things are organized.

There are several reasons to expect that taxonomic categories for abstract items are not as distinct and salient as those for concrete items. First, some research has shown that abstract items (in this case, verbs) are not organized into distinct clusters, but that classes of such items overlap with others on many dimensions (Huttenlocher & Lui, 1979; Miller & Johnson-Laird, 1976). The features of concrete items are highly correlated; that is, given that two items share a feature, they typically share some other features as well, making them more similar to each other than to items in different categories. As a consequence, categories are more distinct from one another. In contrast, a verb may share features with verbs from many different categories to a similar extent. Hampton (1981) has made a similar proposal for abstract concepts in general. Thus, categories of relational concepts may overlap more and offer less constraint on their processing.

Second, taxonomic similarities appear to have little impact on similarity judgments of abstract concept pairs. When people rate the similarity of item pairs that are taxonomically similar and that also share a thematic relation, people tend to focus on the thematic relation for abstract items, but on the taxonomic relation for concrete items. For example, participants often rate *jealousy* and *anger* as similar because jealousy *may lead to* anger. Much less frequently, participants refer to the information that both are emotional states (Wiemer-Hastings & Xu, 2003). In contrast, given an item pair like *cat* and *mouse*, participants will more frequently refer to both being animals than to one *chasing* the other (Wiemer-Hastings & Xu, 2003; Wisniewski & Bassok, 1999). Participants use thematic relations as explanations for concrete item similarity judgments mostly when no taxonomic similarity is present; for abstract item pairs, thematic relations are used frequently also when items are taxonomically related.

One possible explanation for this striking effect is that taxonomic similarity of abstract items, e.g., the similarity of two emotions, or two cognitive processes, is not salient information. This raises the question how useful taxonomic abstract item categories are. Abstract categories may be useful for talking about groups of abstract items, such as events, but they may not reflect actual knowledge organization in memory. For example, the concepts joy, sadness, wedding and farewell could be organized into emotion terms and events, or into concepts of positive versus negative connotation, respectively. There is no salient dimension here along which to unambiguously split

the four words into two groups, and the relations between joy and wedding, or sadness and farewell, are quite salient. If taxonomic classes of abstract items turn out to be of little functional use, the next question would be whether abstract items are organized around some alternative information. The results discussed above suggest thematic relations as one possible source for abstract category organization. Recently, it has been found that when given a choice, individuals routinely sort even concrete items around thematic relations (Murphy, 2001; Lin & Murphy, 2001). For example, when presented with the four items squirrel, mouse, nut, and cheese, people may sort them into pairs of animal and type of food (squirrel – nut; mouse – cheese) instead of animals and foods. Since such thematic relations have already been shown to play a very prominent role in similarity judgments of abstract item pairs, chances are good that they will provide dominant information to their categorization as well.

The main concern of this paper is the question why people do not effectively use taxonomic abstract item categories, such as emotions, cognitive processes, actions, attitudes, attributes, and so on in category-related tasks such as similarity ratings. The main hypothesis was that abstract and concrete item categories differ in the amount of constraint that they place on membership. This was examined in two separate hypotheses. First, it was hypothesized that participants would generate fewer typical examples for abstract than for concrete item categories. That is, we expected that fewer abstract items would be listed by a large number of participants. Related to this issue, we also hypothesized that abstract item categories would be less distinct than concrete item categories. That is, we predicted that members of two abstract item categories would be almost as similar to each other as members of the same abstract item category. Abstract items from the same category may share comparatively fewer category-specific features, and may share relatively more features with members of other categories. Three experiments tested these hypotheses.

Experiment 1

Experiment 1 compared the numbers of types and tokens generated for abstract versus concrete item categories. In accordance with the first hypothesis, it was predicted that significantly more participants would list the same items for concrete categories than for abstract ones.

Method

Twenty participants generated exemplars for 24 commonly used categories, 12 for concrete, and 12 for abstract items. Example categories are *tools*, *pets*, *object attributes*, and *positive emotions*. Category lists were constructed from taxonomic trees and extended to include a variety of categories. Since there is no well-established taxonomic model for abstract items, abstract item categories were taken from ontologies and social categories. Each participant listed exemplars for all categories, to control for individual differences. Categories were listed in random order. There was no time limit. Participants were instructed to list as many exemplars for each category as they could think of.

Results & Discussion

For each category, types and tokens were calculated. Type scores count different exemplars, whereas tokens also count repeated mentions of exemplars. The ratio of both indicates the agreement among participants or the mean production frequency for each exemplar. A highly available exemplar should be mentioned by many participants, resulting in a higher production frequency or token / type ratio. Categories that place strong constraints on cognitive processing would be characterized by higher token / type ratios.

Abstract and concrete categories did not differ in the number of types that were generated. However, more idiosyncratic responses were generated for abstract item categories. Consistent with the prediction, more participants listed the same exemplars for concrete item categories (see Table 1), t (10)=7.93, p<0.001. For the vast majority of abstract item categories, individual exemplars were listed by fewer than 2 participants, whereas on average, at least 3 participants listed exemplars for concrete item categories.

Table 1: Types and Token / Type Ratios from Experiment 1

	Types	Tokens / Type
Concrete	37.75	3.41
Abstract	35.42	1.57

Token / type ratios for concrete item categories varied from 2.25 (*foods*) to 4.78 (*pets*), *SD*=0.79. In contrast, very few participants listed the same exemplars for abstract item categories. The token / type ratio varied only very little across the different abstract item categories (*SD*=0.43), with the token /type ratios ranging from 1.06 (*prosocial actions*) to 1.82 (*social offenses*). The only exception to this dichotomy was one abstract item category, *object attributes*, which is actually relatively concrete, and which had a ratio of 2.71. Without this category, the standard deviation of ratios for abstract categories was reduced to *SD*=0.26.

Overall, it seems that low agreement is a general characteristic of abstract item categories, which could indicate that these categories do not reflect actual organization in memory. Instead, participants may retrieve exemplars out of a different organization, leading to high response variation. A possible confound that may account for the observed differences was that abstract item categories tend to be broader categories than concrete item categories. Experiment 2 varied the broadness of the categories systematically to test this.

Experiment 2

Experiment 2 replicated the procedure used in Experiment 1 with categories of different specificity levels. One set of categories was at a broad, abstract level, another set was more specific. Generally, specific categories contain fewer members and should thus place stronger constraints on exemplar production. The expectation was, accordingly,

that agreement would be higher for specific categories. The categories used in this Experiment are shown in Table 2.

Table 2: Categories Used in Experiment 2

	Broad	Specific
Abstract	Actions	Reasoning
	Mental Processes	Mental Disorders
	Communication	Object attributes
	Events	Character Traits
	Attitudes	Offensive Actions
Concrete	Animals	Wild animals
	Plants	Pets
	Foods	Tools
	Liquids	Office Supplies
	Natural Substances	Beverages

We also selected abstract and concrete item categories that were based on situations. Three such categories were based on *scenes* or *settings*, and three were based on *events* or *scripts*. Both have been shown to be efficient schemata for organizing knowledge in a systematic way (Shank & Abelson, 1977; Tversky & Hemenway, 1983). It has been argued that abstract items are characterized by situational contexts, rather than by internal features (Barsalou & Wiemer-Hastings, in press; Wiemer-Hastings & Graesser, 2000). Abstract concepts typically involve agents, goals, relations, actions and events, and emotional or cognitive experiences. As such, they are akin to abstract situation schemata, and may be organized around situations. Thus, we predicted that abstract, but not concrete, items generated for a given situation, are perceived as similar to each other.

Method

Sixty undergraduate students at Northern Illinois University participated in this experiment. Twenty participants each generated exemplars for broad categories, more specific categories and for situation categories that were either a setting (e.g., *workplace*) or an event (e.g., *wedding*). For the situation-based categories, abstract item instructions asked for "actions, events, or mental processes that could occur in the situation"; concrete item instructions asked for "objects occurring in the situation". The settings and events were identical for both groups to allow for direct comparison. Altogether, there were ten broad and ten specific categories (five abstract, five concrete), and six situation-based categories.

Results & Discussion

As in the first experiment, we evaluated the results through token /type ratios. The data replicated the findings from Experiment 1. Table 3 shows the results. Interestingly, the use of more specific categories (e.g., *offensive action* instead of the broad *action*) did not improve agreement scores, even though it had an effect on the number of types listed.

The token / type ratio for broad category exemplars did not statistically differ between concrete and abstract items, suggesting that broad superordinate categories of concrete items structurally resemble abstract item categories. A significant difference was found for specific categories, where the token / type ratios were significantly higher for concrete than for abstract items ($t(8)=3.76$, $p<0.01$) and also higher than for concrete items listed for broad categories ($t(8)=3.10$, $p<0.05$). That is, switching to a more specific category level increased category constraint for concrete, but not for abstract items.

The data suggest that the abstractness level does not have much of an impact on how many typical exemplars are produced for abstract categories. It seems, then, that the differences observed for concrete versus abstract items reflect a more general difference: Abstract item categories do not have many typical exemplars, and the few typical exemplars have relatively low typicality scores, as measured by the number of participants naming each. These first results suggest that categories may not provide the strong basis for inferences and similarity judgments for abstract items that is usually seen for concrete items.

Table 3: Types and Token / Type Ratios from Experiment 2

		Types	Token / Type
Broad	Abstract	42.2	1.63
	Concrete	34.8	1.78
Specific	Abstract	30.6	1.49
	Concrete	28.6	3.25
Situation	Abstract	36.3	1.60
	Concrete	57.7	2.07

Surprisingly, the pattern did not change much for situation categories. Participants generated substantially more items for concrete situation categories than for other categories, suggesting that this category type places low constraint on concrete items. This is not the case for abstract items. Still, the token / type ratio was significantly higher for concrete than for abstract items, $t(10)=2.72$, $p<0.05$. That is, there seems to be more consensus on what concrete items occur in situations than what abstract processes, events or states may.

Experiment 3

Previous experiments suggested that taxonomic category membership plays a minor role in similarity judgments for abstract items (Wiemer-Hastings & Xu, 2003). We suspect that this is due to overlapping categories, such that members of the same category may be only slightly more similar than members of different categories. Accordingly we predicted that, overall, item pairs should be more similar for items from the same category, but that this effect should be more pronounced for concrete item pairs than for abstract ones. That is, we expected that the difference in similarity for same- vs. different-category abstract item pairs would be significantly smaller than for concrete item pairs.

Alternatively, item pairs used in previous experiments were very untypical exemplars of their respective categories, which could have lowered the salience of their category membership. In the present experiment, we

control for this by collecting similarity ratings only for the most typical items for abstract item categories. If membership in the same taxonomic category does not substantially increase similarity ratings, compared to ratings of members of different categories, then typicality is unlikely the reason for people's neglect of taxonomic similarity in similarity judgments.

With respect to the situation-based categories, we had no strong predictions. However, we suspected that concrete items may be less constrained by such categories than by taxonomic categories, thus that situation-based abstract item categories may actually be more distinct.

Method

We used high-frequent exemplars collected in Experiment 2 for different category types to make sure they would be typical items for the categories. There were two constraints: first, we omitted some items that had been generated for multiple categories (with the exception of *happiness* and *happy* for one set for lack of alternative choices). Second, the exemplars selected from concrete versus abstract item categories were matched in typicality, for fair comparison. This means that the concrete item pairs used in Experiment 3 were not the most typical of their categories, but instead were matched in typicality (as measured by generation frequency in Exp. 2) to the abstract item pairs. Accordingly, the actual differences between these categories are likely strongly underestimated in this Experiment.

Fifty-seven undergraduates from Northern Illinois University participated in this experiment for course credit. Four participant groups were formed at random to judge the similarity of items generated for broad, specific, abstract situational, or concrete situational categories, respectively. For the broad and specific categories, participants were presented with three items from each of five categories, resulting in 15 items presented in random order. Abstract and concrete items were rated by the same participants, but were blocked to avoid contrast effects on the ratings. For the situation-based categories, there were six categories for abstract and concrete items each. Since 18 items had to be rated in pair-wise combinations, concrete and abstract items were presented between-subjects to avoid fatigue effects.

Each item was rated against each other item in the context of the entire list, to allow for category information to affect the ratings. Thus, if category information was accessible from the items, we expected that items associated with the same categories would be rated as more similar to each other than to items not belonging to the category. The order of ratings was varied so that item pairs were presented in ascending vs. descending order, and so that each item was presented equally often in first vs. second position of the item pairs. Ratings were made on a 6-point scale.

Results & Discussion

We predicted that the difference in similarity for abstract item pairs within and between categories would be significantly smaller than for concrete item pairs. A mixed ANOVA tested the effects of within-subject variables abstractness (concrete vs. abstract) and category membership (same vs. different), and between-subjects variable category level (broad vs. specific) on similarity ratings. The mean similarity scores for these variables are shown in Tables 4 for broad and specific categories.

Table 4: Ratings from Experiment 3

	Broad		Specific	
Category	Abstract	Concrete	Abstract	Concrete
Same	3.63	3.88	2.94	3.84
Different	2.74	1.98	2.00	1.45
Difference	0.89*	1.90*	0.94*	2.39*

Category Distinctness of Abstract Items Similarity was rated significantly higher for items of the same categories than those of different categories across all item groups, $F(1, 25)=154.88$, $MSE=0.21$, $p<0.001$. This was a large effect, $\varepsilon^2=0.86$. Further, consistent with our prediction, an interaction of category membership with concreteness was revealed, $F(1, 25)=46.68$, $MSE=.21$, $p<0.01$. This effect was moderately high, $\varepsilon^2=0.65$. As can be gathered from the difference scores in Table 4, concrete item categories had a larger semantic distance overall (difference $M=2.14$) than abstract item categories ($M=0.92$). Similarity ratings among abstract items were significantly higher for same-category pairs than between-category pairs, but, consistent with our predictions, this difference was significantly smaller than for concrete item categories.

Category Type There was a small significant effect of category level, $F(1, 25)=4.77$, $MSE=1.35$, $p<0.05$; $\varepsilon^2=0.16$. Overall, as we expected, specific categories produced greater differences in same-versus different category pair similarity. Table 4 shows that this effect is almost absent for abstract item categories (the interaction of category membership and concreteness approached significance, $p=0.09$). So, specific categories of abstract items are not more distinct than broad categories – their members are quite similar to members of other specific categories. This suggests that low category distinctness is a general problem for abstract items categories that spans different category levels. This finding is consistent with the results from Experiment 2, which showed that category specificity had no impact on the production frequency of the exemplars.

Situations and Abstract Items Table 5 shows the mean similarity scores obtained for situation-based categories. The pattern of similarities is almost reverse from the one obtained for taxonomic categories: Abstract item pairs are more similar for same-category pairs than concrete ones, and abstract item categories are more distinct.

In comparison to taxonomic categories, the most striking difference is the low similarity of concrete items of the same situation-based category. Abstract items that are typical for a given situation seem to be as highly similar to each other as members of the same taxonomic abstract item category. This is an important finding because it suggests that situations put semantic constraints on abstract, but not on concrete items. Accordingly, situation knowledge may

affect abstract item processing (such as similarity ratings) relatively more than concrete items.

Table 5: Ratings for Situation–Based Categories

Category	Abstract	Concrete
Same	3.44	2.28
Different	2.50	1.78
Difference	0.94*	0.50

In an ANOVA, the interaction of concreteness and category membership approached significance, p=0.07. Specifically, only similarity ratings of *abstract* items differed significantly for pairs of the same versus pairs of different categories. The semantic distance was the same for taxonomic and situation categories for abstract items. In contrast, there was a drop in distinctness for concrete situation-centered categories (M=0.50) as opposed to the taxonomic categories.

Conclusions

We have shown that abstract and concrete item categories differ systematically. The observed differences may explain some of the processing differences for abstract and concrete items. In particular, abstract item categories were found to have exemplars with low production frequencies, i.e. very few exemplars are named by two or more individuals. Experiment 2 showed that the token / type ratio for abstract item categories is at the same level as that for general concrete item categories. Thus, asking for exemplars of *mental processes* may be somewhat akin to asking for exemplars of *objects*. However, they differ from broad concrete item categories in at least two important respects: First, more specific subcategories of abstract items do not evoke exemplars with higher production frequencies, as they do for concrete items. Experiment 2 showed that in fact, the token / type ratio dropped slightly from broad (1.63) to specific (1.49) abstract item categories. At the same time, the number of different exemplars mentioned was lowered, suggesting that the more specific categories were indeed smaller categories.

Second, Experiment 3 shows that broad concrete item categories are more distinct than either broad or specific abstract item categories. Thus, we do not think that the difference between abstract and concrete item categories can be reduced to a difference in the abstraction level with abstract item categories being "super-superordinate" categories. An alternative explanation may be that abstract item categories are not organized around typical exemplars which, having most resemblance to the other exemplars, are recollected most often. Instead, it may be that participants have to actively construct the category as an ad hoc category. Memory for abstract concepts does not seem to be organized around taxonomic categories that can be recalled easily using a category label.

Further, the data suggest that abstract categories, regardless of category level, are less distinct from each other than concrete categories. Rated similarity is significantly higher for members of the same category versus members of different categories. However, the difference is much smaller than for concrete item categories (with the notable exception of situation-based categories). At this point, it is important to remember that the concrete exemplars used in the third Experiment were actually not the most typical exemplars of their categories by far, that is, the differences we observed in category distinctness probably substantially underestimate actual differences.

Considering these differences, it does not come as a surprise that taxonomic categories are used less as a basis for similarity judgments of abstract than of concrete item pairs. We also predict based on these data that there would be little agreement in sorting tasks using abstract items since there does not seem to be a strongly organized categorical structure for abstract items.

The data raise the question how meaningful abstract item categories are. Are they merely nominal in function, to enable us to talk about them at an abstract level? Or do they have any impact on the representation and processing of abstract concepts? Our data do not give conclusive answers to these questions but suggest that categorization of abstract items follows different principles and perhaps functionalities from that of concrete items. Referring to related studies, in what follows we will outline a few hypotheses. First, it is informative to link the present findings to studies that explore the content of abstract concepts in comparison to concrete concepts. Second, research on similarity processes, which are presumably involved in categorization, suggests that thematic relations may be an important source for structuring abstract concepts. It is possible that this is linked to the first issue – that the content of abstract concepts is more compatible with thematic than with taxonomic processing.

Conceptual Content

Hampton (1981) found that not all abstract concepts fit a prototypical structure and suggested that this may be due to lower feature correlations. Most likely, this is related to the content of abstract concepts. Hampton suggested on the basis of feature lists that their content consists of social components of situations, e.g., agents, behaviors, and goals. More recent findings are consistent this view. Analyses of conceptual content through property generation tasks suggest that abstract concepts have few "properties" in the classical sense (i.e., perceptual or functional features, parts) but instead have a high percentage of features that are related to situations and to subjective experiences in a situation (Barsalou & Wiemer-Hastings, in press; Wiemer-Hastings, Krug & Xu, 2001; Wiemer-Hastings & Xu, under review). Knowledge of situations is also linked to concrete concepts (e.g., part of our knowledge of chairs is that a person can sit on them to eat or to write), but concrete concepts are further distinguished through a high proportion of entity properties (i.e, external and internal parts, surface features, etc.), of which abstract concepts have very few or none.

Abstract concepts have two specific characteristics that are likely related to the observed lack of distinctive categories for abstract items. One is the smaller set of

features that are used to describe them, which may reflect that the concepts themselves have a lot more overlap than concrete concepts do. This could be directly linked to the high similarity scores of exemplars listed for different categories (Experiment 3). This may be accentuated even more by the other characteristic, namely that features of abstract concepts are significantly less specific than those of concrete concepts (Wiemer-Hastings & Xu). For example, a concrete item may be linked to a specific action (e.g., eating) while an abstract item may simply be linked to *some action*. The differentiation of categories requires a relative large feature basis to arrive at categories that are characterized by somewhat distinct clusters of features. General features are not likely to offer such differentiation.

Hierarchical category structures have been linked to correlated features. Concrete item properties such as parts and functions may be systematically correlated, e.g. because of causal links between their features (e.g., a bee *can fly* and *has wings* – it can fly *because* it has wings). In contrast, situation properties can be more flexibly linked to form a seemingly unlimited variety of abstract concepts. An agent can be found in a large variety of situations and there are few constraints on their specific set-up. Thus, categories governed largely by situational features may all overlap since they would all involve the same kinds of features, to different extents. For example, emotions and cognitive processes would both involve a person, a situation that the emotion / cognition is related to, and an element of introspective experience. As a result, abstract item categories may have *more overlap* with *less distinction*.

Taxonomic versus Thematic Relations

One central purpose of categories is to allow for inferences about novel objects, based on their similarity to known objects. Likewise, categories in memory enable us to generate instances that could serve a particular function. Categories can be structured around taxonomic information, where several items are *a kind of X* (e.g., *gadgets, emotions*) or thematic information, where items fill complementary roles in a proposition. For example, they may be related via an instrument function (*knife, meat*), a causal relation (*cause, effect*) a temporal sequence (*question, answer*) or through a variety of other relations. Both taxonomic and thematic relations are part of our knowledge of abstract and concrete concepts, but there may be a functional advantage in the majority of situations to using one over the other when dealing with concrete or abstract concepts. In particular, as suggested by the findings discussed initially in this paper, taxonomic inferences may be more critical for novel concrete objects or reasoning involving objects, while thematic inferences may be more functional when processing abstract concepts. For example, to specific actions require objects with specific functional or perceptual features; identifying one is facilitated by categories organized around functional and perceptual features. In contrast, abstract item categories may have less practical relevance. For example, to understand an emotion, it may be more important to process events and traits leading up to it than to have quick access to other emotions.

Our data suggest that taxonomic relations of abstract items are comparatively weak. We suggest that abstract items may be organized quite differently from concrete items. For example, Experiments 2 and 3 show that situations provide as much structure for abstract items as taxonomic categories. While concrete items are used across different situations, abstract items that occur in a similar context are perceivably interrelated. We suspect that in the future, cognitive scientists may discover quite different structural principles for abstract item categories, and that they will centrally involve thematic relations.

References

Barsalou, L.W., & Wiemer-Hastings, K. (in press). Situating abstract concepts. In D. Pecher & R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought.* New York: Cambridge University Press.

Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory, 4,* 161-178.

Hampton, J.A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition, 9,* 149-156.

Huttenlocher, J., & Lui, F. (1979). The semantic organization of some simple nouns and verbs. *Journal of Verbal Learning & Verbal Behavior, 18,* 141-162.

Lin, E.L., & Murphy, G.L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General, 130,* 3-28.

Miller, G.A., & Johnson-Laird, P.N. (1976). *Language and perception.* Cambridge, MA: Cambridge University Press.

Murphy, G.L. (2001). Causes of taxonomic sorting by adults: A test of the thematic-to-taxonomic shift. *Psychonomic Bulletin & Review, 8,* 834-839.

Shank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15,* 121-149.

Wiemer-Hastings, K., & Graesser, A.C. (2000). Contextually representing abstract concepts with abstract structures. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 983-988), Mahwah, NJ: Lawrence Erlbaum Associates.

Wiemer-Hastings, K., Krug, J. K. D., & Xu, X. (2001). Imagery, context availability, contextual constraint and abstractness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp.1106-1111).

Wiemer-Hastings, K., & Xu, X. (under review). *Content differences for abstract and concrete concepts.*

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive Psychology, 39,* 208-238.

The Effect of Structure on Object-Location Memory

Carsten Winkelholz (winkelholz@fgan.de)
Christopher Schlick (schlick@fgan.de)
Mark Brütting (bruetting@fgan.de)

Research Establishment for Applied Science (FGAN)
Research Institute for Communication, Information Processing and Ergonomics
Neuenahrer Strasse 20
53343 Wachtberg, GERMANY

Abstract

This paper aims to study the effect of structure in a graphical layout on object-location memory. In two experiments several structures have been examined in respect to the performance of object-location retrieval. The results show that beside simple object-to-object spatial relations also the spatial relation of three objects is encoded in human spatial memory as a noisy distance-angular pair. Further the results show that noise in spatial memory is not symmetric, but seems to be distorted towards a higher accuracy to the horizontal directions.

Introduction

One aspect of human spatial memory is the usage of allocentric frames of references to encode and retrieve the location of an object. This aspect of human spatial memory implicates that the structure of a graphical layout might affect the performance of object-location encoding and retrieval.

Basically the study presented in this paper is motivated by some experiments performed recently in the community of information visualization. One experiment of Travanti & Lind (2001) investigated object location memory in hierarchical information structures across different instances of 2D and 3D displays. The results of their tests show, that the 3D display improves performance in the spatial memory task they designed. They were aware that their result does not prove their hypothesis that the natural appearance of the 3D display used in the test actually affected the improved performance. They speculated that possibly other visual properties of an item in the 3D display were used as a reminder for the memory task. Cockburn (2004) showed that neither the natural appearance nor the different sizes of the items in the 3D display affected the performance of object-location retrieval. In both studies the memory task was to associate alphanumerical letters to the items. Therefore Cockburn suspected that the vertical orientation of Travanti & Lind's 2D display made the formation of effective letter mnemonics more difficult than the horizontal 3D layout, because words and word combinations normally run horizontally left to right. By analyzing these studies we came to the conclusion that one major factor had not been considered - the factor of the object-to-object spatial relations (the structure of the graphical layout respectively).

The effect of layout structure on object-location encoding and retrieval could best be investigated if a computational model of human spatial cognition is considered. Recently some compelling works toward this goal has been published (Wang et al 2002; Johnson et al 2002). This paper shows one application area for computational models of human spatial memory, but also sheds some new light on the requirements of such a model.

Design of the Experiments

The papers cited above inspired the design of the experiments in this study. There were two phases in the cited experiments. In the encoding phase the subjects had to learn associations of alphanumerical letters to one object in the structure. During the encoding phase a click on one of the objects in the display highlighted the object and revealed a letter at the top of the display, which had to be associated to the position of the object. In the retrieval phase the subjects had to find all of the letters, one at a time. A randomly selected letter had been shown at the top of the display area, and the subject had to click the object associated with it.

This design of the experiment has two drawbacks. First the subjects are free to choose the objects in the encoding phase and second that alphanumerical letters are used as retrieval cues. The first point gives subjects the opportunity to develop strategies for learning the object-letter associations. In combination with the usage of alphanumerical letters this increases the probability that subjects create mnemonics through possible abbreviations of words that can be read from a row.

In respect to a cognitive model these are task specific aspects. The study of this paper was interested in more general mechanisms of object-location encoding/retrieval. To meet this goal the design of the experiment had to prevent subjects from further processing object-locations in the encoding phase. This suggested the task of memorizing a randomly created sequence of highlighted objects from the structure. The number of correct repeated sequences is used as a measure of performance. Furthermore allows this kind of memory task an effective analysis of the errors subjects make.

Two experiments were performed. The first experiment investigated the factor horizontal vs. vertical orientated

layout structure and the factor of the existence vs. non-existence of symmetric features in the layout structure. The second experiment focused on the investigation of noise in the encoding of spatial object-to-object relations.

Subjects and Apparatus

30 volunteer subjects (only male, average age 35) were recruited from the staff of our institute to perform both experiments. All subjects had normal or corrected-to-normal vision. Three sets of different structures have been created. Each structure consisted of red spheres of equal size. The layout structures were presented against a black background on a 21'' VGA monitor with a resolution of 1280x1024 pixels. The monitor was in front of the subjects within 2 feet. Subjects were asked to respond by clicking with a mouse. Subjects wore a head-mounted eye-tracking device while they were conducting the experiments.

Experiment 1

The first experiment aimed at showing if the performance of recalling objects is still improved in the horizontal oriented structures, even if in the experimental design no semantic content is used. Further one horizontal structure was added that contains not the symmetric features of the horizontal structure used by Travanti & Lind and Cockburn. Another purpose of this experiment was to show if there is any learning progress in the performance of object-locations encoding/retrieval. It might be possible, that subjects become more familiar with a structure the longer they are exposed to them. In combination with the factor of symmetric features in the structure it might be speculated, that in the presence of symmetric features a subject needs less time to become familiar with the structure.

Materials

Figure 1 shows the three structures that were used in the first experiment. The first two structures are similar to the structures used by Travanti & Lind. Each structure consists of 25 spherical items. The first structure represents a 2D display of a tree-structure, like it is used in most common graphical user interfaces. The second structure represents the structure of the 3D display, where any perspective clues have been removed. The third structure is equivalent to the first one except that it is rotated by 90^0 counterclockwise.

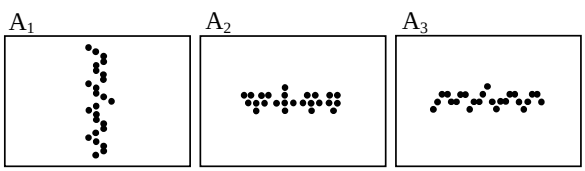


Figure 1: Set of structures used in experiment 1.

Design and Procedure

In each encoding retrieval trial, the subject was presented one structure. After an acoustical signal the computer started to highlight objects of one randomly created sequence. Only one object of the sequence was highlighted at once. The sequences were five items long. The highlighted object differed from the not highlighted objects by color (blue instead of red), increased size and a cross that appeared within its circle shape. The end of a sequence was indicated by an second acoustical signal. Subjects were instructed to repeat the highlighted objects in correct order, by clicking them with the mouse. After five objects had been clicked, another acoustical signal rang out and a short online questionnaire with a subjective rating occurred. Subjects had to rate how confidant they were about their answer and the degree of difficulty to memorize the sequence. The questionnaire was inserted between the tests of two sequences to reduce stress by diversion. Each subject was tested on all structures. The experiments consisted of three blocks. In each block the same structure was tested four times in succession. Between each block there was a break of one minute. Subjects were randomly divided into six groups with five persons, where in each group the order of the three blocks belongs to one of the six possible permutations.

Before the main experiment started, each subject passed through a training run, consisting of two blocks of four sequences. The structures presented in the training run consisted of 16 objects randomly located on the display. The length of the sequences subjects had to learn varied between four and six items.

All sequences for the training run and for the actual test were created randomly only with the property that not the same item occurred in the sequence one behind another. For each subject new random sequences were created. This was done to avoid that for one structure an easy sequence would have been created by chance (e.g. the items of a sequence are only in one row). In general for each structure there might be sequences that are easy to learn, but for some structures these are more likely than for others. And clearly this is a property of a structure that one likes to deduce from its spatial layout. To fix the sequence across subjects would mean that two different factors are controlled. Creating random sequences for each subject means to balance the factor of the sequence among subjects. To fix a sequence across subjects would be interesting to study one specific factor in detail. This was done in parts in the second experiment that is reported in the next section.

Results and Discussion

Accuracy data The number of correct and incorrect repeated sequences for each structure is shown in Table 1.

Table 1: Contingency table (2x3) of correct and erroneous sequences

Structure	A_1	A_2	A_3
Correct seqs.	46	61	63
Erroneous seqs.	74	59	57

The effect of structure approaches significance (2x3 contingency table $p = 0.056$, $\chi^2 = 5.77$). When comparing the numbers of correct repeated sequences between each pair of structures with a one-sided analysis of the corresponding 2×2 contingency tables, the exact Fisher test yields that performance in the horizontal oriented structures are significantly higher ($p<0.05$), whereas the symmetric features in the structure did not show any significant effect.

Learning progress Figure 2 shows the development of the performance by each trial in the same structure.

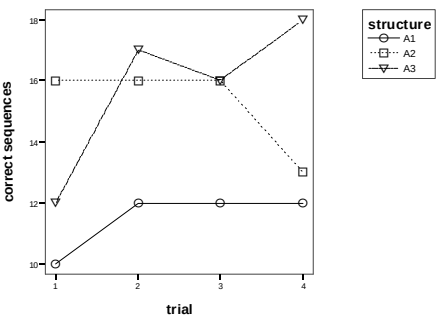


Figure 2: Development of performance in dependence of number of trials.

Structure A_1 with no symmetric features exhibits an increasing performance with each trial, whereas the horizontal oriented structure with symmetric features even shows a decrease in the last trial. The effect of trials on performance is not significant in any structure (2x4 contingency table $\chi 2$ statistical test). Hence, this effect may result from noise in the data.

The most important result of experiment 1 is that it shows that the horizontal oriented structures do improve performance, even if no alphanumerical letters are used as retrieval keys. However, this result may be culture dependent. For example people, who are used to read in columns instead of rows, might be more familiar with horizontal oriented structures.

One culture independent reason for this result might be that the human field of view is more extended into the horizontal direction. This increases noise of allocentric and egocentric memory chunks in vertical directions. If this hypothesis was right, people used to read in columns would profit from a horizontal oriented tree view in two ways: First the horizontal structure would increase performance of object-location retrieval and secondly, inscriptions could be written in columns instead of rows.

Experiment 2

The second experiment aimed at showing how the aspects of human spatial memory, like they are discussed in Wang et al (2001, 2002), affect the performance of object-location encoding/retrieval in dependence on different graphical layout structures. Another purpose of experiment 2 was to collect eye movement data for a more detailed analysis of how subjects encode object-locations.

Materials

The structures used in the second experiment are shown in Figure 3. They are divided into two subsets, because the limited pool of subjects didn't allow testing all permutations needed to prevent order effects.

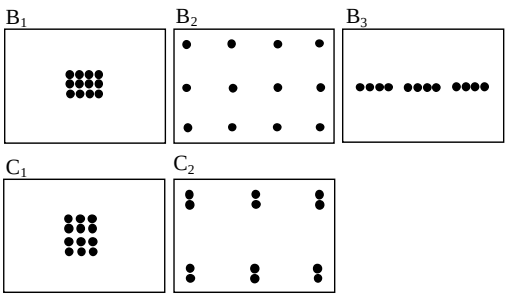


Figure 3: Set of structures used in experiment 2.

Justification The structures in set B and C were created to test some factors assumed to play an important role in the process of object-location encoding/retrieval in structures. The motivation to choose these structures is founded in the assumptions and expectations before the experiments were performed. Mainly the following factors were expected to contribute to the overall performance:

1. Hierarchical features.
2. Noise in the location of an allocentric memory chunk.
3. Noise in the location of an egocentric memory chunk
4. Higher activation of allocentric memory chunks if objects are in spatial vicinity.

The last factor seems plausible, because the effort to assess spatial object-to-object relations is smaller if objects are close together; possibly no eye movement is needed. This last factor would give the spatial narrow matrix B_1 an advantage over the spatial wide matrix B_2 in respect to performance of object-location encoding/retrieval. But also the other factors listed above may contribute. In the linear structure the noise in the memory chunks are more grievous than in the matrices, because there is only one dimension that contributes information, whereas in the case of the matrices also the direction contributes. The tables below show which structure profits by which factor compared to another structure in its set. A + sign in one cell means, that the structure of the row takes an advantage over the structure in the column in respect to the factor of the table, whereas a – sign indicates the opposite The factor of hierarchical features is balanced within each set, so this factor is not included in the tables. (For this purpose the linear structure has been separated into three groups with four objects).

Table 2: Which structure profits by which factor in set B.

Less noise in allocentric memory chunks

	B_1	B_2	B_3
B_1		0	++
B_2	0		++
B_3	--	--	

Less Noise in egocentric memory chunks

	B_1	B_2	B_3
B_1		--	-
B_2	++		+
B_3	+	-	

Higher activation of allocentric memory chunks

	B_1	B_2	B_3
B_1		++	+
B_2	--		-
B_3	-	+	

Table 3: Which structure profits by which factor in set C.

Less noise in allocentric memory chunks

	C_1	C_2
C_1		+
C_2	-	

Less noise in egocentric memory chunks

	C_1	C_2
C_1		-
C_2	+	

Higher activation caused by spatial vicinity

	C_1	C_2
C_1		+
C_2	-	

To estimate the overall performance, the tendencies shown in the tables have to be quantified. Furthermore, not any factor might contribute equally to the overall performance. Without any computational model it can only be speculated about these questions. However, in the setup used in the experiment, it can be assumed that the differences in the noise of the egocentric memory chunks are nearly negligible, because the changes in the average visual angles between the different objects in the scene are small compared to the human field of view. Whereas the directional angular of the allocentric memory chunks possibly covers the whole range. The effect of noise in the allocentric memory chunks in the structure B_1 and B_2 are expected to have an equal effect, because all relative distances are equal. It was expected, that the effect of decrease in performance in the linear structure would be very distinct.

Structure C_1 and C_2 differ only by the distances between the six pairs of objects; the distances between the two objects within a pair are equal. The hypothesis for this structure is that for sequence containing transitions between objects of two far distant pairs it will become more difficult for the subject to encode the location of the object within a pair. This results from a higher noise in the spatial object-to-object relation. To show this effect one predefined sequence was used. This allows analyzing behavior of subjects more efficient. Data from experiments can be used for the parameterization of stochastic models. The regularities found by the algorithms can be analyzed and interpreted (Winkelholz et al., 2003).

Design and Procedure

The experimental design was similar to experiment 1. This time the sequences were six items long. Furthermore, the experiment consisted of two blocks instead of three and in one block each structure from each set was presented once. The first three structures in each block were chosen from set B ordered by one of the possible six permutations. The last two structures in each block were C_1 and C_2, which order again was balanced within groups of subjects.

Except for one sequence in each block all sequences were created randomly for each subject. One sequence for the structures of set C was predefined. Like mentioned above, this was done to be able to analyze experimental tracing data effectively. The sequence was predefined for the structures C_1 and C_2 respectively. The predefined sequence is shown in Figure 4 on the left. It was used in the first block for structure C_1 and in the second block for C_2 or vice versa. By alternating, which structure in the first block starts with the predefined sequence, the effect of remembering the sequence in the second block had been balanced between the structures C_1 and C_2.

Results and Discussion

Accuracy data The numbers of correct repeated sequences are shown in the contingency tables 4 and 5.

Table 4: Contingency table (2x3) of correct and erroneous sequences in set B.

	B_1	B_2	B_3
Correct seqs.	38	34	16
Erroneous seqs.	22	26	44

Table 5: Contingency table (2x2) of correct and erroneous sequences in set C.

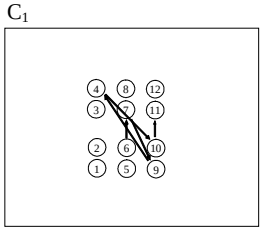
	C_1	C_2
Correct seqs.	35	25
Erroneous seqs.	25	35

The performance in the linear structure is significantly lower than in the structures of the matrices (exact Fisher-test $p<0.001$). Although the number of correct sequences in the narrow structure is a little bit higher than in the wide matrix, this difference is not significant. In table 5 the number of correct and incorrect sequences from the randomly created sequences and the predefined sequence are combined.

Analysis of errors A look at the errors subjects made in their answers gives more insight into the underlying cognitive processes. To analyze the answer sequences for the predefined sequence in set C we used a modified algorithm for variable length markov chains (VLMC) (Ron et al 1996, Bühlmann & Wyner 1998) to parameterize a stochastic model by the answer sequences. Roughly speaking this algorithm can be seen as a filter for subsequences (called contexts) from the data that contain predictive information. We modified this algorithm in a way that only contexts that contain significant predictive information in a statistical sense are included into the model (Winkelholz et al 2004). To apply this algorithm to the answer sequences the objects in the structure has to be assigned to symbols. The contexts of erroneous behavior found by this method in the answer sequences of the structures C_1 and C_2 are shown in Figure 4. In the first column of the table the contexts found by the algorithm are

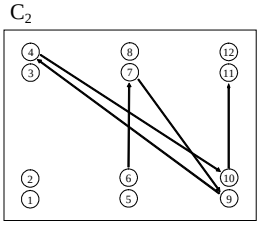
shown in parenthesis, followed by an arrow, and the most probable next symbols that occur in the answer sequences, if this context is given. E.g. "(7,10)->3", means: If subjects had clicked on object 7 followed by object 10, the most probable object they will click next is object 3. If on the right side of an arrow, more than one symbol/number is listed, they are ordered by their probabilities, with the most probable next symbol first. On the right of an arrow possible next symbols are listed, as far as their frequencies for the given context meet one of the two conditions: First, the frequency is significantly higher than for the symbols with lower frequencies. Second, the frequency does not differ significantly from the frequency of the symbol with the next higher probability.

C_1



Context -> Next Symbol	Interpretation
(4)->10, 11	Omission of object 10.
(6)->7,9	Omission of object 7.
(10) ->11, e	Surprised by the end.
(b)->6,7 (7,6) -> 9 (b, 7) -> 6	Changed order of the first two events in the sequence
(4,9)->11 (9)->4,11	The spatial vicinity, caused subject to mistake 9 for 10.

C_2



Context -> Next Symbol	Interpretation
(9)-> 4,3 (3)->10	The spatial vicinity, caused subject to mistake 3 for 4.
(7) -> 9,10	The spatial vicinity, caused subject to mistake 9 for 10.
(7,10) -> 3	If subject mistake 10 for 9 he also mistake 3 for 4.
(4)->10,11 (4,11)->e	Omission of object 10.

Figure 4: Contexts of erroneous behavior found by the parameterization of a stochastic model. Left: The structure with symbols assigned to the objects and the predefined test sequence. Right: Table with contexts and possible interpretation.

In structure C_2 with the more distant pairs there are more contexts concerning with the confusion of the objects within the pairs of the upper left, and down right corners, whereas for structure C_1 there are more contexts concerning the omission of an object. The most notable context for structure C_2 is "(7,10)->3". The angular between the line from 7 to 10 and the line from 10 to 3 is nearly similar to the angular between the lines 7 to 9 and 9 to 4. Therefore this context indicates that subjects used the relative change in angular direction of two transitions as a reminder.

Eye movement data Currently only the eye movement data of the structures C2 and B2 have been analyzed. Only these two structures exhibit spatial distinct features that allow a reliable assignment of fixations to attended features in the structure. In the structure C2 the fixations were only assigned to one pair. The resolution of the eye tracking device was not sufficient to distinguish between fixations within each pair. For the analysis of the eye movement data in the encoding phase of the predefined test sequence the same method as in the analysis of the errors in the answer sequences was used. The pictures obtained from this procedure are shown in Figure 5. Each picture shows the transitions in the eye movement between the pairs of objects, when the object shown as a filled circle is highlighted. The most probable pairs of objects that will be fixated next if one fixation and the highlighted object is given are presented as arrows starting at the currently fixated pair of objects.

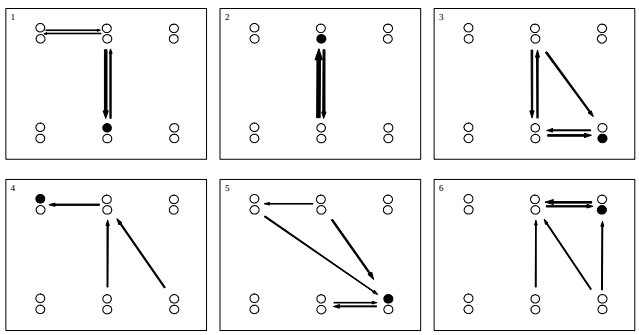


Figure 5: Eye movement data during the encoding phase for the predefined sequence (see figure 4).

The sizes of the arrows are scaled by the frequency of this transition. Although the predefined test sequence does contain two transitions that connect the objects from the upper left corner to the down right corner, there is only one transition in eye movement that connects these pairs directly. Even in the case of a transition from the down right to upper left corner in the test sequence subjects first fixated the group more near to the currently fixated pair of objects (picture 3-4). It was expected that after these transitions in the test sequence occurred, subjects would tend to repeat these transitions by eye movement to create memory chunks for this spatial relation. Instead subjects seem to create spatial relations to the pairs in the middle column. This result becomes more affirmed by taking a look at the eye movement data of the randomized sequences of the structures C_2 and B_2. An overlay of the transitions in the randomized test sequences and the corresponding transitions in the eye movement data are shown in Figure 6.

Although the transitions in the test sequences contain equally transitions between distant objects, these transitions are merely absent in the eye movement data. In both structures most transitions in eye movement are transitions between locations in the vicinity of the two objects in the middle of the screen. In the case of the matrix, movements of the fixation toward objects at the border are very sparse, whereas in the structure with the pairs of objects there are noticeable more fixation movements toward each pair of objects. This also explains the not expected result**,** that there is no significant difference in the performance of the wide and the narrow matrix structure. Possibly, it is sufficient to fixate a location in the middle of the screen to asses most of

the spatial object-to-object relations. Moving attention in the visual buffer to repeat transitions is possible without moving fixation. Therefore the effort to repeat transitions of the test sequences in structure B_1 and B_2 are similar. Different in structure C_2; here subjects needed to move fixation to resolve which object within a pair had been highlighted.

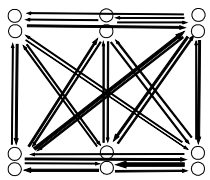
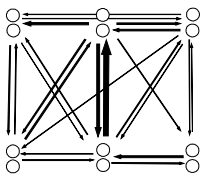
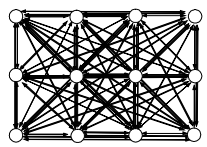
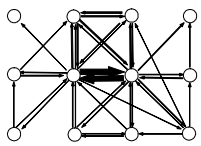
Overlay of all transitions in the randomized test sequences	Overlay of all transitions in eye movements.
	
	

Figure 6: Comparison of transitions in the randomized test sequences with transitions in the eye movement data.

Conclusions and Future Work

The experiments reported in this paper showed how single aspects of human spatial memory affect the overall performance in memorizing tasks of object-locations in layout structures. A computational model that quantifies the interaction of the different aspects of object-location memory is needed to get reliable predictions about the overall performance. The development of such a model within a general architecture of cognition like ACT-R, (Anderson & Lebiere, 1998) enables the implementation of meaningful cognitive models for the application field of information visualization.

The results of the two experiments make the following suggestions with regard to a computational model within the ACT-R architecture:

First, like Wang et al (2002) suggested the model should encode spatial object-to-object relations between the previously and currently attended objects as memory chunks.

Second, also the relation between three objects should be encoded in a memory chunk. In the same fashion as object-to-object relations are encoded this can be done by the visual module whenever attention shifts between three different objects. This memory chunk should be of the form of a noisy angular. Thus the model would show the systematic failures found in the analysis of the answer sequences.

Third, the results from the comparison of the horizontal and vertical oriented structures in the first experiment suggest that noise in the memory chunks of spatial memory is distorted towards a higher accuracy in the horizontal direction. This is a plausible assumption, because the human vision field of view is more extended into the horizontal direction and this should be true for coordinates in all frames of references.

Fourth, eye movement data showed, that subjects need not to gaze at objects they are attending to assess their locations in different frames of references. Therefore it may be disputed, if developers of cognitive models within ACT-R need to control fixation and attention independently. The noise in the assessed object locations should depend on the distance to the current location of fixation.

References

Anderson J. R.; Matessa M. & Lebiere Christian (1998): *The Atomic Components of Thought*. Lawrance Erlbaum Association, Inc. Publishers

Bühlmann, P. & Wyner A.J. (1999). Varaible Length Markov Chains. Annals of Statistics 27 (1), pp. 480-513, 1999.

Cockburn A (2004). Revisiting 2D vs 3D Implications on Spatial Memory. *Proceedings of the Fifth Australasian User Interface Conference* (AUIC2004). Dunedin, New Zealand. January 2004, pages 25-32.

Johnson, T. R., Wang, H., Zhang, J., & Wang, Y. (2002). A Model of Spatio-Temporal Coding of Memory for Multidimensional Stimuli. In W. Gray & C. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 506-511). Mahweh, NJ: Lawrence Erlbaum Associates.

Ron, D., Singer, Y. & Tishby, N. (1996): The Power of Amnesia: Learning Probabilistic Automata with Variable Length. *Machine Learning* 25, 2/3, 1996, pp.117-149.

Travanti , M. & Lind M. (2001). 2D vs 3D, Implications on Spatial Memory. In proceedings of the IEEE Symposium on Information Visualization 2001.

Wang, H., Johnson, T. R., Zhang, J., & Wang, Y. (2002). A study of object-location memory. In W. Gray & C. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 920-925). Mahweh, NJ: Lawrence Erlbaum Associates.

Wang, H., Johnson, T. R., & Zhang, J. (2001). The mind's views of space. *Proceedings of the 3rd International Conference of Cognitive Science* (pp. 191-198).

Winkelholz C., Schlick C. & Motz, F. (2003): Validating Cognitive Human Models for Multiple Target Search Tasks with Variable Length Markov Chains. SAE-Paper 2003-01-2219. In: *Proceedings of the 6th SAE Digital Human Modeling Conference*, June 2003, Montreal, Canada.

Winkelholz C.; Schlick Ch. (2004, submitted): *Statistical Variable Length Markov Chains for the Parameterization of Stocastic User-Models from Sparse Data*. 2004 IEEE International Conference on Systems, Man and Cybernetics

Can Experts Benefit from Information about a Layperson's Knowledge for Giving Adaptive Explanations?

Jörg Wittwer (wittwer@psychologie.uni-freiburg.de)
Matthias Nückles (nueckles@psychologie.uni-freiburg.de)
Alexander Renkl (renkl@psychologie.uni-freiburg.de)
University of Freiburg, Educational Psychology, Engelbergerstr. 41
79085 Freiburg, Germany

Abstract

E-consulting services such as asynchronous helpdesks for hardware and software are a common and comfortable way to get expert advice. However, the constraints of asynchronous communication and the experts' inclination to forget about the exclusiveness of their specialist knowledge may impair the advisory success. Against this background, an assessment tool has been developed which aids helpdesk experts in evaluating an inquirer's background knowledge. In a previous study, it could be demonstrated that the assessment tool increased the effectiveness and efficiency of asynchronous communication. In order to test the mechanisms that make the assessment tool effective, another dialogue experiment was conducted that varied the validity of the information displayed in the assessment tool. The results showed that the information presented to the experts did not only sensitize them for the inquirers' needs but also allowed for specific adaptation to their individual knowledge state. Hence, the validity of the information provided by the assessment tool is crucial.

Introduction

Inasmuch as knowledge becomes ever more specialized and complex, individuals often lack the expertise necessary for making a decision or solving a problem on their own (Nückles & Bromme, 2002). Thus, in many situations, laypersons are reliant on expert advice. The proliferation of the Internet offers new possibilities for laypersons to enlist the assistance of experts. Not only can laypersons retrieve expert information publicly available from the World Wide Web but they can also obtain personal advice from experts in a one-to-one fashion. Helpdesks for hardware and software are a prominent example of e-consulting services that enjoy increasing popularity (Moncarz, 2001). Virtually every large computer company and university computer centre offers helpdesk support, often in a text-based, asynchronous way via electronic mail. The aim of computer consulting is to convey knowledge, which enables the inquirers to solve their problem by themselves, for example, when new and complex software has to be learned or an unexpected technical problem with the computer suddenly occurs. The advisory success heavily depends on the experts' ability to provide intelligible and informative explanations for inquirers with differing levels of experience, ranging from very inexpert to more advanced users (Chin, 2000; Kiesler, Zdaniuk, Lundmark, & Kraut, 2000). Thus, in order to give effective and satisfactory advice, experts should adapt their communication to the knowledge prerequisites of the client (Clark & Murphy, 1982). Both from an educational (e.g., Renkl, 2002) and psycholinguistic perspective (e.g., Clark, 1996), adaptation to a communication partner's prior knowledge is regarded as fundamental for comprehension and learning.

Research on expertise has shown that experts, as compared to novices, possess an extensive and highly differentiated knowledge base that facilitates a rapid categorization of problem situations and the activation of routine problem solving strategies (Chi, Glaser, & Farr, 1988). However, these very characteristics of expert knowledge might interfere with the task of taking into account the limited domain knowledge of a layperson. Hinds (1999) called this phenomenon the 'curse of expertise'. She reported two experimental studies in which experts systematically underestimated the difficulties laypersons faced when performing a complex task. Alty and Coombs (1981) analyzed face-to-face advisory dialogues between computer experts and clients. They found that the computer experts rarely attempted to ascertain the clients' prior knowledge and rarely monitored the clients' comprehension of their explanations. As a result, the clients often did not understand the advice given. From these studies it can be concluded that in order to assure effective advice, experts should be supported in taking into account the knowledge prerequisites and comprehension of the client.

In face-to-face communication, the communication partners can use a variety of situational and interactional cues to monitor their interlocutor's comprehension moment by moment and thereby refine and update their mental model of what the other person knows or does not know (Clark, 1996; Nickerson, 1999). In Internet-based counselling, however, the evaluation of an interlocutor's knowledge and the continuous construction of a mutual understanding are considerably more difficult when compared with face-to-face communication (Clark & Brennan, 1991). First, in asynchronous communication, nonverbal feedback is virtually impossible because the interlocutors cannot see nor hear one another. Second, the costs of message production are higher than in verbal communication because every message has to be typed on a keyboard. Third, there is no set sequentiality between a message and its reply, because the interlocutors' turn taking may be interrupted by messages from third parties, which can impair comprehension (Clark & Brennan, 1991). Given these constraints, the pos-

sibilities to establish a mutual understanding are clearly more restricted as compared with face-to-face communication. On the other hand, asynchronous communication also offers affordances that can facilitate adaptation to a communication partner. It allows for a careful planning and revision of a message before it is sent. There is time to reflect about a communication partner's background knowledge and communicational needs.

The Assessment Tool - A Measure to Support Asynchronous Communication

From the preceding discussion it can be concluded that it would be useful to provide helpdesk experts with a support procedure that compensates for the constraints of asynchronous communication on the one hand, and takes advantage of the affordances on the other hand. When computer experts communicate with clients via an Internet-based helpdesk, they are in an anonymous communication situation with only little information available about the client. Therefore, the procedure should enable the expert to achieve a relatively concise and veridical evaluation of a client's knowledge state right from the start, because the lack of nonverbal feedback, the raised production costs and the limited sequentiality impede the continuous construction of a mutual understanding considerably. With regard to experts' inclination to forget about the exclusiveness of their knowledge, the procedure should encourage them to carefully reflect about a client's knowledge prerequisites in order to facilitate adaptation to the client's communicational needs. The better the computer experts' model of the client's knowledge is, the better the experts can adapt their explanations to the client's knowledge (Clark & Murphy, 1982).

In this paper, an assessment tool will be empirically tested that supports computer experts in constructing a mental model of the client's knowledge state in asynchronous communication (see also Nückles, Wittwer, & Renkl, 2003). The tool consists of a small Internet-based questionnaire by which users who place a technical support inquiry are asked to provide the expert with several self-assessments of their computer expertise (cf. Figure 1). For example, the clients are asked to rate their general level of computer knowledge as well as their knowledge of concrete specialist terms semantically relevant to the topic addressed by their inquiry. The assessment tool can be especially useful to the expert if it enables them to form a picture of the client's knowledge level based on a small number of highly relevant information items. The assessment tool provides the expert with information about the client right from the start, which normally can only be collected during the course of the interaction process. Consequently, it should facilitate the collaborative effort of communication (Clark, 1996). However, the assessment tool can only be effective if the medium of communication allows for careful planning and the revision of one's communicational contributions. Therefore, the assessment tool seems to be especially suitable for asynchronous, written communication because there is time for reflection and revision before a message is sent.

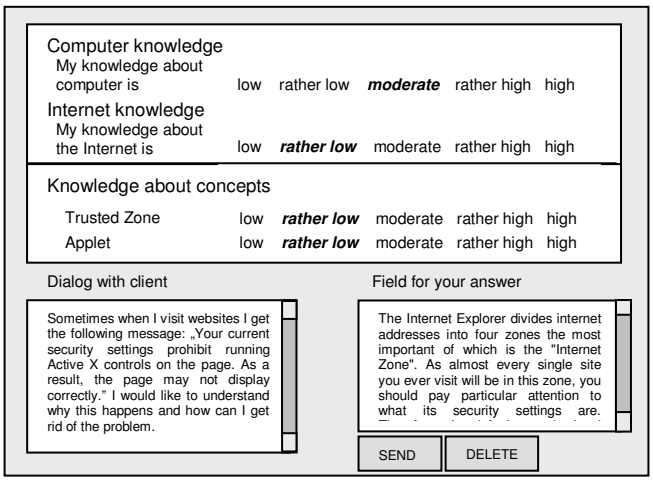


Figure 1: Screenshot of the assessment tool available to the computer expert.

The assessment tool has already been successfully tested in a web-based dialogue experiment between computer experts and clients (Nückles & Stürz, in press). With the assessment tool, the clients acquired significantly more knowledge than the control group without the assessment tool (increased communicative effectiveness). At the same time, they wrote back only half as often in response to the experts' explanations (increased communicative efficiency). Although the study demonstrated that the assessment tool approach was successful, it is unclear which mechanisms led to the increase in communicative effectiveness and efficiency. There are two main theoretical explanations that may account for these findings.

Theoretical Explanations of the Assessment Tool Effect

In Nickerson's theory (1999), the construction of a mental model of another person's knowledge is conceptualized as an anchoring and adjustment process (Tversky & Kahneman, 1974), where one's model of one's own knowledge serves as a default model of what a random other person knows. This default model is transformed, as individuating information is acquired, into models of specific other individuals. Accordingly, one could argue that the assessment tool presented individuating information about the client's knowledge level that provided the computer expert with a relatively specific anchor right from the start of the advisory dialogue. This enabled the expert to calibrate their mental model of the client's knowledge more quickly and accurately than would have been possible without the assessment tool, that is, only on the basis of the client's written questions and comments. According to this explanation, communicative effectiveness was raised because the assessment tool provided the expert with specific information that helped them to adapt to the client's individual knowledge level.

On the other hand, it may be argued that communicative effectiveness was raised *not* because of the *information* pre-

sented, but simply because the assessment tool increased the expert's awareness of the client and counteracted the tendency of de-individuation in Internet-based communication (Gunawardena, 1995). The experts were sensitized to reflect about the client's knowledge, for example which computer concepts are *typically* known by laypersons and which are not. This may have helped them to produce explanations that were more intelligible or informative for the *typical* layperson, irrespective of the *specific* knowledge level of an individual client. According to this explanation, the assessment tool had a more or less non-specific sensitizing effect on the expert. Against this background, the goal of the present experiment was to test whether the availability of *specific* information about the client's knowledge would make a difference at all, that is, support the experts' adaptation and thereby enhance communicative effectiveness and efficiency.

To this purpose, we modified the experimental design employed by Nückles and Stürz (in press). First, instead of using self-assessments, the assessment tool in this experiment provided the expert with objective information about the client's knowledge. Although self-assessments have proven to be good predictors of computer expertise (cf. Richter, Naumann, & Groeben, 2000; Vu, Hanley, Strybel, & Proctor, 2000), they still are not completely valid. Therefore, by using objective data about the client's computer knowledge, we increased the power for detecting a potential effect of specific adaptation. Secondly, a third experimental condition was included, in addition to a communication condition with the assessment tool and a condition without assessment tool. The information displayed in this additional condition was randomly drawn from the pool of knowledge data of clients who had previously participated in the experiment. The random data condition checked to see whether a distortion of the information about the client's knowledge level would impair the communication process. Consequently, the inclusion of this experimental condition would enable us to evaluate whether the *specific information* displayed by the assessment tool would influence the adaptivity of the experts' explanations.

Predictions

Sensitization hypothesis. If the assessment tool mainly had a sensitizing effect on the computer expert, that is, the information about the client was of little surplus value, it should make no difference whether the displayed information was valid or distorted. Accordingly, the mere presence of an assessment tool is supposed to increase the experts' awareness of the client and this alone should help them to improve their explanations. Consequently, in the conditions *with* the assessment tool the clients should acquire substantially more knowledge compared with clients in the condition *without* the assessment tool. Moreover, if the clients received explanations that were more intelligible and more informative compared with the condition without the assessment tool, they should experience less comprehension problems and should be more satisfied with the explanations. Hence, this should lessen their need of writing back in response to an expert's explanation. Consequently, the frequency of questions, and more specifically, the frequency of comprehension questions should be reduced in both conditions with the assessment tool.

Specific adaptation hypothesis. If the information provided by the assessment tool facilitates the adaptation to a specific client's knowledge, both the increase in communicative efficiency and effectiveness should be substantially larger in the condition presenting valid data about the client as compared with the other conditions. In contrast, communicative effectiveness and efficiency should be the lowest in the random data condition, because the distorted information should result in a biased mental model of the client's knowledge and this should impair the expert's adaptation to the client's actual knowledge state to some degree.

Method

The assessment tool. The assessment tool provided the computer experts both with ratings of the client's general computer knowledge and their Internet knowledge (see Figure 1). Apart from these global evaluations, it was also displayed to what extent the client already knew the meaning of two specialist concepts semantically relevant to the understanding of the problem addressed by an inquiry. Thus, the experts had the possibility to adapt their explanations both to the client's general knowledge background and, on a more concrete level, to their prior knowledge regarding a specific inquiry. The values displayed in the assessment tool were determined through an objective and standardized assessment procedure. To this purpose, an updated version of the computer and Internet knowledge test developed by Richter et al. (2000) was constructed and pre-tested on 40 humanities students. In the experiment, the number of items that a client had solved correctly in the general computer knowledge subtest (10 items) and in the Internet knowledge subtest (10 items) was translated into values on the corresponding five-point scales in the assessment tool (cf. Figure 1). For example, if a client had solved only one or two items out of the ten items of the Internet knowledge subtest, this was indicated as a *low* Internet knowledge level. In contrast, if the client had solved nine or ten items of a subtest, this would be represented in the assessment tool as a *high* knowledge level. To assess the client's knowledgeability regarding the specialist concepts they were asked to describe the meaning of each of the concepts. Two raters independently scored the written descriptions for correctness by using the five-point rating scale displayed in the assessment tool (see Figure 1). Inter-rater reliability was .92.

Participants. 60 computer experts and 60 clients participated in the experiment. Computer experts were recruited among advanced students of computer science. The laypersons serving as clients were recruited among students of psychology and the humanities. The results of the knowledge tests showed that the clients covered a wide range of

different knowledge levels. In the general computer knowledge test, a mean of 5.33 correctly solved items was obtained with a standard deviation of 2.50 and a range of 10 items. In the Internet knowledge test, the clients were able to solve 5.80 items on average with a standard deviation of 2.33 and a range of 8 items. Thus, there was ample opportunity for the experts to adapt their explanations to clients with different prior knowledge levels.

Design. Computer experts and clients were combined into dyads that were randomly assigned to the experimental conditions. A one-factorial between-subjects design was used comprising three different conditions: (a) communication with an assessment tool displaying valid information about the client's knowledge (in the following labeled 'valid AT'), (b) communication without assessment tool ('no AT'), and (c) communication with an assessment tool displaying random information about the client's knowledge ('random AT'). Dependent variables encompassed measures of communicative effectiveness (i.e., the client's increase in knowledge) and communicative efficiency (i.e., the number of questions asked by the client in response to an expert's explanation).

Materials. A pool of 20 inquiries was constructed that demanded explanations of relevant Internet topics and problems. Based on expert ratings regarding the familiarity and relevance of the inquiries, six of them were selected for the experiment. Three inquiries required the computer expert to explain a technical concept. The other three were more complex. They asked the expert to instruct the client how to solve a problem and, additionally, to provide an explanation why the problem occurred in order to help the client understand the nature of the problem (e.g., "I'm running Internet Explorer 6. Whenever I try to print a website consisting of several frames, my printer only prints out one frame. I would like to understand why this happens and what I can do so that the frames are printed out all at once?").

Procedure. In the beginning of the experiment, the students serving as clients were administered the general computer knowledge test, the Internet knowledge test, and the concept description task. In addition, their prior knowledge about the six inquiries to be discussed in the communication phase was determined. The students were encouraged to try to answer each of the inquiries if possible. They were informed that they were participating in a study on students' knowledge about computers and the Internet. Thus, it was made certain that the students had no reason to assume that their test results would later be relevant to the communication phase of the experiment. This was important because otherwise the students' self-perceptions of their test performance might have influenced their behavior during the advisory exchange with the computer expert. In the communication phase, the expert and client sat in different rooms and communicated through a text-based interface. The client's task was to sequentially direct each one of the prepared six inquiries verbatim to the expert by typing the prepared wording of the inquiry into the text form of the interface. The expert was asked to answer each inquiry as well as possible. The clients were encouraged to write back and ask as many questions as needed. In the experimental conditions with the assessment tool, the completed form was visible to the expert during the entire course of the exchange. When the client asked a new inquiry, the assessment tool was automatically updated with regard to the client's knowledge about the specialist concepts relevant to the current inquiry (see Figure 1). After the communication phase, the clients were again asked to write down their knowledge about each of the six inquiries. In this way, it was possible to calculate the individual increase in knowledge for each client (cf. Table 1).

Results

Before the client's individual increase in knowledge was computed, it was made sure that the clients had no substantial prior knowledge about the inquiries. The mean scores of the clients' answers collected *before* the communication phase clearly ranged below one (4-point rating scale, cf. Table 1) indicating that, on average, the clients did not know the correct answer to the inquiries prior to the exchange with the computer expert. There were no differences between the experimental conditions, $F < 1$.

Communicative effectiveness. In order to compute the clients' individual increase in knowledge, the mean scores of the clients' answers to the six inquiries prior to the communication phase were subtracted from the corresponding mean scores after the communication phase (cf. Table 1). The maximum score to be attained was three points. An ANOVA performed on the individual difference scores revealed an overall effect of experimental condition, $F(2, 57) = 5.37$, $p < .01$, $\eta^2 = .16$ (strong effect). Following the sensitization hypothesis, a substantial increase in knowledge should be observed in the conditions with an assessment tool but *not* in the condition without an assessment tool. The validity of the displayed information should make no difference. This prediction was represented by the following contrast: valid data: 1, random data: 1, no assessment tool: –2.

Following the specific adaptation hypothesis, the information displayed by the assessment tool should indeed make a difference: The client's increase in knowledge should be larger in the valid data condition compared with the condition without the assessment tool and the random data condition. The smallest knowledge increase would be expected in the random data condition, because the distorted information should impair the expert's adaptation to the client's knowledge level. This linear trend hypothesis was represented by the following contrast weights: valid data: 1, no assessment tool: 0, random data: –1.

The results of the contrast analysis clearly contradicted the sensitization hypothesis and supported the specific adaptation hypothesis. The planned contrast representing the sensitization hypothesis failed to reach statistical signifi-

cance, $F < 1$, whereas the contrast testing the specific adaptation hypothesis was highly significant, $F(1, 57) = 9.99$, $p < .01$, $\eta^2 = .15$ (strong effect). Table 1 shows that the mean values of the clients' increase in knowledge evidently displayed the predicted linear trend with the largest increase in knowledge occurring in the valid data condition and the smallest in the random data condition.

Table 1: Means and standard deviations (in parentheses) of the dependent variables of the experiment.

	Experimental Condition		
Dependent Variable	Valid AT	No AT	Random AT
Mean scores of the clients' answers *before* the communication phase*	0.46 (0.38)	0.68 (0.73)	0.58 (0.50)
Mean scores of the clients' answers *after* the communication phase*	1.97 (0.71)	1.66 (0.55)	1.37 (0.59)
Mean differences of the clients' increase in knowledge	1.52 (0.81)	0.99 (0.78)	0.80 (0.54)
Total number of questions per expert-client exchange	2.15 (1.73)	4.35 (2.98)	4.70 (2.54)
Number of comprehension questions per expert-client exchange	1.75 (1.74)	3.80 (2.44)	3.85 (2.13)

Note. *For each answer up to three points could be assigned (0 = no or wrong answer, 1 = predominantly wrong answer, 2 = roughly correct answer, 3 = completely correct answer).

Communicative efficiency. To obtain a measure of communicative efficiency, we counted the total number of questions the client produced in response to the expert's explanations during the whole exchange, that is, throughout the six inquiries. An ANOVA performed on the total number of questions revealed a significant overall effect of experimental condition, $F(2, 57) = 6.27$, $p < .01$, $\eta^2 = .18$ (strong effect). When the analysis was restricted to the frequency of comprehension questions, that is, to those questions by which the client explicitly articulated a comprehension problem, a similar result was obtained, $F(2, 57) = 6.36$, $p < .01$, $\eta^2 = .18$ (strong effect). To test the sensitization hypothesis and the specific adaptation hypothesis, planned contrasts were computed with the contrast weights reported above. As before, the data analyses yielded no support for the sensitization hypothesis, regardless of whether the total number of questions or the number of comprehension questions was used as the dependent variable, $F(2, 57) = 1.87$, *ns*, and $F(2, 57) = 2.95$, *ns*, respectively. On the other hand, the specific adaptation hypothesis was also confirmed with regard to communicative *efficiency*. The linear contrast was significant when the total number of questions was considered, $F(1, 57) = 10.67$, $p < .01$, $\eta^2 = .16$ (strong effect), and also when the analysis was restricted to the comprehension questions, $F(1, 57) = 9.76$, $p < .01$, $\eta^2 = .15$ (strong effect). With valid data in the assessment tool, the laypersons wrote back only about half as often in response to an expert's explanation as compared to the other experimental conditions (cf. Table 1, last two rows). Thus, only the provision of valid information reduced the frequency of questions by which the client explicitly articulated a comprehension problem or asked for further information. On the other hand, most of the questions occurred in the condition that presented distorted information about the client's knowledge.

Discussion

The dialogue experiment presented in this paper replicates the results found in a previous study (Nückles & Stürz, in press). The approach to support asynchronous communication between computer experts and laypersons by means of an assessment tool has indeed proven to be successful. More importantly, the present findings also allow for conclusions about the mechanisms that led to the increase in communicative effectiveness and efficiency. The clients acquired the most knowledge and asked the fewest questions when the computer expert was presented *valid* data about the client's knowledge. When the information was distorted, the client's knowledge acquisition was impaired. The clients in the random data condition profited the least from the experts' explanations and asked the most questions. These results clearly contradicted the sensitization hypothesis and supported the specific adaptation hypothesis. Thus, it can be concluded that it was in fact the *individuating information* about the client's knowledge that led to the increase in communicative efficiency and effectiveness. From the perspective of Nickerson's anchoring and adjustment model (Nickerson, 1999) the assessment tool improved the communication between expert and client because the information about the client's knowledge provided the computer expert with a *specific* anchor right from the start of the counselling process. This enabled the expert to calibrate their mental model about the client's knowledge more quickly and accurately than would have been possible without such individuating information or with distorted information.

The present findings show that the assessment tool fostered specific adaptation to the clients' knowledge. It is still unclear *how* the experts exactly used the information about the client to produce adaptive explanations. It is plausible to assume that the experts used a 'linear strategy'. For example, they might have reasoned that 'the lower the client's knowledge level, the more extensive my explanations should be' in order to provide the client sufficient context for comprehension (cf. Clark, 1996). Indeed, we found such a correlation between the extensiveness of the experts' explanations and the client's displayed knowledge level ($r = -.32$, $p < .05$). Still, the correlation was rather low. It cannot help to fully understand the cognitive heuristics the computer experts used to adjust their explanations to the client's knowledge. Thus, beyond the tendency to link explanatory

extensiveness to the client's knowledge level, the experts apparently used the information displayed by the assessment tool to adjust their explanations in more sophisticated and individualized ways. One possibility is that the experts referred to the information in the assessment tool to make decisions during the planning phase of an explanation, for example, whether a technical term they intended to use in their answer would already be known by the client, or would have to be introduced in case it was not known (so-called 'pruning', see Chin, 2000). To explore such possibilities, we are currently running a 'think-aloud' study in which we are investigating how the experts developed a qualitative representation of the client's knowledge from the quantitative information provided by the assessment tool and how this qualitative representation was used to generate instructional explanations for the client. In addition, 'thinking aloud' protocols of the client's comprehension processes could help to identify features of the expert's explanations that hinder or enhance the clients' understanding. The identification of features that make an expert's explanation well adapted to a specific knowledge level could be interesting both for the design of advice-giving systems (e.g., Chin, 2000) and intelligent tutoring systems (e.g., Dede, 1986).

The finding that information about a client's knowledge fostered the provision of adaptive instructional explanations might also be suggestive of ways in which Internet-based collaborative settings other than helpdesk communication could be supported. In the realm of distance learning, many universities offer online courses where students of diverse educational backgrounds and with a wide range of different knowledge participate. As the tutors in these courses have to provide instructional explanations for people they – at least initially – do not know, an assessment tool could provide valuable information that may help the tutors to adapt their explanations to the learners' knowledge level. Hence, an assessment tool might also be an appropriate method to support online tutoring (Siler & VanLehn, in press).

Acknowledgments

This project is funded by the German Research Foundation (DFG; NU 129/1-1).

References

Alty, J. L., & Coombs, M. J. (1981). Communicating with university computer users: A case study. In M. J. Coombs & J. L. Alty (Eds.), *Computing skills and the user interface* (pp. 7-71). London: Academic Press.

Chi, M., Glaser, R., & Farr, M. (Eds.) (1988). *The Nature of Expertise*. Hillsdale, NJ: Erlbaum.

Chin, D. N. (2000). Strategies for Expressing Concise, Helpful Answers. *Artificial Intelligence Review, 14 (4-5)*, 333-350.

Clark, H. H. (1996). *Using language.* Cambridge: University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: American Psychological Association.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. F. LeNy & W. Kintsch (Eds.), *Language and comprehension* (pp. 287-299). Amsterdam: North-Holland Publishing Company.

Dede, C. (1986). A Review and Synthesis of Recent Research in Intelligent ComputerAssisted Instruction. *International Journal of Man-Machine Studies, 24*, 329-353.

Gunawardena, C. N. (1995). Social Presence Theory and Implications for Interaction and Collaborative Learning in Computer Conferences. *International Journal of Educational Telecommunications*, *1 (2/3)*, 147-166.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied, 5 (2)*, 205-221.

Kiesler, S., Zdaniuk, B., Lundmark, V., & Kraut, R. (2000). Troubles with the Internet: The dynamics of help at home. *Human-Computer Interaction, 15*, 323-351.

Moncarz, R. (2001). Computer support specialists. *Occupational Outlook Quarterly, 45*, 16-19.

Nickerson, R. S. (1999). How we know - and sometimes misjudge - what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125 (6)*, 737-759.

Nückles, M., & Bromme, R. (2002). Internet experts' planning of explanations for laypersons: A Web experimental approach in the Internet domain. *Experimental Psychology, 49 (4)*, 292-304.

Nückles, M., & Stürz, A. (in press). The assessment tool. A method to support asynchronous communication between computer experts and laypersons. *Computers in Human Behavior.*

Nückles, M., Wittwer, J., & Renkl, A. (2003). Supporting computer experts' adaptation to the client's knowledge in asynchronous communication: The assessment tool. In F. Schmalhofer, R. Young, & G. Katz (Eds.), *Proceedings of the European Conference of the Cognitive Science Society 2003* (pp. 247-252). Mahwah, NJ: Erlbaum.

Renkl, A. (2002). Learning from worked-out examples: Instructional explanations support self-explanations. *Learning & Instruction, 12*, 529-556.

Richter, T., Naumann, J., & Groeben, N. (2000). Attitudes toward the computer: Construct validation of an instrument with scales differentiated by content. *Computers in Human Behavior, 16*, 473-491.

Siler, S. A., & VanLehn, K. (in press). Accuracy of Tutors' Assessments of their Students by Tutoring Context. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.

Vu, K. L., Hanley, G. L., Strybel, T. Z., & Proctor, R. W. (2000). Metacognitive processes in human-computer interaction: Self-assessments of knowledge as predictors of computer expertise. *International Journal of Human-Computer Interaction, 12*, 43-71.

ACT-R is *almost* a Model of Primate Task Learning: Experiments in Modelling Transitive Inference

Mark A. Wood (cspmaw@cs.bath.ac.uk)
Jonathan C. S. Leong (jleong@fas.harvard.edu)
Joanna J. Bryson (jjb@cs.bath.ac.uk)
University of Bath; Department of Computer Science
Bath, England BA2 7AY; United Kingdom

Abstract

We present a model of transitive inference (TI) using ACT-R which strengthens the hypothesis that TI is not dependent on underlying sequential ordering of stimuli, but rather on the learning of productions. We nevertheless find a weakness in the ACT-R sub-symbolic learning system and suggest improvements.

Introduction

The last decade has shown an increasing body of work indicating that ACT-R (Anderson and Lebiere, 1998) can be used to model human learning in an impressive variety of tasks. However, ACT-R has been slow to gain acceptance in mainstream experimental psychology as a useful model, possibly because it does not seem a very good correlate to the physical learning systems we find in the brain.

In previous and concurrent work, we have been exploring another model of task learning which also seems at first blush artificial and not particularly parsimonious, but has also shown an impressively tight fit to human and animal experimental data otherwise unaccounted for. This is the Harris (1988) production-rule-stack model of one of the main testbeds of task learning in the animal literature, the transitive inference task. We developed the two-tier model (Bryson, 2001; Bryson and Leong, 2004), which accounts for all of the Harris model's data, while extending the model to account for both learning and failing to learn this task (a common outcome in live subjects). We've also found potential neurological correlates for the two-tier model.

The two-tier model hypothesises two learning systems: one for connecting perceptual contexts to actions, and another for prioritising which of those perceptual contexts to attend to if more than one are present simultaneously. We realised that this model had similar aspects to ACT-R, which also has two learning systems, one symbolic and one statistical. We therefore decided to apply ACT-R to the transitive inference learning task. Our results show that ACT-R is far better than the standard TI models at accounting for the particular (and somewhat controversial) data set that prompted the Harris (1988) model, and for some individuals provides a better model than Harris (1988), though for others it cannot. Our results lead us to believe that the two-tier model is the best existing model of transitive inference, although ACT-R is close enough that it is probably fixable. ACT-R demonstrates one significant simplification over the two-tier hypothesis, and has one important difference from real mammalian task learning.

Transitive Inference

Transitive inference (TI) formally refers to the process of reasoning whereby one infers that if, for some quality, $A > B$ and $B > C$, then $A > C$. In some domains, such as integers or heights, this property will hold for any A, B or C, though for others it does not (see Wright, 2001, for a recent discussion). TI is classically described as an example of concrete operational thought (Piaget, 1954). That is, children become capable of doing TI when they become capable of mentally performing the physical manipulations they would use to determine the correct answer, a stage they reach at approximately the age of seven. In the case of TI, this manipulation involves ordering the objects into a sequence using the rules $A > B$ and $B > C$, and then observing the relative location of A and C.

Since the 1970's, however, apparent TI has been demonstrated in much younger children (Bryant and Trabasso, 1971) and a variety of animals, from monkeys (McGonigle and Chalmers, 1977) to pigeons (Fersen et al., 1991) — not normally ascribed with concrete operational abilities. The behaviour of choosing A from AC without training after having previously been trained to select A from AB and B from BC is consequently sometimes referred to as "transitive *performance*", and whether it implies sequential ordering at all is now an open issue.

The main motivation for *not* considering TI in animals to be based on a sequential structure is a dataset due to McGonigle and Chalmers (1977), which they have subsequently replicated both with monkeys and children. This data set concerns what happens if subjects demonstrating TI are asked to select between *three* items rather than two. Some individuals show significant, systematic degradation in performance, which cannot be explained by a sequential model. Some researchers have dismissed the triad data as resulting from confusion in the subjects due to the extra item. These criticisms were dealt with in a replication by McGonigle and Chalmers (1992) which provided the main data set used in this paper and by Harris and McGonigle (1994). The fact that the systematicity of the degradation has now been success-

fully accounted for further validates this data set. This data concerns monkeys trained on 4 adjacent pairs drawn from a 5 item sequence, AB, BC, CD, DE.

The Models

Due to space constraints we will not review the more traditional, sequence-based or simple-associative models of TI, but see further (Wynne, 1998; Bryson and Leong, 2004). These models cannot account for the triad data set.

Harris and McGonigle

Harris (1988) showed that both pair and triad TI performance could be accounted for if we assume that monkeys learn a production rule stack. A *production rule* is a basic AI representation which connects a stimulus to a response. A *stack* is a prioritised list. In the Harris model, each monkey learns one rule per possible stimulus, or up to 5 rules in total. One of two actions is associated with each rule, either *select* or *avoid*. If a subject applies the rule $A{\rightarrow}s(A)$ (see A implies select A), then it will simply pick up A, regardless of whether other items are present. However, if a subject applies the rule $A{\rightarrow}a(A)$ it will pick up anything *but* A. If more than one other item is present, the subject is at chance for which object it will pick up. If more than one rule could apply, then whichever rule is higher in the stack (has higher *priority*) will be applied.

Although Harris' hypothesis may seem obscure, it shows a remarkable match to the data. If one assumes that rules are limited to the case that the action refers to the object attended to, then only 16 of the 1920 ($10 \times 8 \times 6 \times 4$) possible stacks of four rules operate correctly on all training pairs (Harris and McGonigle, 1994). All 16 of these stacks also correctly perform TI on *all* pairs automatically, thus already accounting for one of the mysteries of transitive performance.

The degradation some subjects display on the triad tests is a consequence of the random aspect of the *avoid* rules. In fact, triads can be used to discriminate which rule stack an individual subject has learnt. For example, a stack that consists entirely of selects $(s(A), s(B), s(C) \ldots)$ will never make any errors. One that starts with $a(E)$ will be at chance between the other two options whenever E is present in a triad.

Table 1 shows all of the possible discriminable stacks as identified by Harris and McGonigle (1994). Because only the highest-priority applicable rule fires and there are always at least two applicable rules (since there are at least two stimuli), there is no way to discriminate the two lowest-priority rules using triad performance. These stacks therefore only reflect the top three rules of the stacks.

The Two-Tier Model

A successfully trained two-tier model creates a replication of the production-rule-stack model (Bryson, 2001). However, the two-tier model is dynamic, and as such gives us insight into why animals have trouble learning the initial pairs for the TI task, the sorts of mistakes they may make, and the impact of training régimes. The first tier of the two-tier model is a single-vector neural network (NN) which learns the prioritisation of the stimuli. The second tier is a set of small two-item vectors which each learn to associate an action with one of the stimuli. The learning rule for the NNs is a slight simplification of standard delta learning (Widrow and Hoff, Jr., 1960).

Simulations using the two-tier model show artificial subjects successfully learning the training data only about 25% of the time when training pairs are presented in a random order. However, switching to the training régime applied by McGonigle and Chalmers (1992) shown in Table 2, which is standard for primates, the success rate increases to about 75%, which is comparable to live subjects (Bryson and Leong, 2004).

Further, the sorts of errors made by artificial subjects failing to learn are consistent with those shown by live subjects — they tend to confuse the middle pairs. Analysis of the networks shows that this is nearly always a consequence of misprioritising the rules representing the end pairs. An agent can guarantee it always selects A in the pair AB (the only pair A appears in) by learning $a(B)$, and there is a great inclination to learn about middle rules because these are the ones that have the most data (and the most confusing data, since B is sometimes rewarded but sometimes penalised.) However, there are no successful stacks which do not have one end point or the other at the highest priority (see Table 1). The training régime greatly increases the probability of learning correct prioritisation. For further details see Bryson and Leong (2004).

Table 1: Enumeration of Harris and McGonigle Stacks

	Rule Depth				Rule Depth		
#	1	2	3	#	1	2	3
1	s(A)	s(B)	s(C)	5	a(E)	s(A)	s(B)
2	s(A)	s(B)	a(E)	6	a(E)	s(A)	a(D)
3	s(A)	a(E)	s(B)	7	a(E)	a(D)	s(A)
4	s(A)	a(E)	a(D)	8	a(E)	a(D)	a(C)

Table 2: Primate TI training régime (Chalmers and McGonigle, 1984; McGonigle and Chalmers, 1992)

P1	Each pair in order (DE, CD, BC, AB) repeated until 9 of 10 most recent trials are correct. Reject if requires over 200 trials total
P2a	4 of each pair in order. Criteria: 32 consecutive trials correct. Reject if requires over 200 trials total
P2b	2 of each pair in order. Criteria: 16 consecutive trials correct. Reject if requires over 200 trials total
P2c	1 of each pair in order. Criteria: 30 consecutive trials correct. No rejection criteria
P3	1 of each pair randomly ordered. Criteria: 24 consecutive trials correct. Reject if requires over 200 trials total
T	Pair and triad testing

ACT-R

As for the above models, ACT-R also learns production rules, but any number of these rules may have their preconditions for firing satisfied at any given time. In this case, ACT-R's conflict-resolution system selects the rule with highest *utility* value.

Rule utilities are changed by ACT-R's sub-symbolic processing system. It is possible to attach *success* or *failure* tags to productions and when such a rule is fired, ACT-R backtracks to discover which rules fired previously and increments or decrements their utilities respectively. More precisely, the utility of a rule is given by:

$$U = PG - C + \epsilon(s) \tag{1}$$

where G is the *goal value*, C is the *expected cost*, $\epsilon(s)$ is the *expected gain noise* and P is the *expected probability of success*:

$$P = \frac{\text{Successes}}{\text{Successes} + \text{Failures}} \tag{2}$$

In our experiments, rather than make arbitrary changes to ACT-R's many available parameters in an attempt to best fit the data, we have used mostly defaults. The most notable exception to this is that we set the initial Failure count to 1 which, along with ACT-R's default setting of 1 for Successes[1], gives an initial probability of success of 0.5. This change was also made by Belavkin and Ritter (2003) in their Dancing Mouse model. To maximise the number of successful agents, we also tried a range of values for s (which affects the variance of the noise function), finally deciding upon $s = 1$.

One trial consists of two or three items displayed on-screen which the agent encodes into its goal buffer. The goal state is then changed, enabling it to make decisions about which item to pick (see below). Once an item has been picked, either a reward or no reward is displayed appropriately, the agent notes its success or failure respectively and the next trial begins.

We have tested two different approaches to solving the TI problem in ACT-R. In the first, the *select* and *avoid* rules for each item are independent, concurrent candidates for execution. For three displayed items, this corresponds to six conflicting rules that have their preconditions satisfied. Henceforth we refer to this approach as *ACT-R-1*.

In the second, the agent must *focus* on a displayed item before either *select*ing or *avoid*ing it, as in the two-tier model. This results in an extra stage of conflict-resolution for the agent. With three options there are at first three candidate *focus* rules whose actions alter the agent's goal state. This, in turn, satisfies two further rules; *select* and *avoid* for the focus-item. We call this approach *ACT-R-2*.

Results

As with the two-tier model and live subjects, our ACT-R model produces both agents that successfully learn the task and agents that do not. 44 of the 100 agents tested with ACT-R-1 successfully passed the training régime. This compares to 35% of those using ACT-R-2, or 75% of those using the two-tier system. We examine these groups separately.

Successful Agents

The stack model proposed by Harris and McGonigle (1994) attempts to fit the McGonigle and Chalmers (1977) triad data to any of the eight discriminable correct stacks (Table 1). In contrast, the ACT-R agents learn only two possible solutions.

There are two rules which are always successful for all pairs: s(A) and a(E). This means that, provided these rules are discovered by the agent, their utilities will begin to converge to maximal, given by:

$$\begin{aligned} \lim_{t\to\infty} U &= \lim_{t\to\infty} (PG - C) \\ &= \lim_{t\to\infty} \left(\frac{\text{Successes}}{\text{Successes} + \text{Failures}} \right) G - C \\ &= \lim_{t\to\infty} \left(\frac{1}{1 + \frac{\text{Failures}}{\text{Successes}}} \right) G - C \\ &= G - C \end{aligned} \tag{3}$$

where t is the number of trials, and G and C remain constant at ACT-R default values throughout. Therefore any successful ACT-R agent will have these two rules at highest priority. This eliminates half of the Harris and McGonigle stacks, leaving 3, 4, 5 and 6 (Table 1).

In addition, because the top-two rule utilities are converging to the same value, it becomes essentially arbitrary (in fact, governed by the expected gain noise) whether s(A) or a(E) occupies the top stack position for any given choice. In other words, the ACT-R agents do not learn a totally ordered stack, but effectively a pair of stacks. The two possible pairs are:

Hybrid Stack 1 (HS1): s(A)a(E)s(B) & a(E)s(A)s(B)
Hybrid Stack 2 (HS2): s(A)a(E)a(D) & a(E)s(A)a(D)

Table 3 shows each triad with the expected percentage of trials in which each item in that triad is chosen. These probabilities are the same for both Hybrid Stacks, except for the triad BCD. In this case, the format is HS1 / HS2. A 75%/25% split occurs when both A and E are present in the triad. We assume that half the time s(A) has top priority and is thus selected. Otherwise, a(E) has top priority, giving an even chance of A or the other item being selected.

Taking into account the noise added to the system, this model well describes the behaviour of many of the ACT-R agents. In 45% of cases, one of the Hybrid Stacks fitted the agent's distribution better than any of the individual stacks, and a further 47% were best fitted by Stack 5; a contributor to HS1.

Failed Agents

Despite these encouraging results, a majority of the ACT-R agents failed the training régime (Table 2), which is not true of the monkeys (albeit there were only seven

[1]The default is Failures $= 0 \Rightarrow P_{initial} = 1$.

Table 3: Projected percentage choice distributions for Hybrid Stack 1 / 2

Triad	A	B	C	D	E
ABC	100	0	0	-	-
BCD	-	100 / 50	0 / 50	0 / 0	-
BDE	-	50	-	50	0
CDE	-	-	50	50	0
BCE	-	50	50	-	0
ABD	100	0	-	0	-
ACD	100	-	0	0	-
ADE	75	-	-	25	0
ABE	75	25	-	-	0
ACE	75	-	25	-	0
Mean	52.5	22.5 / 17.5	12.5 / 17.5	12.5	0

test subjects) or children (Chalmers and McGonigle, 1984). Virtually all of the agents which failed did so at stage P2a (Table 2), typically having seen less than 300 training pairs in total. There are two exceptions for both ACT-R models which failed at P3.

To best understand why agents fail, we examine each training pair and determine what causes agents to pick the wrong item:

AB Since s(A) almost always has a high utility, errors on this pair tend to be caused by interference from s(B), whose utility is driven up by its success on pair BC. Ironically, then, it is agents who are too successful too soon who fail because of this pair, training stage P2a having the most stringent pass criteria.

BC C is picked when s(C) is too high relative to s(B) or, less frequently, a(C). Occasionally a(B) adds to this interference but, due to the success of s(A), rarely attains a high enough utility.

CD The symmetric case of BC. Here a(C)/a(D) interference is the chief cause of error (see Figure 1).

DE As for AB, if a(D) is discovered early to be a good rule, it interferes with a(E) causing small but significant errors in P2a.

For some agents (around 25%), failure is a result of a combination of the above interferences. If many of the interfering rules are interdependent (eg. s(B), s(C), a(C), a(D)) then this can lead to a more even distribution of errors across all training pairs. Conversely, if two sets of independent rules (eg. s(A), a(B), a(D), a(E)) are interfering, often two training pairs are consistently incorrect, with little or no error on the other two.

As Tables 4 and 5 show, agents confuse the middle pairs far more often than the end pairs (see also **Analysis**). This, in turn, is most often the result of s(C) and a(C), both of which perform incorrectly for one of the middle pairs. We have restricted these data to the last 200 trials carried out by the failed agent. This focuses on the specific phase of training at which the agent failed and removes the noisiest choices, made during P1.

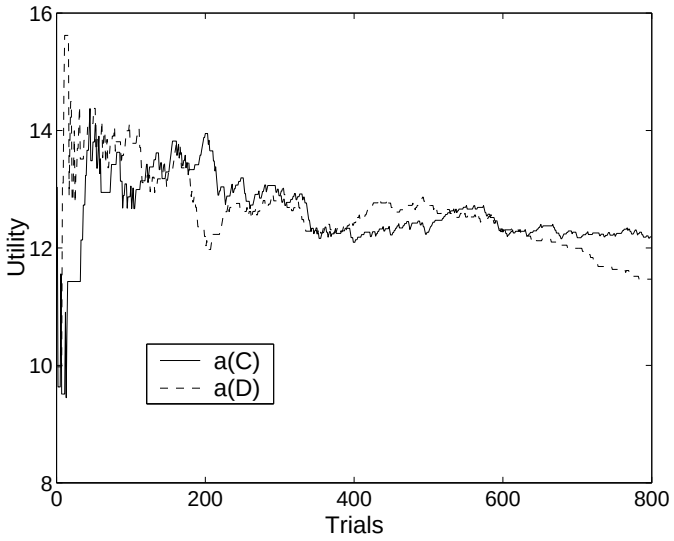


Figure 1: a(C) and a(D) fight for control of the pair CD

Table 4: Aggregate percentage error on each pair

Group	AB	BC	CD	DE	Mean
ACT-R-1	6	25	30	7	17
ACT-R-2	6	43	39	9	24

Table 5: Percentage distribution of failed agents

	Modal Error Pair			
Group	AB	BC	CD	DE
ACT-R-1	2	39	54	5
ACT-R-2	0	42	50	8

Analysis

For ease of statistical comparison, we applied the χ^2 test in the same way as Harris and McGonigle (1994): by excluding item E, which (usually) has an expected value of 0.

For three of the five individual test subjects for whom McGonigle and Chalmers (1977) triadic data is available, one of the hybrid stacks fits better than any of the eight others (Table 6). As explained in the **Successful Agents** section above, and in contrast with the Harris and McGonigle stacks, the hybrid stacks do not represent a total ordering. Thus it would seem that neither do some monkeys form a total ordering, and their choices cannot be perfectly modelled by a simple production-rule system. For Bump and Brown, however, our model is rejected ($p < 0.01$), suggesting that other monkeys do come up with a total ordering, which cannot be well modelled in ACT-R.

The two-tier model does support both models, although its admittedly simplistic learning rule tends to favour the total ordering. These results suggest that

Table 6: Comparison of individual triadic choice data (1977) with both Hybrid and Harris' Stack models

	A	B	C	D	E	χ^2	p(O)
Bill	55	17	20	8	0	-	-
HS2	52.5	17.5	17.5	12.5	0	2.1	n.s.
S 4	60	15	15	10	0	2.8	n.s.
Blue	55	25	14	6	0	-	-
HS1	52.5	22.5	12.5	12.5	0	4.0	n.s.
S 3	60	20	10	10	0	4.9	n.s.
Bump	53	34	8	4	1	-	-
HS1	52.5	22.5	12.5	12.5	0	13.3	< 0.01
S 2	60	30	5	5	0	3.4	n.s.
Brown	36	29	24	11	0	-	-
HS2	52.5	17.5	17.5	12.5	0	15.3	< 0.01
S 7	35	25	25	15	0	1.8	n.s.
Roger	51	26	6	17	0	-	-
HS1	52.5	22.5	12.5	12.5	0	5.6	n.s.
S 5	45	25	15	15	0	6.5	< 0.1

an improved priority-learning rule for either the two-tier model or ACT-R could result in a highly accurate model of TI and possibly task learning in general.

There were just five monkeys who passed criteria and so were included in the triad phase of the 1977 experiment. These five only represented three of the eight stacks in Table 1. This may render the grouped data unrepresentative, but our ACT-R model still displays a better correlation than Harris and McGonigle (1994) of α-choices, as shown in Table 7, where α represents the *correct* choice in a given triad (see also Table 8).

Table 7: Correlation of α-choices to group data

Group	r	p
ACT-R-1	0.688	$p < 0.05$
ACT-R-2	0.692	$p < 0.05$
Hybrid Stack Model	0.616	$p < 0.1$
H & M Stack Model	0.634	$p < 0.05$

Upon closer examination of the choices made for each triad, we see ACT-R closely matching the monkey data for those triads which do not contain the item E (Table 8). For those that do, ACT-R makes more mistakes, suggesting that the monkeys do not have a(E) at as high a priority. This might reflect a primate bias against having identical priorities for rules.

There may be a good reason for favouring priorities over utilities for ordering rules. For example, there is no circumstance in which an ACT-R agent can reach a stable enough solution to reduce error to zero. Suppose such a situation was attained. Then every decision made would result in success and thus increase the utility of the executed rule. Eventually, these rules (of which there must be at least three to produce a correct stack) would converge upon the same value (see Equation 3 above). But since no three rules in a correct stack are independent for all triads, they will start to interfere with each other, causing error to be re-introduced.

This phenomenon is best demonstrated by examining the errors of agents who did not take part in structured training, but were presented with the training pairs in a random order. Here, as is usual, the s(A) and a(E) rules have high, convergent utilities (these rules *are* independent for all training pairs). Then the utility of one (or both) of the other successful rules, s(B) and a(D), will also start to converge. This results in errors made on the *end pairs* since s(B) interferes with s(A) for AB (Figure 2) and a(D) interferes with a(E) for DE. Neither s(C) nor a(C) can attain a high utility, because they will perform incorrectly on one of the middle pairs (BC and CD). The final result is that the middle pairs are chosen consistently correctly, whereas the end pairs have small errors (typically around 15%). This certainly seems somewhat biologically implausible, and contradicts the *End-Anchor Effect* (Bryant and Trabasso, 1971; Wynne, 1998).

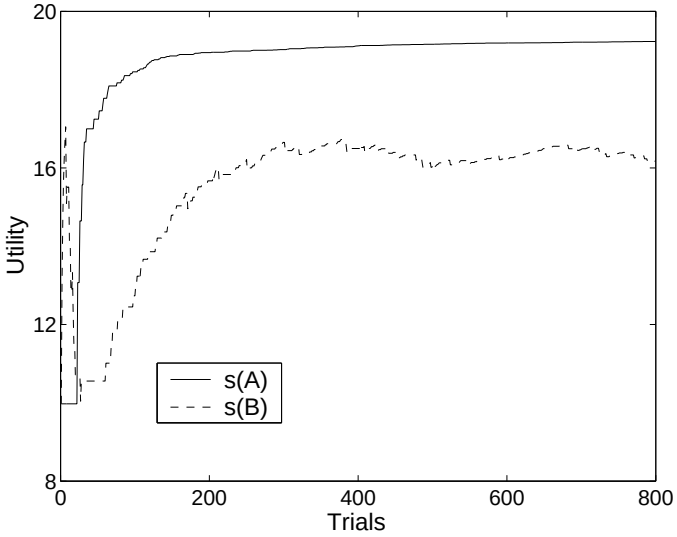


Figure 2: Interference with s(A) prevents the utility of s(B) reaching above a certain level

Conclusions and Further Work

The ACT-R models lack in their ability to represent stable, totally ordered stacks, which some real subjects appear to form. On the other hand, the Harris and McGonigle (1994) stacks lack the flexibility to represent more dynamic solutions to the TI problem. For this reason we conclude that the two-tier model is the best existing model of TI. On the other hand, the fact that there is no significant difference between ACT-R-1 (where no initial item focus is required) and ACT-R-2 (where this focus is required) implies that the two-tier model can be simplified to allow arbitrary numbers of stimulus action pairings, as is the default case in ACT-R.

Table 8: Percentage of items selected - triadic analysis

Triad	Monkeys			ACT-R-1			ACT-R-2			Hybrid Stack Model			H & M Stack Model		
$\alpha\beta\gamma$	α	β	γ	α	β	γ	α	β	γ	α	β	γ	α	β	γ
ABC	80	18	2	83	17	0	86	14	0	100	0	0	94	6	0
BCD	70	26	4	70	29	1	72	26	2	75	25	0	75	25	0
BDE	66	34	0	59	35	6	56	37	6	50	50	0	63	38	0
CDE	62	38	0	49	41	10	48	44	7	50	50	0	56	44	0
BCE	78	22	0	58	42	0	58	42	0	50	50	0	63	38	0
ABD	80	20	0	79	21	0	80	20	0	100	0	0	88	13	0
ACD	86	12	2	90	9	0	91	8	0	100	0	0	88	13	0
ADE	86	14	0	72	24	4	71	26	4	75	25	0	75	25	0
ABE	88	12	0	68	32	0	70	30	0	75	25	0	75	25	0
ACE	80	20	0	76	24	0	74	25	0	75	25	0	75	25	0
Means	78	22	1	70	28	2	71	27	2	75	25	0	75	25	0

There are several obvious next steps. First, as stated in the **Introduction**, the learning rules for priorities in both the two-tier model and ACT-R need improvement, though in different ways. We will be focusing on improving the two-tier model, but would be happy to see or support ACT-R being modified to reflect these results. Also, we suggest two possible improvements to the ACT-R model. Allowing ACT-R to compile its own system of rules from a minimal starting set (Anderson and Lebiere, 1998) may provide a more natural solution, although interpreting the underlying decision processes would be more difficult. Starting with a high initial noise would allow the agents to always discover and benefit from the most successful rules, while rapidly reducing this noise level (in conjunction with the *entropy of success* (Belavkin and Ritter, 2003)) would be necessary in order to obtain a stable enough solution to pass stage P2a of the training régime.

The other obvious next step would be to collect and analyse more triad testing results across a larger number of primates. For our purposes, it would be useful to have triad testing on subjects who fail TI training as well as those who succeed. We are investigating collaborations in this area.

The greatest significance of this work is that it gives further evidence for a non-sequence-based representation underlying the TI task and further supports the utility of the McGonigle and Chalmers (1977) triad data set.

References

Anderson, J. R. and Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Belavkin, R. V. and Ritter, F. E. (2003). The use of entropy for analysis and control of cognitive models. In Detje, F., Dörner, D., and Schaub, H., editors, *Proceedings of the Fifth International Conference on Cognitive Modelling*, pp. 21–26. Universitäts-Verlag Bamberg.

Bryant, P. E. and Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, 232:456–458.

Bryson, J. J. (2001). *Intelligence by Design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agents*. PhD thesis, MIT, Department of EECS, Ch. 9. AI Technical Report 2001-003.

Bryson, J. J. and Leong, J. C. S. (2004). A two-tiered model of primate transitive performance: Separate memory systems for rule-stimulus association pairs and their prioritisations. In preparation.

Chalmers, M. and McGonigle, B. O. (1984). Are children any more logical than monkeys on the five term series problem? *Journal of Experimental Child Psychology*, 37:355–377.

Fersen, L., Wynne, C. D. L., Delius, J., and Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 17:334–341.

Harris, M. R. (1988). *Computational Modelling of Transitive Inference: a Micro Analysis of a Simple Form of Reasoning*. PhD thesis, University of Edinburgh.

Harris, M. R. and McGonigle, B. O. (1994). A model of transitive choice. *The Quarterly Journal of Experimental Psychology*, 47B(3):319–348.

McGonigle, B. O. and Chalmers, M. (1977). Are monkeys logical? *Nature*, 267:694–696.

McGonigle, B. O. and Chalmers, M. (1992). Monkeys are rational! *The Quarterly Journal of Experimental Psychology*, 45B(3):189–228.

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books, New York.

Widrow, B. and Hoff, Jr., M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, 4:96–104.

Wright, B. C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. *Developmental Review*, 21(4):375–422.

Wynne, C. D. L. (1998). A minimal model of transitive inference. In Wynne, C. D. L. and Staddon, J. E. R., editors, *Models of Action*, pages 269–307. Lawrence Erlbaum Associates, Mahwah, NJ.

Coordination of Component Mental Operations in Sequences of Discrete Responses

Shu-Chieh Wu (scwu@mail.arc.nasa.gov)
Roger W. Remington (roger.w.remington@nasa.gov)
NASA Ames Research Center, Mail Stop 262-4
Moffett Field, CA 94035 USA

Harold Pashler (hpashler@ucsd.edu)
Department of Psychology, University of California, San Diego
La Jolla, CA 92093 USA

Abstract

In daily life we often perform sequences of actions, which with practice are accomplished by overlapping mental operations for successive actions. Is it possible to derive performance predictions for such sequences from a characterization of the mental operations for a single stimulus-response pair? We explore this by examining the joint timing of eye movements and manual responses in a typing-like task following Pashler (1994). Participants made separate choice responses to a series of five stimuli spread over a wide viewing area. Replicating Pashler's results, responses to the first stimulus (RT1) were elevated, with inter-response intervals (IRI) for subsequent items rapid and flat across items. The eyes moved toward the next letter about 800 ms before the corresponding manual response (eye-hand span). Analyses of manual responses show multiple components to the RT1 elevation. Analyses of dwell times show that the eyes move to the next stimulus before the completion of all central processing.

Introduction

Current frameworks of human performance modeling often follow traditional theories of human cognition, treating human behavior as a succession of stages composed from a limited number of component mental operations, such as perceptual, cognitive and motor processes. The nature and duration of these mental operations are derived from studies of response time in discrete tasks, which often last less than one second. In the real world, however, tasks are rarely completed with a single discrete action. Rather, they often require the performance of a series of discrete actions integrated into a fluid behavior sequence in response to multiple stimuli during an extended period of time. In the transition from discrete to continuous new behaviors emerge, not previously observed, such as coordination and overlapping among component mental operations. It is an important question for human performance modeling whether models of single-task performance, described at the level of elementary mental operations, are sufficient to characterize behavior in extended, fluid sequences.

The successes of current human performance modeling suggest the answer is yes, at least for highly skilled behavior (e.g., Gray et al., 1991; Matessa et al., 2002). Coordination and overlapping among component operations are simulated by enforcing logical dependencies among operations distributed across different resources, interleaving upcoming operations in the slack time created by queued bottleneck processes, and allowing operations from different resources to proceed concurrently. The success of this approach depends on the underlying assumption that component mental operations inferred from discrete task performance do not function differently in extended task environments. This assumption has yet to be tested. Also, success has been achieved for tasks that are largely perceptual-motor, with good fits obtaining after about 100 contiguous trials (e.g., John et al., 2002).

The goal of the present research is to investigate the coordination of component mental operations in extended task sequences that require a sequence of simple choice responses. To better contrast the coordination among component operations that may arise in extended task performance with the simple progression through set stages thought to underlie discrete task performance, we choose an extended task that consists of a monotonic sequence of identical discrete tasks. This approach helps place the emphasis on the coordination among component operations of different instances of the same task rather than among different tasks. Of all possible cases of coordination, we are especially interested in how movements of the eyes are coordinated with other underlying mental operations. Eye movements are an integral part of most cognitive activities. Their effortless and seamless integration with other components of task performance provides possibly the best example of coordination and the most challenging task for human performance modelers. Yet in existing frameworks the implementation of eye movements (or gaze resources) tends to be greatly simplified. In addition, the way by which eye movements are used is usually based on empirical findings from task conditions where eye movements are specifically made to meet instructions rather than generated naturally in accord with task goals. Little has been known on how task-driven eye movements are coordinated with the succession of stages and processes thought to characterize the underlying mental operations.

In this paper, we present our recent work on how eye movements are integrated with underlying component mental operations in extended tasks. We begin by reviewing existing literature on extended task performance with eye movement measures. Then we present the results of two earlier extended task experiments, followed by a new

experiment designed to address specific issues raised by the previous work. In the end, we discuss the implications of our results with an emphasis on how they inform us on modeling human performance in extended tasks.

Eye movements in extended tasks

Although eye movements occur naturally in almost all daily activities, to characterize the patterns of eye movements researchers in the past have focused activities with a clear script. Examples of such activities range from golf putting (Vickers, 1992), driving (Land & Lee, 1994), to tea making (Land & Hayhoe, 2001), and block-copying (Pelz et al., 2001). A common finding in such observations is that the eyes move in anticipation of upcoming actions during activities that involve scripted behavior.

The existence of preview in extended task performance characterizes the proactive nature of eye movement control. In tasks that require mostly non-visually based decisions, it seems intuitive that the eyes could move away prior to the response as soon as information acquisition is completed. *But, when can the eyes move and what determines it?* Answers to these questions are critical to understanding the coordination between eye movements and other mental operations. As typical fixation durations generally range from 200 to 400 ms, exceeding the time needed for perceptual registration, which can be estimated at around 100 to 150 ms (Salthouse & Ellis, 1980), this suggests that certainly other variables are involved.

Previous Research

Previously, we (Wu & Remington, 2004) examined the coordination between ongoing mental processing and the generation of eye movements in a task requiring multiple manual responses to multiple stimuli on each trial. Specifically, we were interested in two empirical questions. First, in an extended task with multiple stimuli to be responded to, when do the eyes move away from a stimulus? Second, in such an extended task how is the processing sequence affected by difficulty manipulations at separate stages? By independently varying the difficulty of perceptual and central stages we can determine which is on the critical path for the sequence of responses.

We adopted a typing-like task introduced by Pashler (1994). Participants viewed a series of five letters sequentially and responded to each individually in different preview conditions. Pashler manipulated preview to test how the mental processing of two or more stimuli were overlapped in time. He measured the reaction time (RT) to the first stimulus (RT1) and computed the inter-response intervals (IRIs) for subsequent responses. With no preview, RT1 and subsequent IRIs were roughly equivalent and constant across the stimulus sequence. With preview, RT1 was elevated, compare to no preview, while IRIs were constantly low. The same effects were observed regardless of whether 1 or 4 preview items were presented. Pashler interpreted the constant IRIs as an indication of a bottleneck central processing stage of response selection, which would only allow the selection of one response at a time. The fact that IRIs reflected the duration of response selection is further supported by the findings that varying the duration of stimulus recognition and response production had little to modest effect on the durations of IRIs.

Pashler's (1994) task presents a simple example of the operations of three critical mental components (perception, response selection, and response production) and a clear theoretical account for the coordination among them. In this case, characterization of a single task was sufficient to account for the IRI results without further assumptions. The model, however, did not predict the elevated RT1. The experimental paradigm represents a good compromise between the simplicity of typical discrete trial experiments, and real-world behavior.

In our previous work, we adopted Pashler's complete preview condition and incorporated an eye movement component by reducing the size of stimulus letters and increasing the separation between them. Identification of stimulus letters thus required successive saccades and fixations. In two separate experiments, we examined response time, dwell time (fixation duration), and eye-hand span associated with manipulation of the duration of perception and response selection stages.

Our first experiment examined the effect of perceptual difficulty on dwell time. Perceptual difficulty was manipulated by having two luminance conditions for the stimuli, Dim and Bright (5.2 and 46.2 cd/m2, respectively). Participants made sequential fixations to each of the five stimulus characters randomly drawn from the set T, D, and Z, and made choice responses accordingly. Those three letters were mapped to three response keys (V, B, and N) on a PC keyboard and assigned to the first three digits of the right hand. We measured the manual RT to each of the five stimuli and the IRIs. In addition, we derived three eye movement related measures: 1) eye-hand spans, which represent the elapsed time between the initial fixation on a particular stimulus to the moment when the corresponding manual response is generated; 2) dwell time, which represents the duration for which fixation is maintain on a particular stimulus; and 3) release-hand spans, which represent the elapsed time between the end of fixation on a particular stimulus to the moment when the manual response is generated. In fact, dwell times and release-hand spans make up eye-hand spans.

Figure 1 shows mean manual RTs, eye-hand spans, and dwell times as a function of stimulus in our first experiment. The pattern of manual RT results resembled what Pashler (1994) found in conditions with preview; specifically, the elevation of RT1 and constantly short IRIs of subsequent responses. The effect of perceptual difficulty was minimal on RT1/IRIs and appeared to be restricted to S1. Dwell time was lengthened in the Dim condition, though the amount of increase did not reach statistical significance. Results of this experiment confirmed that dwell time encompasses perceptual processes.

Our second experiment examined the effect of response selection difficulty on dwell time. The difficulty of response selection was manipulated by using two sets of stimuli to create two mapping conditions. One set included four alphabets T, D, Z, and Q mapped in this arbitrary order onto keys V, B, N, and M, and assigned to the four digits of the right hand; another set included digits 1, 2, 3, and 4 mapped in this natural order to the same four keys and fingers.

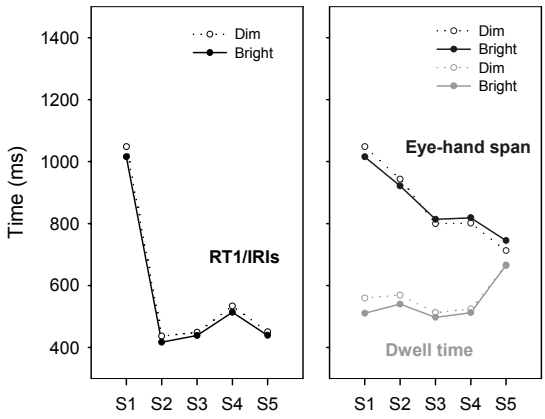
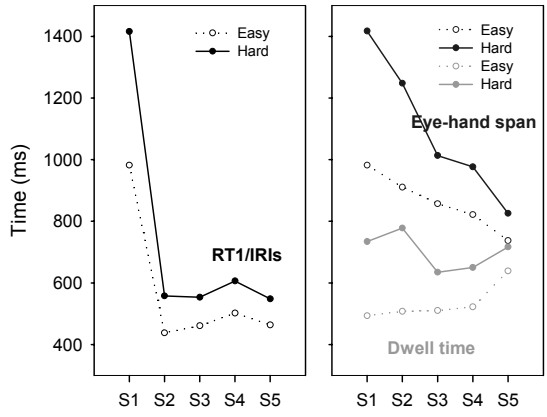


Figure 1. Results of Wu & Remington (2004), Experiment 1

Figure 2 shows mean manual RTs, eye-hand spans, and dwell times as a function of stimulus in our second experiment. Again, the manual RT results replicated the general pattern observed in our first experiment and in Pashler's (1994) study; RT1 was elevated, and IRIs were constant and rapid. In addition, mapping difficulty had a strong effect on manual as well as oculomotor responses. Difficult response mapping resulted in increases in IRIs. It also significantly increased dwell times. Results from this experiment suggest that fixation durations appear to include response selection related processes as well.

Present Experiment

Results from our previous work (Wu & Remington, 2004) provided some answers to the questions posed earlier. In an extended task such as this, the eyes move away at some point during the response selection stage but definitely after completion of the perceptual stage. Results from our



previous work also featured some unexpected patterns of coordination between the eyes and the hand. One in particular is the interrelated temporal constraint among

Figure 2. Results of Wu & Remington (2004), Experiment 2

dwell times, IRIs and eye-hand spans. Except for the Hard mapping conditions in the second experiment, dwell times were mostly constant across stimuli, as were IRIs. In other words, the eyes dwell for a constant duration, and the hand releases responses also at a constant but faster rate. This leads to the observed decrease in eye-hand span across stimuli.

The response of RT1 to the Easy and Hard mapping conditions was also unexpected. Though we always see an elevated RT1, its increase of approximately 400 ms in the hard condition was about twice the increase in IRI and dwell time, which were both about 200 ms. This means that the dwell time did not fully accommodate the increase in RT1. Certainly, this is difficult to account for in a model that assumes that eye movements are triggered at a fixed point in processing. It is difficult to speculate about the reasons for the greater increase without more information about the source of the general elevation of RT1 seen in all our experiments. Thus, the present experiment was designed in part to investigate variables responsible for elevated RT1. In particular we examine the role of planning for a sequence of responses or fixations.

We also attempt to vary the central difficulty within a trial. One explanation for constant IRIs is that the earlier responses are delayed in order to be coordinated with stages in the processing of the subsequent response. It follows that, if no subsequent response is required, eye-hand spans should not be elevated. In the present experiment, we vary central difficulty using a Go/No-Go procedure. On each trial, only 2 or 3 positions contained target characters mapped with a key response. The rest were filled with dummy characters and participants were asked to skip them. We compare dwell time on Go and No-Go responses, and eye-hand spans on Go responses that are preceded and/or followed by No-Go responses to evaluate the impact of central difficulty.

Method

Participants Fourteen undergraduate students recruited from local colleges near NASA Ames Research Center participated in the experiment for course credits.

Apparatus The experiment was conducted using a PC with a 21-inch monitor. Participants were seated in a comfortable chair with their head secured on a head-and-chin rest placed 53.5 cm in front of the monitor. Eye movements were recorded with an infra-red video-based eye tracking system (ISCAN), which outputs data at a temporal resolution of 120 Hz and a spatial resolution of approximately 0.5˚ visual angle.

Stimuli and Display The primary stimulus display consisted of a row of five small characters (letters or symbols) spread over a wide viewing area. The characters were spaced equally (5.5˚ apart) and centered on the middle of the display. Each character subtended 0.34˚ in height and was presented at 11.7 cd/m2.

Design and Procedure Each trial began with the presentation of a white fixation cross (0.3˚) in the center of the display. After the participant had maintained fixation

within a 6˚ radius around the fixation for 500 ms, the fixation was erased and a small filled square (0.34˚) appeared at the leftmost stimulus position. Participants were instructed to move their eyes to fixate the small square when it appeared and maintain fixation at that location. The small square remained for 1 sec, followed by a blank interval of 500 ms. Then the five stimulus characters appeared simultaneously. Participants were asked to look at the characters one at a time, decide what they are, and make responses accordingly. Participants then pressed the spacebar to proceed to the next trial, which began following an inter-trial-interval of 250 ms.

There were six experimental conditions and two control conditions. Trials of different experimental conditions differed in the number of required successive responses in a sequence (one, two, and three), and in the stimulus position on which these sequences occurred (first and second). The six types of trials can be represented as the following: TXXTT, TTXXT, TTTXX, XTXXT, XTTXX, and XTTTX, with T denoting letter stimuli that required a key response (Go stimuli) and X denoting letter stimuli that required no response (No-Go stimuli). Go stimuli were randomly drawn from the letter set T, D, and Z, with the constraint that no letter was repeated in two adjacent positions. This constraint however does not prevent repetition of responses; the same letter could occur in two positions interposed by Xs. Five participants had 40 trials of each type administered in 2 blocks of 120 trials. Nine participants had 60 trials of each type administered in 3 blocks of 120 trials.

Trials in both of the two control conditions consisted of a single target (Go) stimulus in the first position (i.e., TXXXX), though different instructions were given for each. In the first condition, called Respond-Then-Scan (i.e., TXXXX), participants were asked to respond to the first letter stimulus, as before, and fixate each of the rest. In the second condition, called Respond-Only (i.e., T____), they were asked to respond as quickly as possible to the first stimulus only. There were 40 trials in each control condition. The two control conditions were administered after the experimental conditions and in the same order (Respond-Then-Scan first, Respond-Only second) to each participant.

No single aspect of task performance (e.g., manual or oculomotor, speed or accuracy, etc) was emphasized. The only specific instruction given to the participants was to treat each character independently and not group responses.

In all experiments eye movements were monitored and recorded. The recording of eye movements began at the moment when the small square appeared, and ended after the participant had responded to the rightmost stimulus. A calibration procedure was administered before each block of trials to maintain accuracy of recordings.

Results and Discussion

Figure 3 presents mean manual RTs and eye-hand spans as a function of stimulus. Cases where RT1 occurred to S1 (S1-RT1) are plotted separately from cases where RT1 occurred to S2 (S2-RT1). We discuss manual responses and eye fixations separately.

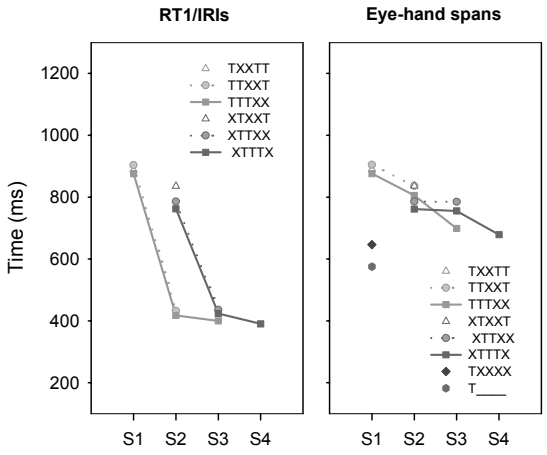


Figure 3. RT1/IRIs and eye-hand span results from the present experiment

Manual Responses The general pattern of elevated RT1 followed by rapid, flat IRIs is apparent in Figure 3. It is striking how closely aligned the curves for all stimulus conditions are. The only significant effect of the arrangement of stimulus was that RT1 was significantly slower when made to S1 (S1-RT1) than to S2 (S2-RT1). The general elevation of RT1 for both S1-RT1 and S2-RT1 suggests that cost is incurred for the first response in a sequence, not just to the first possible stimulus position. These similarities in patterns and magnitudes strongly suggest that the RT1/IRIs patterns are related closely to the production of sequences of responses. Indeed, the fact that RT1s for sequences such as "TTTXX" are equivalent to those for "TXXTT" is a strong indication that the difficulty of the next item has no effect on the current response. In other words, difficulty does not propagate backwards.

There are at least two possible explanations for the difference in RT1 between S1-RT1 and S2-RT1. It is consistent with at least some of the RT1 elevation being due to retrieval of stimulus-response mappings. If one assumes that the No-Go stimulus can elicit retrieval of response mapping for Go stimuli then that retrieval would have been done during S1 processing. This account is similar to accounts of first-trial cost in task switching studies (Logan & Bundesen, 2003). Alternatively, there is more uncertainty associated with S1 targets. If S1 is a non-target then S2 will always be a target. This reduction in uncertainty is a possible confound, though it is difficult to see how it would produce a speed up since the identity of S2 is not known until it is fixated.

Comparisons with the two control conditions provided evidence of sources contributing to the general RT1 elevation. RT1 was fastest (575 ms) in the Respond-Only condition, where participants were instructed to respond only to the first item and ignore the rest. RT1 was 72 ms slower (647 ms) in the Respond-Then-Scan condition, where participants were instructed to respond to the first item and fixate the others in turn. A plausible explanation for this overhead is that the elevated RT1 in the Respond-Then-Scan condition reflects a dual-task cost (cf. Pashler,

Carrie, & Hoffman, 1993), where the response task and the fixation task compete for a limited-capacity resource. There were no instructions as to how to perform manual and eye movement components; participants were free to do them concurrently or in sequence. We cannot say at present whether this overhead in combining the two behaviors, respond and fixate, arises from trying to do the two concurrently or would also be present with a strictly serial strategy. Some evidence suggests that trying to do the manual response concurrently with the fixation scan would cause interference. Pashler et al. observed interference between manual responses and voluntary eye movements in dual-task conditions, where participants were instructed to do both task as rapidly as they could. However, in their experiments substantial cost occurred only for anti-saccades, where subjects had to move away from a newly presented stimulus. A small cost obtained when moving to a specified color. Note that in both conditions the cost could reasonably be ascribed to a decision on the stimulus to determine where to move. In the present experiment the scan is fixed, making it difficult to see how stimulus decision processes could account for cost in the Respond-Then-Scan condition.

It is also hard to see why there should be a dual-task cost unless participants attempted to do the two tasks concurrently. Since there were no constraints or instructions on how to perform the task, any attempt to do them concurrently would have arisen naturally.

Another explanation might be that the Respond-Then-Scan condition forces participants to switch between tasks, resulting in a task-switching cost. However, task-switching costs are generally thought to arise from the retrieval of task-relevant knowledge, usually stimulus-response mappings. It is hard to explain how a switch cost would appear on S1 rather than on S2.

We prefer at present a more general explanation in terms of increased preparation time for the more complex behavior of Respond-Then-Scan. This account also helps explain why RT1 is further elevated in the full-response condition, with 2-3 targets. Here the preparation involves not only the sequencing of an initial response with a subsequent pattern of fixations, but of interleaving the responses.

There was one other significant RT1 effect whose meaning is not clear. S2-RT1 decreased significantly (from 837, 784, to 761 ms, *ps* < .05 based on pairwise *t* tests) as the number of required subsequent responses went up. This decrease was not observed for S1-RT1. It is hard to see how subsequent targets could facilitate a response to a current target. One argument is that the presence of a subsequent target could induce participants to rapidly complete the first item. The eyes fixated the next item prior to responding to the current one. If the next item is not a target they might decide to delay responding, and continue moving the eyes. If it is a target they know they must respond quickly and deal with the new item.

Dwell Time and Eye-Hand Span As in previous experiments, fixation durations remained relatively constant across stimuli. Not surprisingly, fixation durations on target (Go) stimuli were always longer than No-Go stimuli. More interesting comparisons arise when one regards fixation durations as a consequence of the previous stimulus (Figure 4). Here the dwell times suggest that the attempt to interleave the mental operations for successive stimuli pushes cost on to the subsequent stimulus. When the fixated stimulus is a target (a Go stimulus) dwell times were shorter by ~60 ms for targets that were preceded by dummy stimuli (i.e., XT) than by target stimuli (i.e., TT). When the fixated stimulus was a dummy stimulus this difference (TX compared to XX) was ~30 ms. This effect was found in several individual comparisons as well as in an analysis grouping all occurrences of each.

Condition	s1		s2		s3		s4	s5		T preceded by
										T or X
TXXTT	450.4	c	324.6	d	300.1	b	415.9	838.6	a	TT
TTXXT	467.6	a	478.8	c	315.1	d	307.9	768.9	b	XT
TTTXX	460.6	a	484.8	a	443.1	c	349.5	439.5		
XTXXT	339.9	b	394.4	c	344.0	d	329.9	776.8		X preceded by
XTTXX	332.7	b	409.5	a	445.5	c	370.5	527.7		T or X
XTTTX	336.6	b	395.7	a	433.3	a	473.3	581.8	c	TX
TXXXX	534.3		451.9		362.5		334.3	425.6	d	XX

Figure 4. Dwell times for S1-S5 in all conditions

Lengthened dwell times for stimuli preceded by targets suggest that the demand of making manual responses interfered with eye movement related processes. The eyes leave a stimulus prior to the completion of all the processing, such that the remaining processing for the previously fixated item delays one or more operations on the subsequent stimulus. A more detailed explanation rests on assumptions about the underlying resource architecture, which specifies the operations that can occur in parallel and those that must be done sequentially. The effect can be explained by adopting the common assumption that perceptual, cognitive, and motor operations execute in parallel, constrained only by logical or data dependencies. By this account, dwell times for the second stimulus are lengthened because cognitive resources required for stimulus-response mapping for the first stimulus postpone central processes on the second. Since central cognitive operations logically require data from perception, the inference is that this time is shorter by ~70 than that required for response selection. With continued explorations of similar factors it should be possible to obtain parameter estimates for processing operations that would permit a full model of extended task performance based on individual trial data.

Other aspects of the eye-hand span results resembled those found in previous experiments. As before, eye-hand spans decreased across the stimulus/response sequence. Figure 5 shows the results of the two constituents of eye-hand spans, dwell times and release-hand spans. There are several notable findings. First, it is evident that the difference among RT1s in cases where RT1 occurred to S2 was mainly due to the difference in release-hand spans. If one assumes that release-hand spans represent the time taken to complete remaining processes after fixation is terminated, it is foreseeable that release-hand spans may also include processes necessary for programming and coordinating response sequences. The fact that eye-hand spans decreased at a constant rate suggests that the coordination may not be restricted to each pair of responses. In the present set of experiments the maximal number of responses is set at five. It is possible that participants could

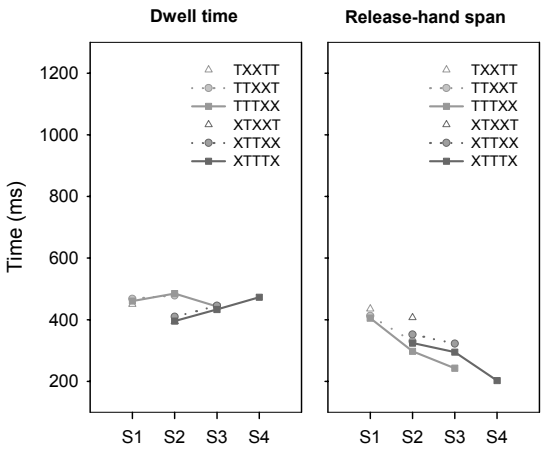


Figure5. Mean dwell time and release-hand span results

plan for five responses. Whether the trend will hold for longer sequences has yet to be tested.

General Discussion

The conditions of the experiment were designed to identify variables contributing to the elevation of RT1, and provide insight into the relationship of eye movements to manual responses by examining the effects of stimuli that required no manual response. Our previous experiments showed large eye-hand spans indicating that substantial processing remained on previous item after the eyes had moved. Analyses of dwell time responses to manipulations of stimulus-response compatibility suggested that dwell times encompassed central processes associated with response selection. Here dwell times for targets were elevated by ~70 ms when the preceding stimulus required a response. A straightforward account in terms of stage processing might estimate that the processing remaining after the eye movement is ~70 ms + the time for perceptual processing on the next task. Given a reasonable estimate of perceptual processing time of ~150 ms, it would seem that ~220 ms of central processing remain after the eyes move.

However, this explanation has difficulty accounting for the smaller increase (~30 ms) found on No-Go fixations in the same condition. That there is any effect of previous target at all is evidence that central processing is required to decide whether or not to respond to the No-Go stimulus. It might be assumed that the smaller effect for No-Go dwell times indicates more than postponement. That is, there may be interference between response-related processes on the two adjacent target stimuli. Since evidence for postponement is well known in dual-task studies, more evidence will be required to determine whether interference is acting here, rather than a more complicated postponement process.

Conclusions

We have evidence that RT1 elevation is due to a combination of factors including preparation for eye movement sequences, preparation for hand response sequences, and retrieval of stimulus-response mappings. Dwell times indicate that there is imperfect time sharing of the processing and response to successive stimuli. Mental operations on the previously fixated stimulus result in delays in processing the subsequent stimulus. These data can provide numeric estimates of internal processing times required to fully model these results.

Acknowledgements

This research was supported by funding from the Airspace Operations Systems (AOS) Project of NASA's Airspace Systems Program.

References

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. Human-Computer Interaction, 8, 237-309.

John, B., Vera, A., Matessa, M., Freed, M., & Remington, R.W. 2002. Automating CPM-GOMS. In Proceedings of the ACM SIGCHI 2002 Conference on Human Factors in Computing System: Changing Our World, Changing Ourselves (pp.147-154). New York: ACM.

Land, M. F., & Hayhoe, M. 2001. In what ways do eye movements contribute to everyday activities? Vision Research, 41, 3559-3565.

Land, M. F., & Lee, D. N. 1994. Where we look when we steer. Nature, 369, 742-744.

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task cuing procedure? Journal of Experimental Psychology: Human Perception and Performance, 29, 575-599.

Matessa, M., Vera, A., John, B., Remington, R. W., & Freed, M. (2002). Reusable templates in human performance modeling. In W. Gray & C. Schunn (Eds.), Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.

Pashler, H. 1994. Overlapping task operations in serial performance with preview. Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology 47, 161-191.

Pashler, H., Carrie, M. & Hoffman, J. (1993). Saccadic eye movements and dual-task interference. The Quarterly Journal of Experimental Psychology, 46A, 51-82.

Pelz, J., Hayhoe, M., & Loeber, R. 2001. The coordination of eye, head, and hand movements in natural task. Experimental Brain Research, 139, 266-277.

Salthouse, T. A., & Ellis, C. L. 1980. Determinants of eye-fixation duration. American Journal of Psychology, 93, 207-234.

Vickers, J. N. 1992. Gaze control in putting. Perception, 21, 117-132.

Wu, S.-C., & Remington, R. W. (2004). Coordination of component mental operations in a multiple-response task. In S. N. Spencer (Ed.), Proceedings of the Eye Tracking Research and Applications Symposium 2004. New York, NY: ACM SIGGRAPH.

Visual Analogy: Reexamining Analogy as a Constraint Satisfaction Problem

Patrick W. Yaner (yaner@cc.gatech.edu)
Ashok K. Goel (goel@cc.gatech.edu)
Artificial Intelligence Laboratory
College of Computing, Georgia Institute of Technology
Atlanta, GA 30332-0280

Abstract

Holyoak and Thagard proposed that the retrieval and mapping tasks of analogy can be viewed as constraint satisfaction problems, and described a connectionist implementation of their proposal. In this paper, we describe another constraint satisfaction method for the two tasks in the context of visual analogy: in our method, the source cases are organized in a discrimination tree, and all the source cases are searched *at once*. We also present an evaluation of the method for retrieval and mapping of 2-D line drawings from an external memory. The evaluation is based on structural constraints, and uses subgraph isomorphism as the similarity measure. One result is that a decomposition of the retrieval task into feature-based reminding and structure-based selection appears to provide little computational benefit over just selection.

Introduction

Holyoak and Thagard proposed that the retrieval (Thagard, Holyoak, Nelson, & Gochfeld, 1990) and mapping (Holyoak & Thagard, 1989) tasks of analogy can be productively viewed as constraint satisfaction problems. Their proposal incorporated structural, semantic and pragmatic constraints and used graph isomorphism as the primary similarity measure. Their mapping system, called ACME, and the complementary retrieval system, named ARCS, provided connectionist implementations of their proposal. In ACME, nodes are constructed for each map hypothesis (between a source element and a target element), with inhibitory and excitatory links between different nodes, and the network is run until it reaches quiescence. The work described here builds on Holyoak and Thagard's proposal but seeks a different solution to the retrieval and mapping tasks. While we also view the retrieval and mapping tasks as constraint satisfaction problems (CSPs), our method for addressing the tasks (i) organizes the source cases in a discrimination tree, (ii) uses (general-purpose) heuristics to guide the search, (iii) performs a backtracking search, and (iv) searches all the source cases *at once.*

The goal of our current work is to develop a computational theory of *visual* analogy. Analogies transfer relational knowledge from a source (or base) case to a target problem. Depending on the nature of the target and the source, the knowledge transferred in an analogy may pertain to different kinds of relations, for example, causal, functional or teleological relations. In visual analogy, the pertinent relations are spatial relations among visual elements. In a different part of the project, we have developed a technique for transfer of spatial knowledge, given a target problem and a source case and given a mapping between the two (Davies & Goel, 2001, 2003). In the part described in this paper, we focus on the retrieval and mapping tasks.

Our methodology is to start with simple problems and incrementally add complexity to them. This incremental nature of the methodology is manifested in three ways: firstly, visual knowledge can be of many forms, such as depictive bit-mapped representations, sketches, or animations, but our work deals specifically with diagrammatic knowledge represented symbolically as discrete geometric elements and the spatial relations between them; secondly, though visual analogies, like analogies more generally (as proposed by Holyoak and Thagard), can involve semantic and pragmatic constraints, we start with just the structural constraints imposed by requiring source and target to match structures; and thirdly, from a graph theoretic perspective, there may be more than one sort of graph isomorphism measure that may be the ideal measure, such as maximal common subgraph, but we begin with subgraph isomorphism as our metric.

The retrieval task, in this work, assumes a computer-based library of 2D line drawings, takes as input a query (target) in the form of a drawing (and no other information), and gives as output the source drawings that are most similar to the target. The mapping task takes as input a target problem and a source case, and gives as output correspondences between the basic elements of the source case and the target problem.

Retrieval

Following earlier work on analogical retrieval—e.g., MAC/FAC (Forbus, Gentner, & Law, 1995)—our retrieval architecture supports a two-stage process for diagram retrieval: reminding (or initial recall), and selection. The architecture consists of (up to) six basic components: an initial stage generating feature vectors, a process that generates a semantic network describing the contents (spatial structure in this case) of an drawing, a process that matches a target's description (semantic network) to source descriptions from memory, a working memory with potential sources to match with the target, and finally, an interface to the rest of the analogy system in which this retrieval would be taking place.

The reminding task takes as input a target example and returns as output references to stored drawings whose feature vectors match that of the target. The stored drawings are indexed by feature vectors describing their spatial elements; the feature vector for the target is constructed dynamically. References to those drawings with sufficiently similar feature vectors (according to some appropriate criteria, as explained below) are brought into the working memory. In the selec-

tion stage, the semantic networks of the drawings in working memory are matched with that of the target example. Drawings whose descriptions match the target description sufficiently well are collected and returned.

While the reminding stage of the retrieval process uses a vector of features—i.e. a vector of attribute-value pairs—as a heuristic to gauge the potential of a source drawing matching the target drawing, the selection task uses the spatial structure of line drawings—i.e. the qualitative arrangement of the various shapes in them—to actually match the target to the source drawings.

Visual cases are represented in three distinct ways: the the drawings themselves, the feature vectors, and the network of spatial relations. The representation of the drawings themselves is simply object-based: a list of each visual element, such as lines, triangles, etc., and their specific geometric properties (location, and so on). The feature vector is a multiset of the object and relation types contained in a semantic network. A multiset is a set that can contain more than one of each element (e.g. $\{2 \cdot A, 3 \cdot B, \ldots\}$). Given a semantic network describing an drawing, a feature vector in our system would look something like this: $\{3 \cdot \texttt{rectangle}, 2 \cdot \texttt{circle}, 3 \cdot \texttt{leftOf}, 1 \cdot \texttt{contains}, \ldots\}$.

A drawing is recalled, in the first stage, if the multiset of shape and relation types contained in it is a superset of that of the target. The method scans all stored drawings, calculating whether or not the multiset of objects and relations in the target is a subset of the multiset of objects in each source drawing, and returning those for which this is the case. That is, if Q is the feature vector for the target, and $S_1, S_2, \ldots, S_k$ are the feature vectors of the drawings currently in memory, then the method returns those drawings for which $Q \subseteq S_i$.

Figure 1 illustrates a simple 2D line drawing and its representation in terms of spatial relations in our system. The system at present recognizes four types of spatial elements: individual lines, triangles, rectangles, and ellipses (circles and squares are special cases of ellipses and rectangles, and are not treated as being of a separate type). Also, it presently recognizes five types of relations among the elements: `left-of`, `right-of`, `above`, `below`, and `contains`. The automatic generation of a semantic network for a target drawing works by taking the input drawing (in XFig format) and comparing every pair of shapes using the available predicates. If a particular predicate holds, a link is added between the associated nodes in the semantic network, with the appropriate label. As an example, the semantic network in Figure 1 would represent the drawing shown above it.

Memory Organization

When a source drawing is added to memory, several things happen. First, its description is generated, the network of relations describing the spatial layout of the drawing, as well as its feature vector. Second, once this network is generated, each "term" in the network, by which we mean a link (relation) together with its incident nodes (elements), is added to a discrimination tree. This allows the selection method to match individual terms in the target with all terms of the same form that appear across *all* source drawings in memory, thus allowing all of the descriptions of all of the drawings to be searched at once.

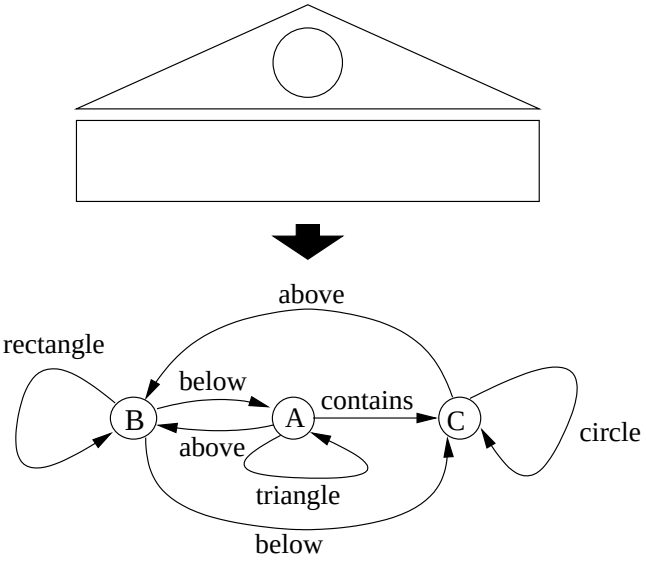


Figure 1: An example of a three-node semantic network in our language. Each pair of objects is tested, and links added for each relation that holds.

The selection method, described below, builds a set of potential assignments for each target element, and in evaluating these, it looks to see what terms each source element is involved in, and this involves the index into memory by individual terms. The overall scheme is to build a representation of all possible mappings, and reduce this list by screening out the ones that don't work. Ones that don't work are screened out because they do not satisfy the constraints imposed by the problem. This is constraint satisfaction, and this is what it means to solve the problem by constraint satisfaction.

Constraint Satisfaction

The core of the system is the selection process. The process finds a correspondence between the target drawing and the source drawings in working memory, eliminating drawings for which no correspondence can be found.

The selection problem is essentially one of matching objects (variables and constants) in the target and the source under the constraints imposed by the terms in which they appear. The target has a set of variables (its objects, the nodes in the semantic net) to be matched to some constants (i.e. values) from the sources and the relationships between these variables impose constraints on the values to which they can be matched. This is constraint satisfaction. This algorithm works by maintaining an index of all the terms across all of the source descriptions. It recalls individual terms from memory and puts them together to form the complete matching. When a source drawing is stored in memory, its description is generated and indexed in this way, by each term that appears in it. There is a separate table for each type of term, i.e. one for `left-of`, one for `above`, etc.

Treating the target elements as variables to be assigned values, the potential values are the nodes from the source descriptions in memory, all of which are considered at once. That is, the method is not performing a separate test on each

source in memory, but, rather, it is running a search procedure on the entire memory considered collectively. The constraints on the values assigned to the variables (the target nodes) are precisely those imposed by the subgraph isomorphism problem: if nodes A and B from the target are to be matched with nodes X and Y from memory, respectively, then, first, X and Y must be in the same description; second, all relations that hold between A and B must also hold between X and Y, respectively. If these constraints are met, then A can be matched with X and B can me matched with Y. Here the constraints are all either unary (say, A is a circle—a type constraint), or binary (say, A is left of B—a relational constraint). The only exception is the constraint that all values be from the same description, but this can be inferred from the binary constraints.

This matching process works in three phases: initialization of domains, reduction of domains, and finding the matching, where matching means subgraph isomorphism. The first phase initializes the target domains to sets of values that have the same incoming and outgoing edges. The second phase reduces these domains by eliminating values that are not all in the same drawing. These two phases reduce the selection of values for each variable. The third phase actually computes the isomorphism using constraint satisfaction and backtracking.

The first phase (initialize domains) works by finding nodes in memory that "look similar" to the target nodes: if a target node A is incident on, say, three links whose labels are R, S, and T, then the algorithm builds a list of all nodes in memory—across all the source descriptions—that have at least three incident links with labels R, S, and T. The second phase (reduce domains) works by ensuring that the set of source descriptions (document IDs) that are represented in the domain of (list of values for) each variable is the same. This serves to eliminate any value from the domain of any variable that does not come from a description represented in every other variable's domain.

These two stages are the "real" first stage of the algorithm, and our results, described in (Yaner & Goel, 2003), showed that the feature-vector-based first stage was really quite redundant, and offered little improvement. Viewed as such, this first stage applies two heuristics to the sources from memory: (i) prune any individual element (as opposed to entire drawings) that don't have the same "signature" (as just described) as the corresponding target element, and (ii) prune any terms whose associated drawings are not represented in every target element's domain. The latter one enforces subgraph isomorphism. It is important to note that these are both logically implied by the similarity metric that the last phase, described below, implements. It would be an interesting experiment to look at other heuristics that prune out mappings that might have otherwise been returned by the last phase.

The last phase (find matchings) is the one that actually does the work. The basic procedure is one that generates matchings, checking them for consistency as it goes, and backtracking when necessary. The test, here, is actual subgraph isomorphism: if A is related to B in the target, then the relations (links, edges) between $m(A)$ and $m(B)$ must include at least those that held between A and B, where $m(*)$ is a mapping from target to source. This algorithm returns all valid mappings. The idea is that the first two phases have restricted the set of possible mappings so that there aren't nearly as many, now, as there would have been if a pure depth-first search had been done.

In general, the time complexity of depth-first search, such as this is, is on the order of $O(k^d)$ in the worst case, where k is the branching factor of the state space and d is the maximum depth. In this case the depth is the number of elements in the query, and the branching factor is the number of elements across all sources in memory. However, the space complexity, as with depth-first search in general, is only $O(kd)$, i.e. it's linear in the size of the problem. Note that this is a backtracking search, however, so large portions of the state space are cut off at each step. With 42 test images in memory, the number of objects in a drawing ranging from 3 to over 50 (the average was about 12), the number of terms in the description ranged from a couple of dozen to over eight thousand. There were 21 queries with this test set, with two to five spatial elements in each, and up to several dozen terms. With this test data, the system was retrieving drawings in about 9.32 seconds on average (across all 21 drawings), doing an average of about 1.49 million memory accesses (to the index of terms across all the drawings) per retrieval.

Galatea

Since the system does retrieval essentially by producing all possible mappings that it is capable of finding, we adapted a version of the system to the mapping process for use in a system called *Proteus*, a visual analogical reasoning system. The transfer stage of *Proteus*—implemented in a system called *Galatea*—is described in Davies and Goel (2001, 2003).

Galatea solves problems represented in a high-level visual language called Covlan (Cognitive Visual Language). The system solves these problems by analogy to existing problems whose solutions are mapped out as a sequence of transformations on the knowledge states that are represented in this language. *Galatea* solves the problem by taking a mapping between the initial knowledge states of the source and target and mapping the transformations and generating the intermediate knowledge states (and mappings between them), and thereby constructing the rest of the transformations and knowledge states leading to the solution to the target problem. The mapping system, then, needs to connect the initial knowledge state of the source and the target drawing. From the perspective of retrieval and mapping, the relevant issues pertaining to *Galatea* are: (1) what is that knowledge representation, and (2) what are the nature is of the required mappings?

Covlan consists of knowledge states, primitive elements, primitive relations, primitive transformations, general visual concepts, and correspondence and transform representations. In Covlan, all knowledge is represented as propositions. In this paper we will only be concerned with the primitive elements and the primitive visual relations. The primitive elements are `polygon`, `rectangle`, `triangle`, `ellipse`, `circle`, `arrow`, `line`, `point`, `curve`, and `text`. There is also a `set` element type, with members that have `in-set` relations back to the set they are members of, though these do not correspond to visible entities—this is purely for grouping purposes. Each element is represented as a frame with attribute slots such as `location`, `size`, `orientation`, and

`thickness`, but these attributes will not concern us, since mappings between attribute values are not part of the required mappings, and thus representing them in the semantic network is not necessary.

Primitive visual relations represented are touching, `above-below`, `right-of-left-of`, `in-front-of-behind`, and `off-s-image`. A typical knowledge state is represented with a node corresponding to that knowledge state (e.g. `L14-simage1`), and elements (which may be sets) are represented with `contains-object` relations from the knowledge state element to the visual elements themselves.

We describe next some example problems originally designed for *Galatea*. The first example problem is a fairly simple one: dividing a pizza into some number of slices based on analogy to the problem of dividing up a cake into some number of pieces. In this case, there is a cake (or pizza), and a set of people in the initial problem state. Set members are not mapped, and the division is made in transformations in later problem states, so the only possible mappings are cake to pizza and set of people to set of people, or cake to set of people and set of people to pizza. The problem, as represented in *Galatea*, does not contain any visual relations between the set of people and the cake (or pizza), and thus there is nothing constraining the mapping to be the "correct" mapping. The latter mapping will probably lead to failure in the transfer stage, but both are returned by our system.

A more complex and interesting example is based on Gick and Holyoak's fortress/tumor problem (1980). In this problem, we have an army attacking a fortress over mined roads, and the general decides to split his army to avoid setting off the mines, and a target case in which there is a patient with a tumor and a doctor who wants to kill the tumor with radiation. The supposed analogy is to split the beam (somehow) to avoid killing the healthy tissue that is in the way. The visual representation of these problems has a fortress (and a tumor represented similarly) and four roads (sections of the body surrounding the tumor), and an army represented by an arrow (a ray of radiation represented similarly). The "correct" analogy maps the set of roads to the set of body parts, the fortress to the tumor, and the army to the ray. However, there being three of each thing to match, and the particular representation chosen not using the visual relations (though it could have), there was nothing constraining the mapping, and all six possible correspondences were returned. Had visual relations constrained it, the number of possible mappings would have been smaller.

Mapping

Galatea has set up the requirements for the mapping task such that only visual elements are to be mapped, not attribute values, and so attribute values (which can be represented as propositions, and hence can be represented in a semantic network) are not included in the input to the mapping system. In addition, members of sets are (generally) not to be mapped, and so any visual element on the left-hand side of an in-set relation can be pruned from the mapping system's input, as well. With these two constraints, the mapping system was run on several sample problems, two of which were described above. Four other problems of similar nature and size were also run on this system.

function GENERATEMAPPINGS

1: *sourceRels* ← first simage from source problem
2: *targetRels* ← target problem simage
3: *sRelLables* ← names of all relations represented in *sourceRels*
4: *tRels* ← remove from *targetRels* all relations that don't match one in *sRelLabels* and all relations involving a literal (i.e. attribute-value pairs)
5: *tRelLabels* ← names of all relations represented in *tRels*
6: *sRels* ← remove from *sourceRels* all relations that don't match one in *tRelLabels* and all relations involving a literal (i.e. attribute-value pairs)
7: *sNodes* ← list of all nodes (elements) from *sRels*
8: *tNodes* ← list of all nodes (elements) from *tRels*
9: *domains* ← GENERATEDOMAINS(*sNodes*, LENGTH(*tNodes*))
10: *rDomains* ← GENERATEDOMAINS(*tNodes*, LENGTH(*sNodes*))
11: $fMappings$ ← FINDPROJECTIONS(*sRels*, *tNodes*, *tRels*, *domains*)
12: *rMappings* ← FINDPROJECTIONS(*tRels*, *sNodes*, *sRels*, *rDomains*)
13: *rMappings* ← reverse each of the mappings returned in *rMappings* so that they map source onto target properly instead of target onto source
14: **return** $fMappings \bigcup rMappings$

Algorithm 1: Generate Mappings

The mapping algorithm (see Algorithm 1) works as follows: the outer procedure (generate mappings) first retrieves the named source and target representations from memory, then applies the above heuristics to it, and finally generates the mappings and returns them. Since it computes subgraph isomorphism, as above, we run it both ways—attempting to map source onto target, and also attempting to map target onto source and reversing the returned mappings. Thus it is possible to find the target within the source or vice versa, finding the source within the target. The algorithm for FINDPROJECTIONS is identical with that of the third phase, "find matchings", above.

It's important to note that this does not actually solve the mapping problem; it particular, it returns *all* mappings, so that an additional search or evaluation stage is necessary to find the relevent ones. This is where pragmatic and semantic constraints may start to enter back into the picture. Our work to date has begun with only structural constraints as an experiment, and we plan to reintroduce other constraints as the larger problem context is reintroduced.

At any rate, the cake/pizza example described above, when run through this system, came up with two mappings: one that maps the cake to the pizza and the set of people to the set of people, and one that maps the cake to the set of people and the other set of people to the pizza:

Cake *maps-to* Pizza
Set12 *maps-to* Set14

Cake *maps-to* Set14
Set12 *maps-to* Pizza

Set12 is the set of people in first cake problem knowledge state, and Set14 is the set of people in the first pizza problem knowledge state. *Proteus*, recall, does not map members of sets, and so the individual people are not mapped onto each other, only the sets. The first one, obviously, is the "correct" one, the one that would lead to a successful transfer and evaluation of the problem solution.

The fortress/tumor problem was more interesting. The heuristics pruned out the set of roads and body parts, as well as the shapes and sizes and positions of all the elements, and so the only details left to influence the mappings were the fact that the elements were part of the problem. There were three elements, thus, remaining, for each one: Fortress and Tumor, Soldier-Path and Ray, and Set1 (the set of roads) and Set2 (the set of body parts surrounding the tumor), and six mappings produced:

Fortress *maps-to* Tumor
Soldier-Path *maps-to* Ray
Set1 *maps-to* Set2

Fortress *maps-to* Tumor
Soldier-Path *maps-to* Set2
Set1 *maps-to* Ray

Fortress *maps-to* Ray
Soldier-Path *maps-to* Tumor
Set1 *maps-to* Set2

Fortress *maps-to* Ray
Soldier-Path *maps-to* Set2
Set1 *maps-to* Tumor

Fortress *maps-to* Set2
Soldier-Path *maps-to* Tumor
Set1 *maps-to* Ray

Fortress *maps-to* Set2
Soldier-Path *maps-to* Ray
Set1 *maps-to* Tumor

Now, this really represents all correspondences between three things and three things. The primary reason for this is that the representation chosen for this particular problem does not involve any reference-frame relations such as `left-of` or `right-of`. If it had, these relations would constrain the mappings.

Discussion

In the introduction, we mentioned Holyoak and Thagard's ACME system (1989) and noted the similarities and differences between our work and theirs. ANALOGY (Evans, 1968) was an even earlier AI program that performed the task of finding similarities and differences between visual cases. It performed simple geometric analogies of the kind that appear on many intelligence tests. Let us suppose that each of A, B, C, D, E and F is an arrangement of simple geometric objects, e.g., a small triangle inside a large triangle, a small circle inside a larger circle, etc. Given an analogy A:B, and given C and multiple choices D, E and F, ANALOGY found which of D, E, and F had a relationship with C analogous to that between A and B. It represented the objects and the spatial relationships between them in the form of semantic networks, which enabled it to compare the spatial structure of the various arrangements. However, since ANALOGY performed an exhaustive and linear search of the mappings, its method cannot scale up to any realistic problem.

While ANALOGY was an early program that matched symbolic descriptions of two drawings and found similarities and differences between the drawings, MAGI (Ferguson, 2000) and JUXTA (Ferguson & Forbus, 1998) are two recent systems that find mappings between symbolic representations of two drawings (or two portions of the same drawing). These systems use truth maintenance as the mechanism for keeping track of new constraints and retracting old conclusions.

Our decomposition of the retrieval task into feature-based reminding and structure-based selection is similar to that of MAC/FAC (Forbus et al., 1995). The similarity is specially striking because in its current stage ours deals only with structural constraints; as noted in the introduction, we plan to explore and exploit semantic and pragmatic constraints in the next stage. However, in contrast to MAC/FAC, the experiments described in (Yaner & Goel, 2003) indicate that the two-stage decomposition of the retrieval task provides little computational benefit over just one-stage retrieval based on structure-based selection.

In computer-aided design, FABEL (Gebhardt, Voß, Gräther, & Schmidt-Belz, 1997) was an early project to explore the automated reuse of diagrammatic cases. In particular, TOPO (Börner, Eberhard, Tammer, & Coulon, 1996), a subsystem of FABEL, used the maximum common subgraph (MCS) of the target drawing with the stored drawings for retrieve similar drawings. Gross and Do (1995) describe a method for retrieving designs that contain a given design pattern in the domain of architectural design. Gross and Do's heuristic method is very simple: given two drawings, it compares the type and number of spatial elements and the spatial relations by counting. Their method is roughly equivalent to the first stage in the two-stage retrieval process.

In computer vision, Grimson and Huttenlocher (1991) developed a similar method for object recognition. They begin with a model with a set of features, such as a set of potential edges in some arrangement, and sensor data with a set of sensor features (edges, vertices, etc.); a lot of sensor features might be noise. The task is to find a set of sensor features that comes from one (and the same) object. Their method matches model features to sensor features under some transformations within specific limit of tolerance. The model imposes constraints, for instance, by its arrangement of features. Although they do not describe it as constraint satisfaction, their method in fact is in assigning values to variables under unary and binary constraints imposed by the arrangement by using a backtracking depth-first search.

Constraint satisfaction methods have become common in AI: Prosser (1993) describes methods of constraint satisfaction with backtracking; and Bayardo and Schrag (1997) provide evidence of applicability of constraint satisfaction with backtracking for real-world intractable problems in planning and scheduling. Our method of constraint satisfaction with

backtracking, with the case memory organized into discrimination trees, builds on the work of Ounis and Paşca (1998). They view the general problem of associative image retrieval as one of computing projections over conceptual graphs representing their content. Although they do not describe it as a constraint satisfaction method, their algorithm, in fact, is doing constraint satisfaction to compute the projection. However, their method is limited to constraint satisfaction with generate and test with no backtracking.

Conclusions

We have described a constraint satisfacton method for the retrieval and mapping tasks of analogy. Our method (i) organizes the source cases in a discrimination tree, (ii) uses (general-purpose) heurstics to guide the search, (iii) backtracks (if and when needed), and (iv) searches all the source cases at once. We also presented an evaluation of the method for the retrieval and mapping of diagrams from an external memory.

Our laboratory-scale experiments, with drawings containing only up to fifty spatial elements and their representations containing only up to eight thousand terms, indicate that the method of constraint satisfaction is fast and appears quite promising for use in practice. On the one hand, we fully expect that the complexity of the task will significantly worsen for larger drawings and larger libraries of drawings, but, on the other, we also expect that it should be possible to develop significantly faster methods for the task. For example, we expect that use of spatial aggregations and abstractions to organize the representation of the spatial structure of a drawing in the form of a linked hierarchy of semantic networks would partition the search space performance especially for large, complex drawings (e.g. Papadias, Kalnis, & Mamoulis, 1999). In addition, more sophisticated constraint satisfaction techniques such as forward checking and intelligent variable ordering, to name just a couple of common ones, can be brought to bear on the problem as well, taking advantage of structure in the knowledge representation and the search space.

References

Bayardo, R. J., Jr., & Schrag, R. (1997). Using CSP look-back techniques to solve real-world SAT instances. In *Proc. AAAI-97* (pp. 203–208). Providence, Rhode Island: AAAI Press.

Börner, K., Eberhard, P., Tammer, E.-C., & Coulon, C.-H. (1996). Structural similarity and adaptation. In I. Smith & B. Faltings (Eds.), *Advances in Cased-Based Reasoning: Proc. 3rd European Workshop on Cased-Based Reasoning* (Vol. 1168, pp. 58–75). Lausanne, Switzerland: Springer-Verlag.

Davies, J., & Goel, A. K. (2001). Visual analogy in problem solving. In *Proc. IJCAI-01* (pp. 377–382). Seattle, WA: Morgan Kaufmann Publishers.

Davies, J., & Goel, A. K. (2003). Representation issues in visual analogy. In *Proc. 25th Annual Conf. Cognitive Science Society.* Boston, MA: Lawrence Erlbaum Associates.

Evans, T. G. (1968). A heuristic program to solve geometric analogy problems. In M. Minsky (Ed.), *Semantic Information Processing.* Cambridge, MA: MIT Press.

Ferguson, R. W. (2000). Modeling orientation effects in symmetry detection: The role of visual structure. In L. R. Gleitman & A. K. Josh (Eds.), *Proc. 22nd Annual Conf. Cognitive Science Society.* Philadelphia, PA: Lawrence Erlbaum Associates.

Ferguson, R. W., & Forbus, K. D. (1998). Telling juxtapositions: Using repetition and alignable difference in diagram understanding. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in Analogy Research* (pp. 109–117). Sofia, Bulgaria: New Bulgarian University.

Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*(2), 141–205.

Gebhardt, F., Voß, A., Gräther, W., & Schmidt-Belz, B. (1997). *Reasoning with complex cases* (Vol. 393). Boston: Kluwer Academic Publishers.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Grimson, W. E. L., & Huttenlocher, D. P. (1991). On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(12), 1201–1213.

Gross, M. D., & Do, E. Y.-L. (1995). Diagram query and image retrieval in design. In *Proc. 2nd Int'l Conf. on Image Processing.* Crystal City, VA: IEEE Computer Society Press.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*(3), 295–355.

Ounis, I., & Paşca, M. (1998). RELIEF: Combining expressiveness and rapidity into a single system. In *Proc. 21st Annual ACM SIGIR Conference* (p. 266-274). Melbourne, Australia: ACM Press.

Papadias, D., Kalnis, P., & Mamoulis, N. (1999). Hierarchical constraint satisfaction in spatial databases. In *Proc. aaai-99.* Orlando, FL: AAAI Press.

Prosser, P. (1993). Hybrid algorithms for the constraint satisfaction problem. *Computational Intelligence*, *9*(3), 268-299.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constrain satisfaction. *Artificial Intelligence*, *46*, 259–310.

Yaner, P. W., & Goel, A. K. (2003). Visual case-based reasoning I: Memory and retrieval. In *Proc. IICAI-03.* Hyperbad, India: Springer-Verlag.

Cognitive processes of artistic creation: A field study of a traditional Chinese ink painter's drawing process

Sawako Yokochi (b031214d@mbox.nagoya-u.ac.jp)
Graduate School of Education and Human Development
Nagoya University, Furo-cho, Chikusa-ku
Nagoya, Japan 464-8601

Takeshi Okada (j46006a@nucc.cc.nagoya-u.ac.jp)
Institute for Advanced Research and
Graduate School of Education and Human Development

Abstract

How are art works created by artists? In this study, we focused on the drawing processes of a Chinese ink painter through a field study and a field experiment. In the field study, we observed processes of fusuma drawing in a temple, and in the field experiment, we asked the painter to draw sixteen pictures (eight drawings in BLANK condition and eight drawings LINES condition). We analyzed those drawing processes and found that: (1) this artist seems to gradually form a global image of the drawing as he draws each part one by one; (2) lines that the audience drew seem to create new constraints for his drawing and force him to create new patterns; and (3) moving his brush in the air before actually drawing lines on the paper seems to serve one of the following functions: Positioning (where to draw), rehearsal (how to draw), and image generation (what to draw).

Introduction

It is widely believed that only talented people can create great works of art. Despite this, the psychology of creativity has demonstrated that ordinary cognitive processes underlie the emergence of images or concepts (Weisberg, 1993). Although cognitive psychology ought to be interested in such creative cognitive processes, few empirical studies on the artistic creative process have been conducted (Getzels & Csikszentmihalyi, 1976). Among the few studies that have been conducted, some are relatively old and pre-date the information processing revolution that has occurred in the field (e.g., Eindhoven & Vinack, 1952). More recently, studies have used techniques such as interviewing in order to understand the creative process but have neglected on line methods (e.g., Mace & Ward, 2002; Cawelti, Rappaport & Wood, 1992). Despite these efforts, creative cognitive processes are not yet well understood. At this early stage of cognitive study on artistic creation, it seems that multi-method approaches are most appropriate. For example, Getzels & Csikszentmihalyi (1976) have approached creativity from several perspectives by using several test batteries, such as IQ tests, creativity tests, personality tests, and observations and interviews of art-making processes.

Viewing the state of creative study, in the present study, we try to answer the question, "How does a painter create his/her works?" We offer a case study based on observations, interviews, and a field experiment with detailed cognitive analyses of the drawing processes of a *Suibokuga* (Chinese ink painting) painter.

Method

Subject: Mr. K. is a *Suibokuga* painter in his early 60's with about 18 years of experience of painting in that style. He usually draws *Sansuiga*, which are traditional Chinese landscapes of mountains and valleys, on fusuma (Japanese sliding doors) or folding screens in temples and shrines. He has also exhibited his works at museums in the USA and France in addition to many places in Japan. He has a special style of drawing. He improvises his drawing in front of audiences by incorporating random lines that the audience drew onto blank paper.

Period of observation: This field study was conducted from 1998 to 2001, with a follow-up interview conducted in 2003. We observed his drawing processes and collected substantial on-line data about his drawing. Also, we investigated his drawing processes through conducting a field experiment.

Data described in this paper: In this paper, we focus on the following two data sets in this field study: (1) process data of a fusuma drawing in temple X; and (2) data from a field experiment.

In the temple, spending about one and a half-hour, the painter drew a picture of mountain and river across four fusuma sliding doors. We set up two video cameras from both sides of the fusuma doors to capture his drawing process. After he finished his drawing, we interviewed him about his drawing process. In this case, he did not ask the audience to draw random lines because the master of the temple asked him not to do so.

In the field experiment, we asked him to draw eight pictures created from fifteen random lines drawn by two experimenters (we call this the LINES condition) and eight pictures created on blank paper (we call this the BLANK condition). The themes of the paintings are the four seasons. We asked him to draw two pictures of each season in each condition: spring; summer; fall; and winter. The order of task presentation was counter-balanced by condition. The order of the season for each task was randomized. We recorded the processes of his drawing with two video cameras. He drew three or four pictures in his studio in a day. It took a total five days between June and December to complete the field experiment. Usually it took about 20 to 30 minutes for him to finish a picture. In the third day of the experiment, he reported that he

could not concentrate on drawing and drew just one picture. In the second day of the experiment he thought a picture in the BLANK condition was not good enough. Therefore he drew another picture with the same theme once more in the final day.

Goal of this study: This study describes the drawing process of a *Suibokuga* painter through a field study. Unlike laboratory experiments, field studies can be problematic with regarding to variable control. In addition, since this is a single case study, we also cannot generalize our findings to all artists. However, through field studies such as this, we can propose new hypotheses or offer useful insights with high levels of ecological validity. Especially, in domains where few previous studies exist, starting from field studies can be very useful in order to find important questions and hypotheses and to lead to further research projects that follow realistic and meaningful directions.

Results and Discussion

The following three main features were identified through our field study;

(1) The painter seems to form a global image of the drawing gradually as he draws each part one by one;
(2) The painter draws pictures in fairly patterned ways. Lines that the audience drew, however, seem to create new constraints for his drawing and force him to create new patterns;
(3) The painter often moves his brush in the air before actually drawing lines on the paper. Based on our data analyses, we describe three possible functions of these movements.

Processes of Drawing Images

Mr. K draws his paintings very smoothly and quickly. Although it might look as if he had already formed an image of the entire picture before starting to draw, our analyses of the drawing process and an interview with him revealed that he starts drawing with a local image of the picture. Then, he gradually forms a global image as he draws each part one by one.

When we interviewed him just after he finished drawing fusuma doors in the temple, he said, " Not the entire picture. Starting from here, the pine tree that I first drew, then there and this bridge and here, then the cedar trees above the stairway. Then the roof of the hat. I had an image of only those parts at the beginning"(See Figure 1). It seems that he does not form the entire image before he starts drawing. How can he draw so smoothly without forming the whole image or complete plans in his mind before starting to draw? We analyzed his drawing processes in detail to answer this question.

Figure 2 shows the process of his drawing on the fusuma doors of the temple. The circled numbers on the fusuma doors indicate where and in what order he drew. The circled numbers on the tatami mats indicate where and in what order he moved. We divided the process into five sections based on his movements. The first four sections were segmented when he moved backward to survey the entire picture for more than one minute. The rest of his drawing processes were combined into one section, because he moved backward and forward very often without long pauses. In the first section, he sat on a tatami mat and started drawing a tree on the left-most part of the fusuma door. After he drew the central part of the left fusuma doors for about 22 minutes, he stepped back in order to see the entire picture. Then he started drawing on the second door from the right and paused to observe what he drew many times. When this part of the picture became more formed, he moved backward and looked at the picture occasionally. At almost the end of his drawing in the last section, he moved back and forth frequently, adding a few lines here and there. This analysis of his drawing processes and his interview in the temple suggests that he gradually formed his plans for the painting while he was drawing. Although this is a single case analysis, we observed that he drew the fusuma doors in this way on many other occasions.

Mr. K cannot look at the entire picture without stepping backward when he draws on such big fusuma doors. Although he can take in the entire picture when he draws on a small-sized paper, he still has to spend a certain amount of time planning and monitoring when he draws, even though he can see the entire picture at a glance. Therefore, we measured the duration and timing of pauses in the data from the field experiment in order to infer his planning and monitoring process as while drawing. We divided drawing processes into small cycles. One cycle consisted of the period from his soaking the brush in the sumi ink plate, lifting up, drawing on the paper, and soaking it in the ink plate again. We counted the distribution of pauses by length and found that the frequency drastically dropped above nine seconds. This suggests that there might be some functional difference in pauses shorter than nine seconds and those longer than nine seconds. The frequent occurrence of the shorter pauses probably indicates that he moves the brush from one place to another or ink plate, etc. And, the less frequent occurrence of the pauses longer than nine

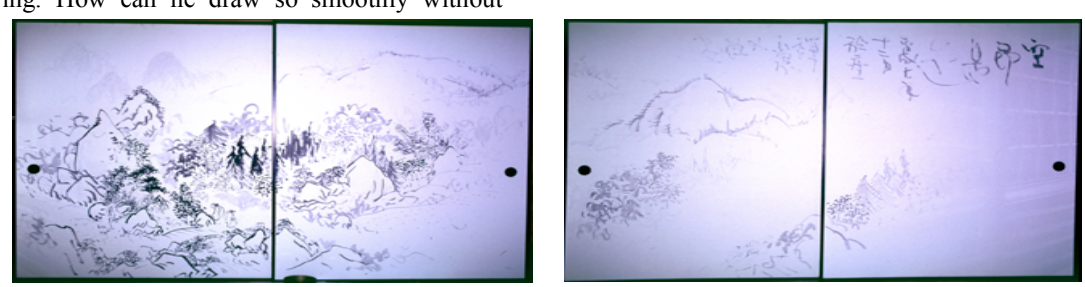
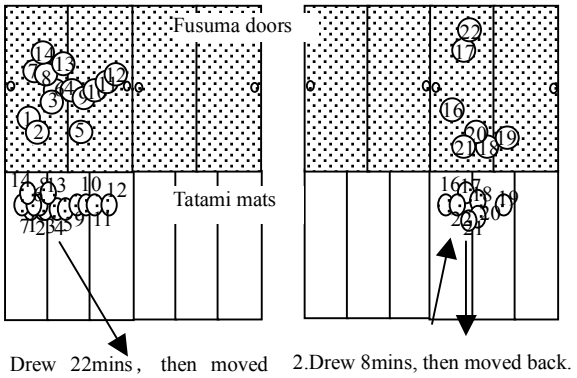


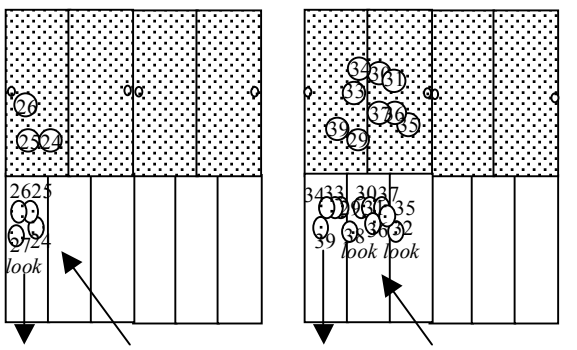
Figure 1: Picture on fusuma doors at temple X

seconds would mean that he spent time thinking about the



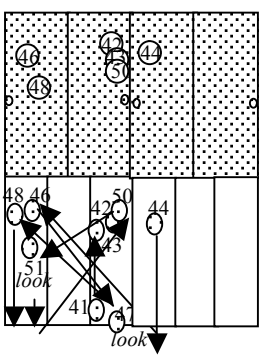
1. Drew 22mins, then moved back.

2.Drew 8mins, then moved back.



3.Drew 5mins, then moved back.

4. After intermission, drew 11mins, then moved back.



5.Moved back and forth (5mins.).

Figure2: Processes of drawing on the fusuma doors in temple X

pictures, planning and monitoring his drawing processes[1].

Table 1 shows the data from the field experiment. When counting pauses equal to or longer than nine seconds, we found that there was about the almost same number of pauses in the first and the second half of his drawing in each condition, t (7) = -2.37, p = .050, t (7) = -1.80, p = .862 (See Table 1).

Table 1: Mean number of pauses (nine or more seconds) during drawing

	Whole drawing	First half: Second half	Before drawin with lines
BLANK condition	5.0	1.4 : 2.4	
LINES condition	11.4	5.6 : 5.8	4.5

This suggests that he plans and monitors his drawing through the entire process of drawing. There were more pauses in the LINES condition than in the BLANK condition, F (1, 7) = 19.166, p = .003. When we focused on the frequency of pauses just before he drew from random lines, we saw about the same frequency of the pauses as a difference between each condition, F (1,7) = 3.163, p = .119. This probably means that he needs to think about local drawing plans in order to incorporate those random lines into his picture when he creates pictures from random lines.

In summary, it appears that the painter plans and monitors through the entire process of drawing. He first forms a mental image of a small area (creates a local drawing plan), and gradually forms the entire mental image of the picture as he draws each object.

Lines as Constraints

Analyses of the contents and patterns of Mr. K's drawing suggest that he drew pictures in a fairly patterned way. Through our observation, we found that he drew objects one by one. In the field experiment, he started to draw his paintings from a tree in fifteen out of the sixteen pictures. Then rocks, houses, people and mountains followed. We observed in many other occasions that he drew pictures in the same way. It suggests that he uses some strategies in order to draw certain objects in a relatively stable order in various situations. However, when we interviewed him, he said, "All of the pictures that I created from random lines are more unique and nicer than those created in a traditional way." What kind of difference is there between both conditions? We investigated the differences in time of drawing and the number of drawing cycles between pictures in the LINES condition and pictures in the BLANK condition (See Figure 3 and Table 2).

First, the mean time of drawing (except for the time of painting shadows or shading ink lines which always occurs at the end of his drawings) was calculated in each condition. In the BLANK condition, the mean time of drawing was about ten minutes (M = 640.13 sec, SD = 170.91 sec), and, in the LINES condition, it was about eighteen minutes (M = 1050.38 sec, SD = 199.40 sec). The time of drawing in the LINES condition was significantly longer than the time of drawing in the BLANK condition, t (14) = 3.87, p < .01. We also counted the number of drawing cycles in each condition and calculated the mean number. The mean number of drawing cycles in the LINES condition was significantly higher than that in the BLANK condition, t (14) = 3.91, p

[1] Our criterion gains plausibility from experiments in previous studies. For example, Chase & Simon (1973), with perception and memory tasks in chess, presumed the long time interval while glancing chess pieces placed on the board was needed to combine several chunks, and the short time interval to access to a single chunk. Thus, it is also reasonable to think that the difference of time interval reflects upon the processes of thinking during drawing.

Table2: Differences between the BLANK condition and the LINES condition

Measures	BLANK condition Means and (*SD*s)	LINES condition Means and (*SD*s)	*p* of *t* Tests
Time of drawing(sec)	640.13 (170.91)	1050.38 (199.40)	< .01
Number of drawing cycles	30.0 (8.80)	43.5 (4.92)	< .01
Time of one cycle(sec)	23.3 (7.50)	25.5 (6.85)	n.s.

< .01. These results indicate that it takes more time and more drawing cycles to create new pictures from random lines.

This result suggests that these lines somehow influenced drawing. Therefore, we investigated how these lines were used in his drawing. There were fifteen random lines drawn by the experimenters on each paper in the LINES condition. With an average of 9.3 out of fifteen lines, he would create new object starting from others lines. In the other 5.7 instances he incorporated the other's line into an existing object. Thus, the random lines triggered his drawing process and created new constraints on his drawing.

There seemed to be some differences in terms of quality between pictures in the LINES and BLANK conditions. To check this possibility, we asked twenty undergraduate students to rate their impressions of the paintings using a semantic differential method.

The procedure is as follows: Twenty undergraduates who did not major in art were presented pictures randomly with twelve word pairs of opposite meaning as a paper and pencil task. All words were adapted from adjectives used in the study of emotions when appreciating pictures (Ichihara, 1968) and interviews of the painter. Subjects were asked to rate the pictures based on a seven-point scale for each word pair.

Factor analysis with a principal factor solution was used to create scales across the word pair items. The three distinct factors with an eigenvalue above 1.0 were recovered and the ratio of variance contribution was 65%. These factors were rotated with Varimax and the factor loading was calculated (See Table3).

Four items are strongly correlated with the first factor, which we term *good composition*: modulated / non-modulated; well composed / poorly composed; focused / unfocused; and well-balanced / ill-balanced (*alpha* = .82). The second factor, which we term *liveliness*, is strongly correlated with the items: lively / dull; static / dynamic; energetic / non-energetic; and powerful / power less (*alpha* = .77). The final factor, which we term *simplicity*, strongly correlated with the items: clear cut / mixed up; simple / complex; relaxed / crowded; and light / heavy (*alpha* = .73).

We conducted a single-sample version of Hotelling's T^2 to compare their rating scores of paintings from the two conditions (See Figure4).

Figure 3: Picture in the BLANK condition (top) and picture in the LINES condition (bottom)

The mean scores of *good composition* and *simplicity* in the BLANK condition were significantly higher than those in the LINES condition, *F*s (1, 159) = 93.838 and 28.479, respectively, *p*s < .001. This result indicates that pictures in the BLANK condition are well composed. Also, because there is fair amount of white space in these pictures, it creates the impression of simple picture. The painter draws the BLANK pictures with the style of traditional *Sansuiga* paintings. On the other hand, the mean score of *liveliness* in the LINES condition was higher than that in the BLANK condition, *F* (1, 159) = 4.153, *p* < .05. This result indicates that pictures in the LINES condition were characterized by liveliness and were dynamic. Thus, the character of LINES pictures is different from traditional *Sansuiga* paintings.

Mr. K also thinks that this way of drawing is more exciting than the traditional way. When we interviewed him asking why he wanted to draw from random lines, he answered:

> "Creating from random lines, I have to incorporate the others' world into my world... I have to use them with my lines...Seriousness! I enjoy playing this game in earnest. There is not just myself. I get serious about drawing in this way. Yes. I am highly motivated with this way."

Thus, these lines seem to create new constraints for his drawing and force him to create new patterns.

Roles of Hand Movements in Drawing Processes

From our observations in the field studies, we noticed that the painter moved his brush in the air very often before he actually drew lines on paper. We wondered why he did so.

This kind of hand movement is not unique to this painter. For example, Henry Matisse moved his brush in a similar way in the video, "Matisse: Voyage". When we talked with researchers in architectural design and in art education, they agreed with us that painters or designers often draw in the air before they draw on paper. This kind of hand movement is not even unique to painters. Sasaki & Watanabe (1983) also found that when writing Kanji characters, Japanese people often moved their fingers in the air. They interpreted this phenomenon to mean that

Table3: Result of Factor Analysis
(Varimax rotated factor pattern)

	Score: 7 ------ 1	Ⅰ	Ⅱ	Ⅲ	SMC
Good composition	Modulated--Non-modulated	**.70**	.27	.05	.57
	Focused--Unfocused	**.80**	.11	.13	.67
	Well-balanced--Ill-balanced	**.80**	.11	.16	.67
	Well-composed--Poorly-composed	**.79**	.07	.15	.66
Liveliness	Lively--Dull	.13	**.86**	-.03	.75
	Static--Dynamic	.06	**-.65**	.28	.50
	Energetic--Non-energetic	.20	**.85**	.13	.77
	Powerful--Power less	.42	**.62**	-.31	.65
simplicity	Clear cut--Mixed up	.56	-.15	**.63**	.72
	Simple--Complex	.29	-.16	**.62**	.49
	Relaxed--Crowded	.37	.20	**.70**	.67
	Light--Heavy	-.15	-.18	**.75**	.62
	Contributions	.410	.319	.271	

people use their body to remember Kanji Characters. Thus, it would be reasonable to hypothesize that moving in the air would have some important function not only when writing Kanji characters but also when drawing pictures.

We identified the timing of when he moved his brush in the air to investigate the role of the movement in his drawing. The cycles of drawing that we mentioned above were divided into three sections in order to identify the timing of his brush movement in the air. The first section, *beginning* section, was from his soaking the brush in the sumi ink plate until just before putting it on paper. The second section, *middle* section, was from his starting to draw until finishing to draw. The final section, *end* section, was from his lifting up the brush from the paper until just before soaking it in the ink plate. Then, we counted the number of brush movements in the air for each section.

The following coding scheme was used to identify brush movements. If the painter moved his brush more than once in a circle in the air, except for changing the posture of holding his brush or moving the brush from one place to another, we identify the movement as drawing in the air. A main coder coded the drawing processes of all sixteen pictures. After being taught this scheme and practicing coding independently, another coder coded one picture. The consistency between two coders was 90%. The percentage of intra-coder consistency of the main coder was 96%. Thus, the scheme was considered reliable.

Table 4 shows the mean number of drawing in the air and the percentage in each section in each condition. Although the frequency of drawing in the air in the LINES condition is higher than in the BLANK condition,

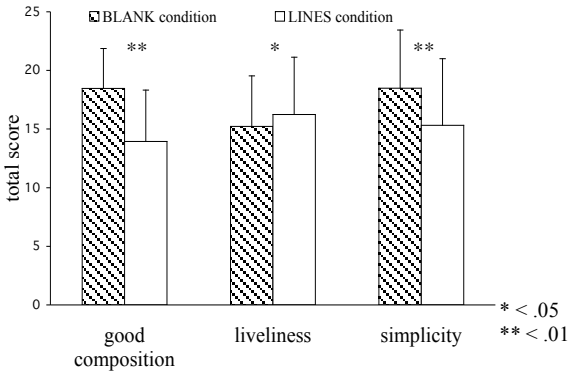


Figure4: Mean scores of three factors in each condition. Error bars represent 1*SD*.

the percentage of the drawing in each section is about the same between the two conditions. In the beginning section, the percentage of the drawing in the BLANK condition was 60% and that in the LINES condition was 56%. In the middle section, the percentage of drawing in the BLANK condition was 35% and that in the LINES was 36%. This indicated that Mr. K often draws in the air at the beginning and middle of drawing cycles. Thus, it would be reasonable for us to assume that drawing in the air has some important functions in drawing processes since they occur before the painter actually draws on paper.

Next, we focused on the relationship between pauses and drawing in the air. The percentage of pauses with drawing in the air in the BLANK condition was 59% and that in the LINES condition was 86%. This suggests that he often moves the brush in the air in order to think about drawing plans to incorporate lines into his picture. Furthermore, in the LINES condition, the percentage of pauses with drawing in the air, when he added on to others' lines, was 97% and when he drew without adding lines to others' lines was 59%. These results suggest that by moving the brush in the air, he generates a mental image to facilitate incorporating others' lines.

In order to further investigate the function of the drawing in the air, we interviewed him about his drawing process while showing a video record of his drawing a *Sansuiga* picture. While watching a part of the videotape in which he was drawing in the air, he said to us,

> "I might be checking how I feel when I touch the brush. Umm... Is this my habit? I always do this, don't I...I may move my hand in the air to rehearse my brush stroke...I always draw in the air before starting to draw on the paper. This seems to be my habit, doesn't it? Although I do not draw any actual objects on the paper, through drawing the form in the air, I can judge if the balance of the objects is OK. I have never realized my habit before you pointed it out. But, now I noticed it..."

This quote tells us that he probably moves his hands in order to plan how to use his brush and actually draw the image of objects in his mind. This is a quite reasonable candidate function of this hand movement. But, we need to be careful before making any conclusions on this issue based on the data from this field study. It would be, however, worth proposing some plausible hypotheses for future research. At this moment we propose the following

Table 4: Percentage of drawing in the air in three different sections

	Beginning section	Middle section	End section
BLANK condition	19.1 (60%)	11.0 (35%)	1.5 (5%)
LINES condition	29.0 (56%)	18.5 (36%)	4.0 (8%)

three functions as good candidates. First, by drawing in the air, the painter decides where to put the brush on the paper. We call this *positioning*. Second, the painter rehearses his brush movement so that he can draw smoothly. This is related to how to draw. We call this *rehearsal*. Third, by drawing an object in the air, the painter generates a mental image of what he plans to draw next. We call this *image generation*.

We could not confirm these hypotheses with this field study, because we could not control variables systematically. Further studies are needed to investigate the roles of drawing in the air.

General Discussion

This study focused on a traditional art, Chinese ink painting. Mr. K has an enormous amount of knowledge of the painting style and draws pictures using this knowledge. However, knowledge is not enough to create new pictures improvisationally and smoothly. When he drew the picture in temple X, he went backward to look at the entire picture. Also he occasionally covered this picture in progress with his hands to narrow down the space of focus. That is, he limited the drawing space to make planning or monitoring the picture easier. Thus, he could gradually form a mental image of a picture as the actual drawing on the paper progresses.

Knowledge and skills accumulated in years of expertise enable an artist to create artworks fairly quickly and smoothly. It seems that each brush of drawing evokes a local image of *Suibokuga* in Mr. K's memory. He creates his pictures combining those images based on certain rules that he learned from books or his experience. This process is highly effective when producing certain kinds of artwork.

On the other hand, artists often become bored while producing similar works too many times. When bored, artists want to try something new to stimulate their artistic motivation. In this *Suibokuga* painter's case, the method of asking the audience to draw random lines and incorporating them into his own picture is one such example. Creation of new patterns in artistic works seems to emerge through artists' intentional manipulation of constraints in a creation process. We found that even in a case of traditional art, artists sometimes conduct this kind of manipulation intentionally.

Artistic creation requires hands-on activities. Just having an image or a concept is not enough. In order to implement an image or a concept into an actual artwork, an artist needs to use his/her body. Sasaki et al. (1983) suggested that people would imagine the figure of Kanji characters by moving their hands. In the study of embodied representation, Barsalou (1999) has argued that sensorimotor processes, such as body movement, could affect the cognitive processes. Similarly, body movement in artistic creation, such as moving a brush in the air, also seems to play an important role in creative processes. In this way, artistic creation is a highly embodied process.

We could not make strong conclusions with this field study, since we could not control variables systematically. In addition, because this is a single case study, we also cannot generalize our findings to artists. However, we believe that our findings offer an essential first step towards future studies of the process of artistic creation from cognitive perspectives. We are currently conducting other studies regarding the creative process of Japanese contemporary artists in order to uncover potential similarities and differences between traditional and contemporary arts in an effort to generalize our hypotheses.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Cawelti, S., Rappaport, A., & Wood, B. (1992). Modeling artistic creativity: An empirical study. *Journal of Creative Behavior, 26* (2), 83-94.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Csikszentmihalyi, M. & Getzels, J. W. (1971). Discovery-oriented behavior and the originality of creative products: A study with artists. *Journal of Personality and Social Psychology, 19* (1), 47-52.

Eindhoven, J. E., & Vinacke, W. E. (1952). Creative processes in painting. *The Journal of General Psychology*, *47*, 139-164.

Getzels, J. W., & Csikszentmihalyi, M. (1976). *The creative vision: A longitudinal study of problem finding in art*. New York: Wiley.

Ichihara,Y. (1968). Kaigakansho no sinrigakutekikenkyu (1): Semantic differential shakudo ni kansuru kenkyu [Psychological study of the appreciation pictures (I): on semantic differential scales]. *The journal of social sciences and humanities. (Jimbungaku-ho), 62*, 79-90.

Mace, M. A., & Ward, T. (2002). Modeling the creative process: A grounded theory analysis of creativity in the domain of art making. *Creativity Research Journal, 14* (2), 179-192.

RM Arts, Le Centre Georges Pompidou, & La Sept. (Producer), & Baussy-Oulianoff, D. (Writer & Director). (1987). *Matisse: Voyage* [Video]. Chicago: Home Vision Entertainment.

Sasaki, M., & Watanabe, A. (1983). "Kusho" kodo no shutsugen to kino: Hyosho no undokankakuteki na seibun ni tuite [An experimental study of spontaneous writinglike behaviour ("Kusho") in Japanese]. *Japanese Journal of Educational Psychology, 31* (4), 273-282.

Suwa, M., & Tversky, B. (1997). What do architects and students perceive in their design sketches? *Design Studies*, *18* (4), 385-403.

Weisberg, R. W. (1993). *Creativity: Beyond the myth of genius*. New York: W.H. Freeman.

Honorifics in Japanese Sentence Interpretation: Clues to the Missing Actor

Yuki Yoshimura (yyuki@cmu.edu)
Department of Modern Languages, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213 USA

Brian MacWhinney (macw@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213 USA

Abstract

Cross-linguistic research in the framework of the Competition Model (MacWhinney & Bates, 1989) has shown that case-marking is the major cue to sentence interpretation in Japanese, whereas other cues such as animacy and word order are much weaker. Japanese is a pro-drop language. Many Japanese sentences are grammatical without subjects and objects. When subjects are absent, case-markers are also unavailable to use. However, Japanese honorific and humble verbs may provide important information to determine the agent when the case-marking cue is absent. This study examined the usage of honorific and humble verbs as the agreement cue in Japanese sentence interpretation by native speakers in comparison to their usage by second language learners of Japanese.

Introduction

Japanese uses specific affixes on the verb to mark social relations of power and solidarity. These markings are called honorifics. This use of formal grammatical markings is a unique feature of Japanese that has often been used as evidence for the operation of links between language, culture and thought (Whorf, 1967). Apart from these fascinating links to culture, honorifics can also serve the more mundane function of helping to distinguish the actor of the transitive verb. This paper is aimed at discovering the role of the verbal agreement cue in processing by native speakers. We are also interested in tracking the acquisition of this cue by second language learners of Japanese. Our study is couched within the framework of the Competition Model (MacWhinney & Bates, 1989) which emphasizes the relation between statistical regularities in the language and the strength of these cues for both first (L1) and second (L2) language speakers.

In order to interpret a transitive sentence, we have to identify the actor or agent. In the English sentence, *the doctor met the patient*, native speakers interpret *the doctor* as an agent who was engaged in meeting someone. This is because nouns placed before verbs are considered to be the actor in English. On the other hand, Japanese uses a completely different set of cues to determine the actor or agent. Although it has a basic SOV word order as in *the doctor the patient met*, Japanese also allows other word orders such as *the doctor met the patient* (SVO), *the patient the doctor met* (OSV), and *the patient met the doctor* (OVS). Each of these sentences yields the same interpretation with *the doctor* as the agent. Instead of relying on a word order cue, Japanese has case markers, such as *ga* (subject marker), *wa* (topic marker), *o* (object marker), and *ni* (dative marker) to mark case roles. The exact grammatical characterization of these participles has been the subject of dispute for years among Japanese linguists (Kuno, 1973), but there is little disagreement regarding the general importance of case role markings in the language. In general, a noun followed by the subject marker *ga* is likely to have an agentive role in any word order in Japanese. For example, a sentence like *kanja* (patient) *ni* (dative marker) *atta* (met) *isha* (doctor) *ga* (subject marker), in the order of OVS provides a meaning *the doctor met the patient*, although OVS order is not canonical in Japanese.

Cue Competition

Some theories tend to emphasize the universality of syntactic types across languages and the importance of a single "basic" order within languages (Chomsky, 1981). However, from the viewpoint of processing models, online sentence interpretation must rely at least initially on surface cues to role marking, and these cues vary markedly across languages. As we have already seen, English sentence interpretation relies heavily on word order (Bates & MacWhinney, 1989). This reliance would seem to support the central role of a fixed word order, as conceived in generative linguistic theories. However, other languages do not follow this pattern. A series of previous studies in the framework of the Competition Model have shown that case marking is the dominant cue in Japanese, Hungarian, and German, whereas subject-verb agreement cue is important in Italian, French and Spanish, and animacy distinction is the crucial determiner of interpretation in Chinese sentences (MacWhinney & Bates, 1989). English is unique in this sense in that it is the only well-studied language that depends so heavily on a word order cue.

Cue usage also varies developmentally within a single language. Children first focus on conspicuous cues that they can pick up easily (Slobin & Bever, 1982). Gradually they shift their cue usage to those that have high availability and

reliability in the language (McDonald, 1989). For example, Japanese children first focus on animacy distinction, because they already possess fairly clear ideas about which nouns are animate and which are inanimate. In contrast, learning of the case-marking system is a lot more complex than animacy, although it emerges eventually as the dominant cue in Japanese. Thus Japanese children rely on the animacy cue to interpret sentences first and then they later shift their cue usage to the case-marking cue which has high availability and reliability, i.e., it is a cue which is often present and which usually provides a correct interpretation.

Typically, languages make use of several cues for marking case roles. In most sentences, these cues agree with each other to guide a correct interpretation, though cues sometimes compete against each other. Sentences with inanimate subjects, such as *The study looked at Japanese children*, are quite common in written English. Here, word order provides a cue indicating that *the study* is the agent, whereas the animacy cue suggests that *Japanese children* should be the agent, because an animate noun is generally preferred as agent because of its dynamicity. Despite this cue competition, word order wins over animacy because word order has the highest availability and reliability in English (McDonald, 1987). On the other hand, when word order and animacy compete in Japanese sentences, animacy wins over word order because animacy is stronger in Japanese. Thus, in the parallel Japanese sentence *Japanese children* would be the agent (Sasaki & MacWhinney, in press). Similarly, when all of case marking, word order and animacy compete, case marking wins over animacy, and animacy wins over word order (case > animacy > word order), because case is the strongest cue in Japanese (Sasaki & MacWhinney, in press). Because Japanese word order is so flexible, it is the weakest cue in Japanese.

Unlike children's speech or child first language acquisition, in adult Japanese language use, in addition to theses basic cues, new cue emerges to compete with the other basic cues. This new cue is the honorific cue which is used along with verbs as morphological markings. Honorific and humble verbs are not used in children's speech because children are not yet expected to fully understand hierarchical society in Japanese culture that is reflected in the language use. However, the appropriate use of honorific and humble verbs becomes crucial in order to survive in adult Japanese society.

Cultural and Linguistic Interactions

Like Spanish and Italian, Japanese allows frequent omission of subject and object nominals. In English, pro-drop sentences, such as *∅ saw the black cat*, are considered to be ungrammatical. However, *neko* (cat) *o* (object marker) *mita* (saw) in Japanese is completely grammatical. Some Japanese linguists even claim that it is inappropriate to use the word "pro-drop" to describe Japanese constructions because subjects are not dropped but absent from the beginning (Kaneya, 2002). In Japanese, sentences without subjects and objects are completely grammatical. For example, the following short dialogue is very common in Japanese conversation.

A: "kuroi neko mita?"
[black cat saw?]
B: "un, mita."
[yeah, saw]

The subject is absent in utterance A, and both the subject and object are absent in utterance B. Yet, they are both grammatical (see more of these examples in Kaneya, 2002). Linguistically, word order cannot be an important cue because subjects and sometimes objects are both dropped in Japanese.

Japanese also marks cultural preferences regarding the status of the grammatical third person through the morphology of verbs and adjectives. Some verbs and adjectives carry information identifying the agent. However, the shape of this information is limited in specific ways, because Japanese culture inhibits stepping into others' psychological or physiological territory (Kamio, 1995). Verbs and adjectives that describe a third person's mental state have a special conjugation. Adjectives are usually used with *–garu* for third person, and verbs for third person are used in the *–teiru* form. For example, adjectives like *hoshii* (want, desirable), *ureshii* (happy), *itai* (painful) are all used with the adjectival third person marker *–garu* as in *hoshi-garu*, *ureshi-garu*, *ita-garu*. Verbs like *omou* (think), and *komaru* (have trouble) are used in the *–teiru* form in *omo-tteiru*, and *koma-tteiru*. It is ungrammatical to say *kanja* (patient) *ga* (subject marker) *ureshi* (happy-1st person, dictionary form), because a first person adjective cannot describe a third person subject. Rather, *ureshii* here should be *ureshi-garu* (3rd person). Therefore, even if subjects are dropped, sometimes verbs and adjectives will provide sufficient information to determine the agent.

Similarly, honorific and humble verbs are very useful cues that can be used for a variety of both transitive and intransitive verbs in Japanese. Honorific verbs cannot be used for the first person, but only for the second or the third person particularly for superiors. Humble verbs can be used only for the first person or the speaker's in-group members. For example, *o-hanashi-ninari-mashita* (honor + talk + honor + past tense) may indicate that the agent is someone superior to the speaker, and cannot be either the speaker or someone inferior to the speaker. Similarly, *o-hanashi-itashi-masu* (honor + talk + humble + non-past) indicates the action of the (humble) speaker or the speaker's (humble) in-group members. Even without overt mention of the subjects, these honorific and humble verbs provide evidence that is sufficient to identify the agent.

Considering the fact that subjects are frequently absent in Japanese, these verbal markings of honorific status should be one of the more reliable cues in Japanese. Although these cues are not always available, they should always be reliable when they are available particularly in adult speech.

Previous studies (Sasaki & MacWhinney, in press) have shown that the case-marking cue is the dominant cue in Japanese. However, when subjects and objects are absent, case markers are also naturally absent. When the dominant cue is unavailable, other cues must be used instead. As we have seen already, other cues such as animacy and word order have been examined in relation to the case-marking cue. However, the use of Japanese verb marking for honorific status has not yet been examined.

This study has two goals. The first is to measure the use of the honorific agreement cue by native Japanese speakers in comparison to case-marking cue and word order. Honorific and humble expressions are used only when there are social and psychological distances between the speaker and the listener, or between the speaker and the target person addressed. The availability of the honorific and humble verb cue would be high, particularly in adult speech under a hierarchical pressure, although overall availability of the honorific verb agreement cue in general speech may not be higher than that of the case-marking cue. Therefore, we can hypothesize that honorific verb agreement cue may not be stronger than the case-marking cue, yet it should be an important cue when case is absent.

The second goal of the study is to examine how second language learners acquire this verb agreement cue. Unlike native speakers, second language learners have not yet developed an entrenched usage of the case-marking cue. Moreover, when they first begin to pay attention to honorific marking, they may at first tend to overestimate and overgeneralize its importance, because the instruction is focusing specifically on this structure which learners tend to master in a short period of time.

Methods

Participants

Twenty native Japanese speakers, 16 advanced level Japanese learners as a foreign language, and 29 intermediate level Japanese learners participated in the study. Native Japanese speakers were recruited in Pittsburgh with a mean age of 31.6, and with a mean of length of residence in the United States less than 3 years. L2 learners were recruited from advanced and intermediate levels of Japanese courses at Carnegie Mellon University. Both intermediate and advanced learners have learned honorific and humble verb systems in class, though advanced learners have been exposed to them approximately for a year longer than intermediate learners. Some of advanced learners had an experience of studying abroad in Japan. No learners have an experience of studying abroad more than three months.

Stimuli

Three factors controlled in the study were word order, case-marking, and honorific cues with three levels for each. Word orders consisted of NNV, NVN, and VNN (N=Noun, V=Verb). Thus there were always two nouns and one verb used in every condition. The three levels for case-marking factor were nominative, dative and zero. When the first noun is marked with a nominative case and the second noun with a dative case, it is described as Nom_Dat condition, and Dat_Nom is used for the reversed case marking condition. Zero indicates the zero case marking condition. Honorific agreement cues were manipulated using simple transitive verbs such as *call*, *meet*, and *talk* with three levels of agreement: agreement-yes, agreement-no, and agreement-missing. Agreement-yes and agreement-no indicate whether the verb used in each condition agreed grammatically with the first noun. The agreement-missing condition contains plain verbs without modification of honorific or humble styles. Half of agreement-yes and agreement-no conditions used honorific verbs and the other half used humble verbs along with noun features differing in positional superiority. In order to control animacy effects, all nouns were animate. Each condition consisted of a noun combination differing in occupational superiority such as teacher-student, general-soldier, and president-employee.

For example, in the condition of order-NNV, case-zero, agreement-yes, sentences like, *sensei* (teacher) *gakusei* (student) *ohanashi ninari masu* (talk-honorific), was used. In this condition, we can see whether participants used either the case-marking or the verb agreement cue to interpret the sentence. Case is zero, so it provides no clues to determine the agent, whereas the verb agreement cue suggests that the honorable person should be the agent. In another condition where case is available, e.g., *sensei* (teacher) *ga* (subject marker) *gakusei* (student) *ni* (dative marker) *o hanashi itashimasu* (talk-humble), we can see clear competition between case and verb agreement cue. Case suggests *sensei* (teacher) marked with *ga* (subject marker) to be the agent whereas verb agreement with humble verb suggests *gakusei* (student) to be the agent although it is marked with a dative marker.

In addition to these three factors, filler sentences were inserted to guarantee that subjects treated the task in a natural fashion. The filler sentences excluded honorific agreement factor and superiority features in nouns. They controlled only word order and case-marking factors. The fillers also served the function of breaking up any tendency to lock into processing for specific verb types.

All three factors were fully crossed with three levels for three participant groups: native Japanese speakers, advanced learners, and intermediate learners. The full-factorial design of 3x3x3 was manipulated with the total of 54 sentences in addition to 18 filler sentences.

To create each sentence, the total of forty eight words was used: six base nouns (three superior and three subordinate nouns) with three case-marking forms (nominative, dative, and zero), and five base verbs with three agreement forms (honorific, humble, and plain) for experimental sentences, and five nouns (no superiority difference) with three case-marking forms for filler sentences. A male native Japanese speaker recorded the sentence components with a normal reading speed with no accents, and digitized with 16-bit

monaural .wav format at a 22-kHz sampling rate using CoolEdit 2000. Each sound file of nouns and verbs was combined into the appropriate sentence pattern using E-Prime 1.1 with complete random orders. The intonation patterns of combined words for all statements were indistinguishable, which prevented listeners from using any prosodic cues.

Procedure

In this task, all participants sat in front of a computer, and heard from a headphone a series of sentences that were composed of two nouns and a simple transitive verb. As the sentence began, the computer screen displayed pictures describing two nouns in each sentence. The pictures remained displayed until participants pressed a key indicating their choice of one of the pictures as the agent. Pictures were accompanied by words describing the pictures. For example, "teacher" (in Japanese) was shown on top of a picture of "teacher". This is to decrease non-native speakers' processing load. The subject identification task by itself for non-native speakers may put heavy processing load on their memory particularly when the cue competition is high, so they were instructed not to worry about memorizing words they did not know.

Participants were asked to choose or identify the picture that performed the action described in each sentence. They were instructed to choose either of the two nouns as agent, and to push a button corresponding to either of two pictures shown on the computer screen. If they thought the person on the right side did an action, they pushed the right button. If they thought the person on the left side did an action, they pushed the left button. They were instructed to press the button as quickly as possible after they heard a sentence. The response as a choice of the first noun was measured after 5 practice sentences.

All participants were tested individually in a small quiet room and were asked to complete the same task, although native Japanese speakers and non-native speakers were given different instructions before the task. Native Japanese speakers were told that sometimes sentences were culturally inappropriate or grammatically incorrect, and they were asked to respond quickly using their intuitions in the case they found some sentences unnatural. As it was mentioned earlier, these sentences were set up to test the usage of cue competition. For non-native Japanese speakers, additional instructions were given. In order to refresh learners' memory and make them comfortable about the use of honorific and humble verbs, each non-native participant went through a brief review on verb conjugations. They were also briefly informed about social and hierarchical differences between roles such as general vs. soldier, and president vs. employee, which were used in the experiment.

Results

ANOVAs were performed using percentage of choice of the first noun as agent as the dependent variable. The main

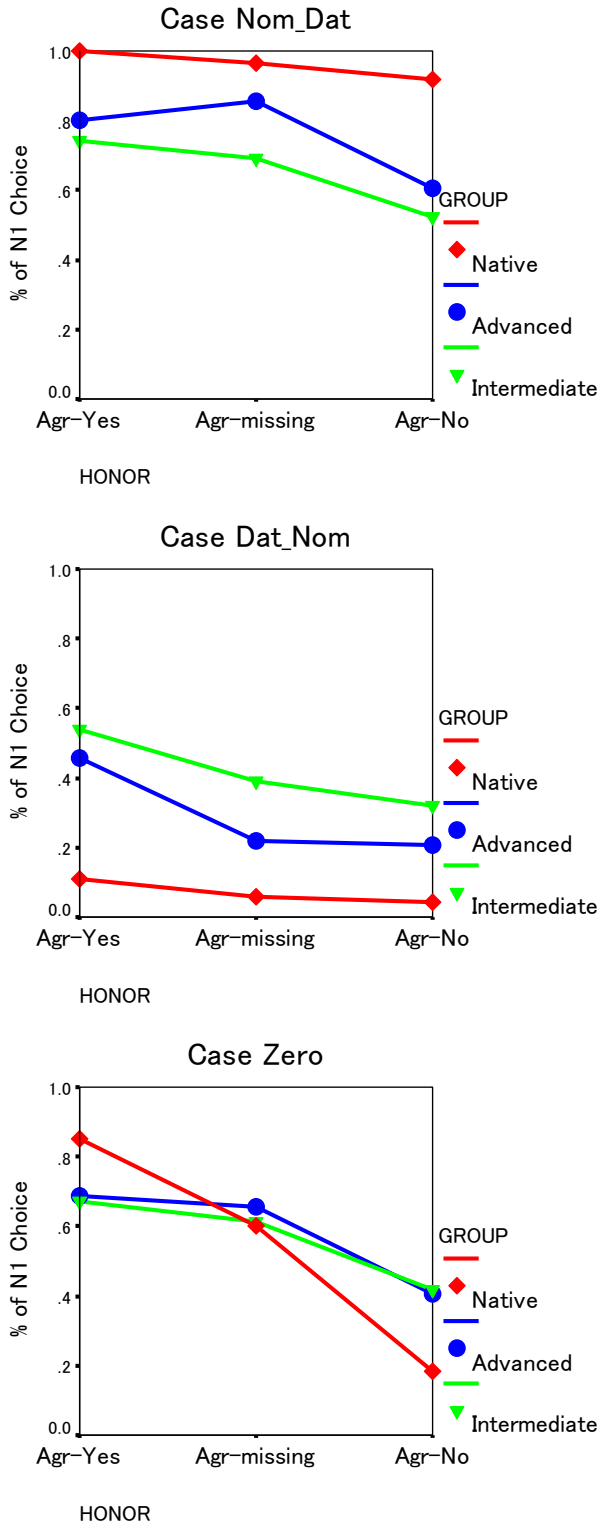


Figure 1: The percentage choice of the first noun as agent in the condition of (1) Nom_Dat case (the first noun is marked with the nominative case, and the second noun is marked with dative case) crossed with the agreement cue, (2) Dat_Nom case and (3) Zero case (neither nouns were marked with case).

effects of case and honorifics were significant (Case: $F(2, 61) = 164$, $p < 0.001$; Honor: $F(2, 61) = 17.7$, $p < 0.001$), but word order was not. The non-significance of the word order effect shows that none of the groups, including L2 learners, relied on the word order cue. The interaction between case and honorific agreement was also significant ($F(4, 59) = 20$, $p < 0.001$). The use of nominative case for the first noun and dative case for the second noun is shown in Figure 1. In the Nom_Dat condition, native speakers consistently used the case-marking cue, despite the presence of an agreement cue. L2 learners also made use of the case-marking cue, though they relied much less on case than did native speakers. Native speakers' first noun choice for Nom_Dat was over 95% in all three agreement conditions, whereas the first noun choice of advanced and intermediate learners dramatically decreased in the condition where verbs do not agree with the first noun feature (advanced: 60%; intermediate: 52%).

Similarly, in the Dat_Nom case condition, native speakers consistently showed the strong use of the case-marking cue to select the second noun as agent. The second noun is marked with a nominative case in this condition, so the lower percentage of the first noun choice indicates the heavier reliance on using the case-marking cue. The first noun choice by native speakers was less than 11% in all agreement conditions, whereas L2 learners' first noun choice was higher. When the verb agrees with the honorific status of the first noun, learners' first noun choice increased noticeably (advanced: 46%; intermediate: 54%). This shows that learners placed heavy reliance on the honorific agreement cue even when it contradicted the case-marking cue.

In the zero case-marking condition, the first noun choice by native speakers showed a clear decline from agreement-yes to agreement-no conditions (agreement-yes: 85%; agreement-missing: 60%; agreement-no: 18%). Because the case-marking cue was unavailable in this condition, native speakers relied on the honorific verb agreement cue to determine the agent. On the other hand, learners' usage of verb agreement cue was not as robust as native speakers'.

Discussion

The patterns of Japanese native speakers' performance in the experiments showed that the case-marking cue was still the dominant cue in Japanese. Despite the presence of the honorific verb agreement cue, native speakers consistently chose nouns marked with the nominative case as agent. However, when case was absent, the honorific verb agreement cue became an important and reliable cue to determine the agent. Even though Japanese has a canonical SOV word order, we did not find any effects of the word order cue in the absence of case. This confirms the results from the previous studies. Importantly, we demonstrated within a single experiment both inattention to word order and attention to honorific agreement. Thus, it appears that honorific agreement is the second major cue in Japanese sentence processing, after case-marking.

The second important finding of the study is that cue availability determined cue strength. Native speakers' first noun choices were not entirely controlled by the agreement cue. There was 85% first noun choice in the agreement-yes condition and 18% in the agreement-no condition. This suggests that even native speakers are sometimes unsure about the correct use of honorific and humble verbs. As we have already discussed earlier, honorific and humble verbs are not frequently present in younger people's daily linguistic input until they start working in businesses. Some of the participants of native speakers in the study were graduate students who have no experience working in businesses. Therefore, their cue usage of the honorific verb agreement cue might not have been as strong as we could find in speakers from the business environment. Further study of the use of this cue by speakers from the business community may help us understand the extent to which increased availability of the cue could lead to an increase in its relative strength when it is placed in conflict with case-marking.

The third finding of the study relates to the cue usage patterns by L2 learners. As we predicted, their usage of both the case-marking cue and the honorific agreement cue was more variable than that of native speakers. Interestingly, they overused the honorific agreement cue even when the case-marking cue was available to use. On the other hand, they did not use the honorific cue as much as they could, when case was absent. This suggests that learners' use of the case-marking cue has not yet stabilized at native speaker levels after about two years of learning Japanese. Moreover, the overuse of the honorific cue indicates that learners tend to focus on a single cue when this is at the focus of an instructional module or experiment. This tendency to focus on individual cues may prevent them from fixing the relative strength of each cue among all available cues in the target language. To counteract this tendency, instructors may need to present learners with input that illustrates competition between the relevant cues. In particular, learners need more experience with sets of sentences in which case marking is either present or absent and in which honorific agreement is either present or absent.

References

Chomsky, N. (1981). *Lectures on government and binding*. Cinnaminson, NJ: Foris.

Kamio, A. (1995). Territory of information in English and Japanese and psychological utterances. *Journal of Pragmatics, 24*, 235-264.

Kaneya, T. (2002). *Nihongo ni shugo wa iranai*. Tokyo: Kodansha.

Kuno, S. (1973). *The structure of the Japanese language*. Cambridge, MA: MIT Press.

MacWhinney, B., & Bates, E. (Eds.). (1989). *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.

McDonald, J. L. (1987). Sentence interpretation in bilingual speakers of English and Dutch. *Applied Psycholinguistics, 8*, 379-414.

McDonald, J. L. (1989). The acquisition of cue-category mappings. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of language processing* (pp. 375-396). New York: Cambridge University Press.

Sasaki, Y., & MacWhinney, B. (in press). Language acquisition research based on the Competition Model. In M. Nakayama & R. Mazuka & Y. Shirai (Eds.), *Handbook of Japanese Psycholinguistics*. Cambridge: Cambridge University Press.

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A cross-linguistic study of word order and inflections. *Cognition, 12*, 229-265.

Whorf, B. (1967). *Language, thought, and reality*. Cambridge, MA: MIT Press.

Angular Disinhibition Effect in a Modified Poggendorff Illusion

Yingwei Yu and Yoonsuck Choe
Department of Computer Science
Texas A&M University
3112 TAMU
College Station, Texas 77843-3112, USA
{yingwei,choe}@tamu.edu

Abstract

Visual illusion can be strengthened or weakened with the addition of extra visual elements. For example, in Poggendorff illusion, with an additional bar added, the illusory skew in the perceived angle can be enlarged or reduced. In this paper, we show that a nontrivial interaction between lateral inhibitory processes in the early visual system (i.e., *disinhibition*) can explain such enhancement or degradation of the illusory percept. The computational model we derived successfully predicted the perceived angle in a modified Poggendorff illusion task with an extra thick bar. The concept of disinhibition employed in the model is general enough that we expect it can be further extended to account for other classes of geometric illusions.

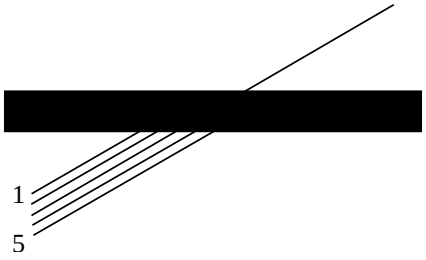


Figure 1: **The Poggendorff Illusion.** The original Poggendorff illusion is shown. The five lines below the horizontal bar are labeled 1 to 5 from top to bottom. Line 3 is physically collinear with the line on top. In this example, line 4 is perceived to be collinear.

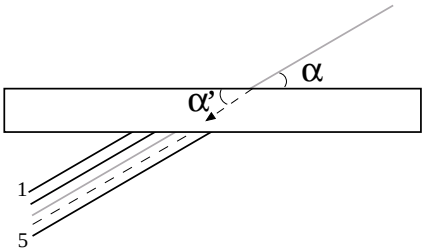


Figure 2: **The Angle Displacement in the Poggendorff Illusion.** The actual angle α ($= 30°$) and the perceived angle α' ($> 30°$) are shown. The gray line shows the straight line penetrating the bar. The dashed line below shows the perceived direction in which the line on top seemingly extends to.

Introduction

Visual illusions are important phenomena because of their potential to shed light on the underlying functional organization of the visual system. For simple illusions, a simplistic explanation can be sufficient, but when multiple effects exist in an illusion, the final percept can be quite complex. For example, when we perceive an angle, our perception of the angle is usually greater than the actual angle (*expansion* effect), but when there are multiple lines and thus multiple angles, the expansion effect can be either enhanced or reduced.

Such an interference effect can be demonstrated in a modified Poggendorff illusion. In the original Poggendorff illusion (see, e.g., Tolansky 1964; Morgan 1999), the top and the bottom portions of the penetrating thin line is perceived as misaligned (figure 1). Figure 2 shows how such a perception of misalignment can occur. The line on top forms an angle α with the horizontal bar, but the perceived angle α' is greater than α (i.e., exaggerated). As a result, the line on top is perceived to be collinear with line 4 on the bottom, instead of line 3 which is physically collinear. However, when an additional bar is added, the perceived illusory angular expansion effect is altered: the effect is either reduced (figure 3) or enhanced (figure 4) depending on the orientation of the newly added bar. Understanding the functional organization and the low-level neurophysiology underlying such a nontrivial interaction is the main aim of this paper.

Neurophysiologically speaking, in the original case where two orientations interact, lateral inhibition between orientation cells in the visual cortex can explain the enlargement in perceived angle. However, as we have seen in figures 3 and 4, when an additional orientation response is triggered, lateral inhibition alone cannot explain the complex effect. Our observation is that this complex response is due to disinhibition, i.e., inhibition of another inhibitory factor resulting in effective excitation (Hartline et al. 1956; Hartline and Ratliff 1957, 1958; Stevens 1964; Brodie et al. 1978). Unlike simple lateral inhibition between two cells, we explicitly accounted for disinhibition in our computational model to describe the complex interactions between multiple orientation cells. The resulting model based on the neurophysiology of the early visual system was able to accurately predict the perceptual performance for the modified Poggendorff illusion.

The rest of the paper is organized as follows. First, a neurophysiological motivation for our computational model is presented, followed by a detailed mathematical description of the model. Next, the results from the computational experiments with the model is presented and compared to psychophysical data, followed by discussion and conclusion.

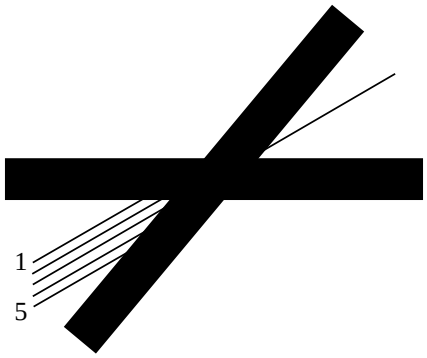


Figure 3: **The Poggendorff Illusion with an Additional Thick Bar of 50 °.** The Poggendorff figure with an additional bar at 50° is shown. In this case, line 2 is perceived to be collinear (i.e., $\alpha' < 30°$).

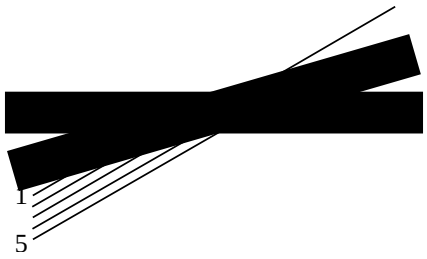


Figure 4: **The Poggendorff Illusion with An Additional Thick Bar of 20 °.** The Poggendorff figure with an additional bar at 20° is shown. For this case, unlike in figure 3, line 4 (or to some, line 5) is perceived to be collinear ($\alpha' > 30°$). (The α' in this case is slightly greater than in original Poggendorff figure.)

Computational Model of Disinhibition in the Visual Cortex

Let us first consider how orientation columns in the visual cortex interact in response to several intersecting lines. For each line at the intersection, there are corresponding orientation columns that respond maximally, with a Gaussian response distribution. As multiple simple cells are activated by the different lines at the intersection, the response levels will interact with each other through lateral connections. Thus, there are two issues we want to more precisely address: (1) what exactly is the activation profile (or the response distribution) of the orientation-tuned cells, and (2) how these cells interact with each other through the lateral connections.

The Activation Profile of Orientation Columns

Each simple cell in the primary visual cortex responds maximally to visual stimuli with a particular orientation. The response of these cells to different orientations can be modeled as a Gaussian function:

$$y = y_0 + \frac{A}{\sigma\sqrt{\pi/2}}\exp(-2\frac{(x-x_c)^2}{\sigma^2}), \qquad (1)$$

where y_0 is an offset; x_c is the center (or mean); σ is the standard deviation; and A is a scaling constant (Martinez et al. 2002).

It also comes to our attention that the cell tuned for a certain orientation, say α, should respond to the opposite orientation, which is $\alpha + 180°$. However, experiments have shown that the peak at the position $\alpha+180°$ is somewhat smaller than the peak at α (Alonso and Martinez 1998). To accurately model this, we need two Gaussian curves to fit the responses of a cell to a full range of orientations from $-180°$ to $180°$.

The fitting curve can be written as follows:

$$\begin{aligned} y &= y_0 \\ &+ \frac{A}{\sigma\sqrt{\pi/2}}\exp(-2\frac{(x-x_c)^2}{\sigma^2}) \\ &+ \frac{AK}{\sigma\sqrt{\pi/2}}\exp(-2\frac{(x-x_c+\pi)^2}{\sigma^2}), \end{aligned} \qquad (2)$$

where K is the rate of activation for the opposed direction ($K < 1$). Such an asymmetric response enables the the simple cells to be sensitive to the direction (as well as orientation).

Using the equation, we can now visualize the response profile of simple cells tuned to orientations ranging from 0 to 360°. Figure 5 shows the responses of orientation columns tuned to -90°to 270° (x-axis) to inputs of two different orientations, 0 and 30°. Figure 6 shows the responses of the same set of orientation columns to inputs of two orientations of 0 and 150°. From these two figures, we can observe that for each specific orientation input, the excitation is tuned at that value with a peak in the Gaussian curve, and at the same time, the opposite orientation tuned cell shows a lower peak response. The asymmetry in responses occur in both an acute angle (figure 5) and an obtuse angle (figure 6). Note that even though the difference in orientation between 0° vs. 30° and 0° vs. 150° is 30° in both cases, the response profile greatly differs in the 0° vs. 150°case.

This is an improvement over conventional excitation profile models such as Gabor filters (Daugman 1980), which make no distinction between these two angles in the two figures. Using the more accurate response profile, we will next investigate how these response profiles can interact.

Column Level Inhibition and Disinhibition

Our observation that the angular enlargement sometimes seems to be weakened when there are more than two bars or lines in the Poggendorff illusion (figure 3) led us to hypothesize about the potential role of a recurrent inhibition effect, i.e., disinhibition. Basically disinhibition is the inhibition on other inhibitory factors, resulting in a net excitatory effect at the target. Experiments on the Limulus (horseshoe crab) optical cells showed that the final response of each receptor resulting from a light stimulus can be enhanced or reduced due to the interactions through inhibition from its neighbors. Note that disinhibition has also been found in vertebrate retinas such as in tiger salamanders (Roska et al. 1998) and in mice (Frech et al. 2001). In the following, the Limulus neurophysiology giving rise to disinhibition is summarized, followed by the description of our computational model based on the Limulus model.

Hartline-Ratliff's model of disinhibition Experiments on Limulus optical cells have shown that lateral inhibition effect

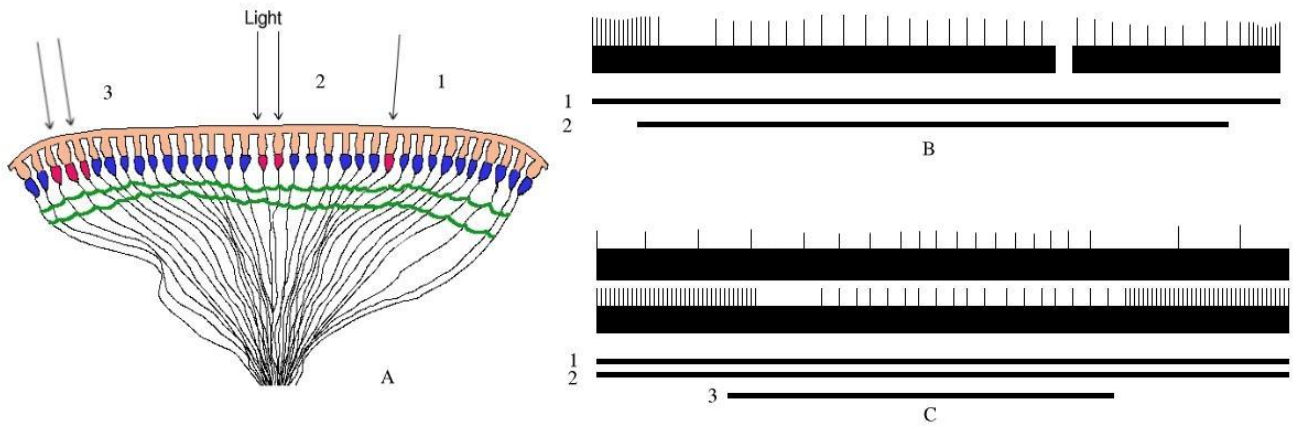


Figure 7: **Lateral inhibition in Limulus optical cells.** The figure shows the disinhibition effect in Limulus optical cells. (a) The retina of Limulus. Point light is presented to three locations (1, 2 and 3). (b) The result of lighting position 1 and 2. The top trace shows the spike train of the neuron at 1, and the two bars below show the duration of stimulation to cell 1 and 2. When position 2 is excited, the neuron response of position 1 gets inhibited. (c) Both 1 and 2 are illuminated, and after a short time, position 3 is lighted. The top two traces show the spike trains of cell 1 and cell 2. The three bars below are input duration to the three cells. As demonstrated in the figure, when position 3 is lighted, neurons at position 2 get inhibited by 3, so its ability to inhibit others get reduced. As a result, the firing rate of neuron at position 1 gets increased during the time neuron at position 3 is excited. This effect is called disinhibition. Redrawn from (Hartline and Ratliff 1957).

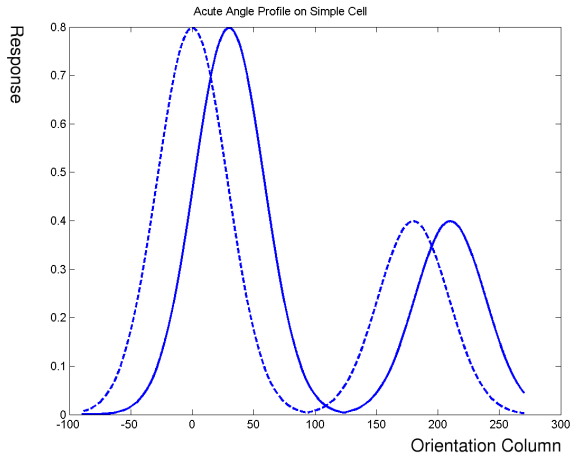


Figure 5: **The Activation on Simple Cell by an Acute Angle (**30°**).** The dotted curve is the responses of the orientation columns (x-axis) to a horizontal line of 0°, and the solid curve is the responses to 30° line.

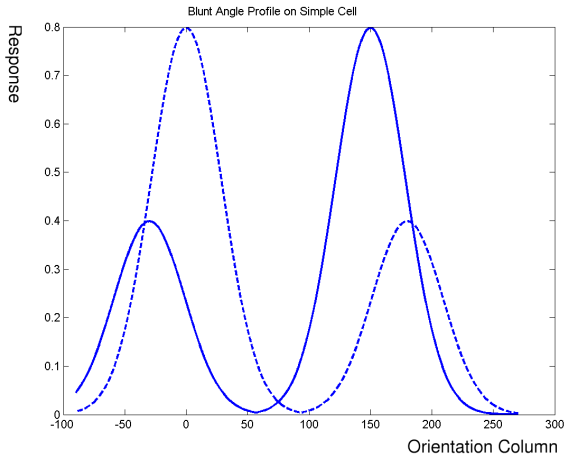


Figure 6: **The Activation on Simple Cell by Blunt Angle.** The dotted curve is the responses of the orientation columns (x-axis) to a horizontal line of 0°, while the solid curve is the responses to a 150° line.

is recurrent (figure 7; see Hartline and Ratliff 1957, 1958). The final response of a specific neuron can be considered as the overall effect of the response from itself and from all other neurons. Conventional convolution operation using lateral inhibition alone does not account for the effect of disinhibition which plays an important role in the final response. The final response of each receptor resulting from a light stimulus can be enhanced or reduced due to the interactions through inhibition from its neighbors, which may be important. (Such disinhibition effects have been found to play an important role in brightness-contrast illusions Yu et al. (2004).)

Hartline and his colleagues also did significant mathematical modeling of the Limulus optical cell response. The Hartline-Ratliff equation describing disinhibition in the Limulus can be written as follows (Hartline and Ratliff 1957, 1958; Stevens 1964):

$$r_m = \epsilon_m - K_s r_m - \sum w_{m \leftarrow n}(r_n - t_{m \leftarrow n}), \qquad (3)$$

where r_m is the response, K_s is the self-inhibition constant, ϵ_m is the excitation of the m-th ommatidium, $w_{m \leftarrow n}$ is the inhibitory weight from other ommatidia, and $t_{m \leftarrow n}$ the threshold.

Brodie et al. extended this equation to derive a spatiotemporal filter, where the input was assumed to be a sinusoidal grating (Brodie et al. 1978). This model is perfect in predicting Limulus retina experiments as only a single spatial frequency channel filter, which means that only a fixed spatial frequency input is allowed (Brodie et al. 1978). Because

of this reason, their model cannot be applied to a complex image, as various spatial frequencies could coexist in the input. In the following section, we will build upon the Hartline-Ratliff equation and derive a filter that can be used in modeling orientation columns.

A simplified model of disinhibition Based on the Hartline-Ratliff equation above, we derived a model for two-dimensional disinhibition as follows (Yu et al. 2004):

$$\mathbf{r} = W^{-1} \times \mathbf{x}. \tag{4}$$

where $\mathbf{r}$ is the output vector, $\mathbf{x}$ is the input vector and W is the weight matrix:

$$W_{ij} = \begin{cases} -w(|i,j|) & \text{when } i \neq j \\ 1 & \text{when } i = j \end{cases}, \tag{5}$$

where $w(i,j)$ is the kernel function (usually a difference-of-Gaussian) defining the inhibition rate from the j-th neuron to the i-th neuron. Based on this simplified model of disinhibition, we can now more easily derive the disinhibition effect at the orientation column level.

Applying disinhibition to orientation cells Cells occupying the same single orientation column in the cat visual cortex are known to inhibit each other (Blakemore and Tobin 1972). From this, we can postulate that a group of cells tuned to the same orientation representing different lines (e.g., intersecting lines) may compete with each other through inhibition.

Now let us consider the mathematical description for the inhibition at the column level. Suppose a group of orientation cells tuned to orientation α receives n lines as their inputs. The initial excitation $E_{\alpha i}$ for a cell inside this group α can be calculated as follows:

$$\begin{aligned} E_{\alpha i} &= y_0 \\ &+ \frac{A}{\sigma\sqrt{\pi/2}} \exp(-2\frac{(\alpha - x_c(i))^2}{\sigma^2}) \\ &+ \frac{AK}{\sigma\sqrt{\pi/2}} \exp(-2\frac{(\alpha - x_c(i) + \pi)^2}{\sigma^2}), \end{aligned} \tag{6}$$

where y_0 is an offset, A is a scaling constant for the Gaussians, σ is the standard deviation, K is the rate of activation of the opposite direction, and $x_c(i)$ is the orientation of the i-th input line. In this way, we can calculate the excitation E of the cell to the i-th line on a certain group of cells tuned to α. All those parameters in this equation are fairly standard parameters, which does not require a precise tuning.

Using the Hartline-Ratliff equation (Hartline and Ratliff 1957) for recurrent lateral inhibition and the simplified model of disinhibition (Yu et al. 2004), the final response R of cell i in orientation column α can be obtained as follows:

$$R_{\alpha i} = E_{\alpha i} - W \times R_{\alpha i}, \tag{7}$$

where W is a constant matrix of inhibition rate (or weight, controlled by a free parameter η: $w_{ij} = \eta$ if $i \neq j$, and 0 otherwise). From this, we can finally derive the response equation which accounts for the disinhibition effect:

$$R_{\alpha i} = (I - W)^{-1} \times E_{\alpha i}, \tag{8}$$

where I is the identity matrix.

By applying the orientation α to all the columns, the projection of each line to the columns should shift a little bit depending on the strength of the activation of each line. Thus, the final perceived line orientation γ can be obtained by finding the maximum response after the inhibition process:

$$\gamma_i = \text{argmax}_{\alpha \in C} R_{\alpha i} \tag{9}$$

where γ_i is the perceived orientation for the i-th line, R is the responses of i-th neuron tuned to orientation α and C is the set of all the orientation columns in layer 4 of the visual cortex.

Experiments and Results

Prediction of Angle Expansion without Additional Context

To test the model in the simplest stimulus configuration, we used stimuli consisting of one thick bar and one thin line. The thick bar was fixed at 0°, and the thin line was rotated to various orientations while the perceived angle was measured in the model. The enlargement effect of the angle varied depending on the orientation of the thin line. As shown in figure 8, we can observe that there are three major characteristics of this varying effect. First, for the acute angles, there is an increment in the angle of the perceived compared to actual, but for the obtuse angles, the perceived angle is less than the actual angle. Second, the peak is around 20° for the largest positive displacement, and around 160° for the largest negative displacement. Third, there is an obvious asymmetry in the displacements between the acute angles and the obtuse angles. Note that the peak at 20° is greater in magnitude than the dip at 160°. As compared in figure 8, these results are consistent with results obtained in psychophysical experiment by Blakemore et al. (1970).

Prediction for the Modified Poggendorff Illusion

Disinhibition effect is the key observation leading to our extension to the angular expansion model based on lateral inhibition alone. Because of disinhibition, when more than two lines or bars intersect, the perceived angle of the thin line will deviate from the case where only two lines or bars are present. Figure 9 shows the prediction of our model (solid line) when a second thick bar of varying orientations was added to the original Poggendorff illusion (see figure 3 and 4 for an example). If disinhibition effect did not exist, the solid line would have come out flat, however, there is an interesting peak and a valley in the predicted response. The effect demonstrated in figure 3 is accurately predicted by the peak near 20°, and the effect in figure 4 by the valley near 50°. So, at least for these two cases, we can say that our disinhibition-based explanation is accurate. However, does the explanation hold for an arbitrary orientation? To test this, we conducted a psychophysical experiment to measure human perceptual performance and compare the results to the model prediction (the results are shown as data points in figure 9).

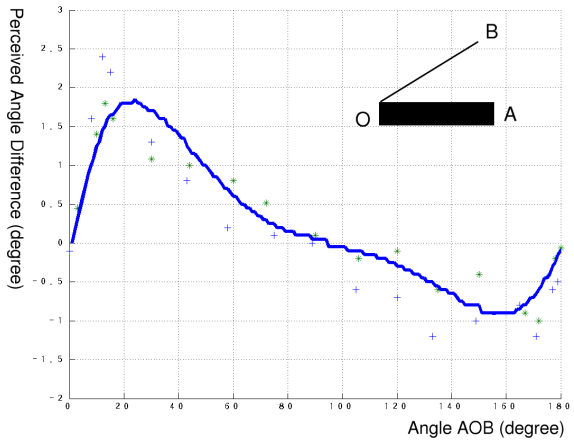


Figure 8: **The Variations of Perceived Angle Between Two Intersecting Lines.** The x-axis corresponds to the angle $\angle AOB$ (inset), from 0 to 180°. The y-axis is the difference between the perceived angle and the actual angle. The solid line is the result predicted by our model, and the data points * and + are data from human subjects in Blakemore et al. (1970). The curve was generated in two iterations, with the following parameters: $\eta = 0.009$ and $\sigma = 1.0$ for the first pass; $\eta = 0.005$ and $\sigma = 0.5$ for the second. The other parameters remained the same for both iterations: $y_0 = 0.0$ and $K = 0.5$.

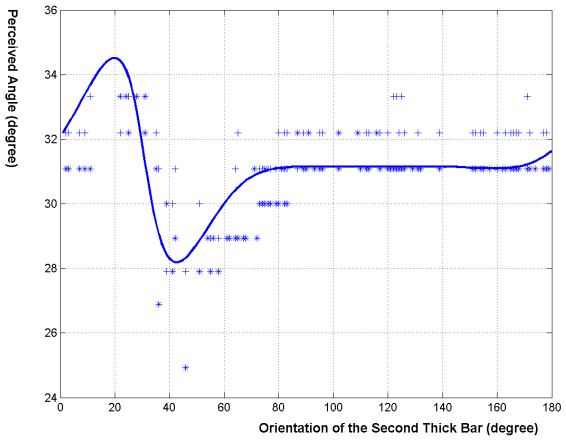


Figure 9: **Perceived Angle in a Modified Poggendorff Illusion.** The results from the computational model (solid line) and human experiments (data points marked * and +) on a modified Poggendorff illusion (figure 3) are plotted. The second thick bar was rotated while the perceived angle was measured. The x-axis indicates the angle of the second bar. The y-axis shows the perceived angle of the thin 30° line. The model prediction and the human data are in close agreement. The parameters used in this experiment were as follows: free parameter: $\eta = 0.02$; standard parameters: $y_0 = 0$, $\sigma = 0.5$, and $K = 0.5$.

Experimental methods Two subjects with normal vision participated in the experiment (the authors YC and YY). An LCD panel with a 1024×768 resolution, which is supposed to be high enough to avoid line aliasing aritfacts, was used to display the stimuli. The computer program displayed two thick bars and one thin line on the screen, similar to the stimulus in figure 3. The first thick bar was fixed in the center of the screen at 0°, and the width was 100 pixels. The thin line, 5 pixels in width, intersected the horizontal bar at a fixed angle of 30°. The second thick bar, 100 pixels in width, intersected at the same point as the the other two, where as the angle was varied from trial to trial. The program also displayed up to 10 thin lines (all 30°) below the horizontal bar, from which the subjects were asked to choose the one that is the most collinear to the thin line above the bar. The subject was allowed to click on the line of choice, and the perceived angle was recorded for each click, and a new stimulus was generated. A total of 101 trials were recorded for each subject.

Results Figure 9 shows the result of the psychophysical experiment (data points * and + for YC and YY, respectively), along with the prediction of the model (solid line). The peak (near 20°) and valley (near 50°) are apparent in the experimental data, and the overall shape of the curve closely agrees with the model prediction. The results show that our model of angular interaction based on disinhibition can accurately explain the modified Poggendorff illusion, and that low-level neurophysiology can provide us with insights into understanding the mechanisms underlying various visual illusions. Note that for this experiment, our disinhibition model is more comprehensive than the calculation method of simply summing up two Poggendorff effects by two separate bars. First it is because disinhibition is the summing up between all the pairs of bars at neuronal level, and second, simply summing up the effects of two bars omits the interactions between the lateral inhibition effects.

Discussion

We have presented a model based on angular inhibition by considering the disinhibition effect. The soundness of the theoretical extension lies in the fact that it is grounded in physiological and psychological facts. First, at the cellular level, lateral inhibition and disinhibition effects are found in the visual column of cat (Hubel and Wiesel 1962; Blakemore and Tobin 1972) and it is known that the opposite directions of the same orientation evoke an asymmetric response (Alonso and Martinez 1998). Our prediction of the angle variations for acute and obtuse angles shows asymmetric properties and matches these experiments. Second, our model can correctly predict the disinhibition caused by more than two lines intersected and the results match with our own experimental observation using the same kind of stimuli.

Besides the Poggendorff illusion, our model has the potential for explaining other geometric illusions, such as the café-wall illusion. Fermüller and Malm (2003) showed a variation of the café-wall illusion where adding some dots in strategic places significantly reduced the perceived distortion. Such a correctional effect can be explained by our model. Because the newly introduced dots give rise to a new orientation component (as the second thick bar did in our modified Poggendorff illusion), the disinhibitory effect caused by that new orientation can reduce the distortion formed by the existing orientation components.

Even though the disinhibition model presented in this paper is largely motivated by low-level neurophysiology, disinhibition can potentially serve a more general function. For ex-

ample, disinhibition can also be applied to higher brain functions such as categorization and memory (see Vogel (2001) for a model of associative memory based on disinhibition).

Conclusion

In this paper, we presented a neurophysiologically based model of disinhibition to account for a modified version of the Poggendorff illusion. The model was able to accurately predict a subtle orientation interaction effect, closely matching the psychophysical data we collected. We expect the model to be general enough to account for other kinds of geometrical illusions as well.

Acknowledgment

This research was supported in part by Texas A&M University, by the Texas Higher Education Coordinating Board grant ATP#000512-0217-2001, and by the National Institute of Mental Health Human Brain Project grant #1R01-MH66991.

References

Alonso, J., and Martinez, L. M. (1998). Functional connectivity between simple cells and complex cells in cat striate cortex. *Nature neuroscience*.

Blakemore, C., Carpenter, R. H., and Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature*.

Blakemore, C., and Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Exp. Brain Res.*

Brodie, S., Knight, B. W., and Ratliff, F. (1978). The spatiotemporal transfer function of the limulus lateral eye. *Journal of General Physiology*, 72:161–202.

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20:847–856.

Fermüller, C., and Malm, H. (2003). Uncertainty in visual processes predicts geometrical optical illusions. *Vision Research*.

Frech, M. J., Perez-Leon, J., Wassle, H., and Backus, K. H. (2001). Characterization of the spontaneous synaptic activity of amacrine cells in the mouse retina. *Journal of Neurophysiology*, 86:1632–1643.

Hartline, H. K., and Ratliff, F. (1957). Inhibitory interaction of receptor units in the eye of Limulus. *Journal of General Physiology*, 40:357–376.

Hartline, H. K., and Ratliff, F. (1958). Spatial summation of inhibitory influcences in the eye of limulus, and the mutual interaction of receptor units. *Journal of General Physiology*, 41:1049–1066.

Hartline, H. K., Wager, H., and Ratliff, F. (1956). Inhibition in the eye of limulus. *Journal of General Physiology*, 39:651–673.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, 160:106–154.

Martinez, L. M., Alonso, J., Reid, R. C., and Hirsch, J. A. (2002). Laminar processing of stimulus orientation in cat visual cortex. *Journal of Physiology*.

Morgan, M. (1999). The poggendorff illusion: a bias in the estimation of the orientation of virtual lines by second-stage filters. *Vision Research*, 2361–2380.

Roska, B., Nemeth, E., and Werblin, F. (1998). Response to change is facilitated by a three-neuron disinhibitory pathway in the tiger salamander retina. *Journal of Neuroscience*, 18:3451–3459.

Stevens, C. F. (1964). *A Quantitative Theory of Neural Interactions: Theoretical and Experimental Investiga-tions*. PhD thesis, The Rockefeller Institute.

Tolansky, S. (1964). *Optical Illusions*. London: Pergamon.

Vogel, D. (2001). A biologically plausible model of associative memory which uses disinhibition rather than long-term potentiation. *Brain and Cognition*, 45:212–228.

Yu, Y., Yamauchi, T., and Choe, Y. (2004). Explaining low-level brightness-contrast illusions using disinhibition. In *Biologically Inspired Approaches to Advanced Information Technology (Bio-ADIT 2004; LNCS)*. New York: Springer. In press.

When Holistic Processing is Not Enough: Local Features Save the Day

Lingyun Zhang and **Garrison W. Cottrell**
lingyun,gary@cs.ucsd.edu
UCSD Computer Science and Engineering
9500 Gilman Dr., La Jolla, CA 92093-0114 USA

Abstract

Is configural information or featural information more important for facial identity recognition? How are the skills for processing these types of information developed? To investigate these issues, Mondloch et al. designed three sets of face images based on a single face, "Jane", to measure featural, configural, and contour processing. These stimuli were tested on human subjects of different ages in a same/different task. We test our model [Dailey et al., 2002] of face processing on these stimuli. We find that our model is overly holistic: It finds the configural differences the easiest to detect, while adult human subjects find featural changes the easiest to detect. We then introduce a representation of the important parts of the face (eyes and mouth) to our holistic model. We find that only a relatively small amount of holistic representation, compared to parts representations, is necessary to account for the data.

Introduction

We have developed a model of face processing that accounts for a number of important phenomena in facial expression processing, holistic processing and visual expertise [Dailey and Cottrell, 1999, Cottrell et al., 2002, Dailey et al., 2002, Joyce and Cottrell, 2004]. Here, we investigate the model's ability to account for human sensitivity to variations in faces that are considered theoretically important for face identification. Face processing is typically described as *holistic* or *configural*. Holistic processing is typically taken to mean that the context of the whole face has an important contribution to processing the parts: subjects have difficulty recognizing parts of the face in isolation, and subjects have difficulty ignoring parts of the face when making judgments about another part. Configural processing means that subjects are sensitive to the relationships between the parts, e.g., the spacing between the eyes. We will use the two terms *configural* and *spacing* interchangeably in this paper. Holistic processing can easily be captured by a model that uses whole-face template-like representations as ours does: interference from incongruent halves of a face occurs when making judgments about a different part (e.g, expression on top when a different expression is on bottom [Cottrell et al., 2002]). However, configural effects related to spacing information are attenuated by the alignment procedure that we typically use, which warps the image so the eyes and mouth are always in the same three positions.

Diamond and Carey [Diamond and Carey, 1986] were among the first to discriminate between the types of processing involved in face/object perception and recognition. Based on studies looking at the inversion effect to faces, landscapes and dogs in both dog novices and dog experts, they proposed that first-order relational information, which consists of the coarse spatial relationships between the parts of an object (i.e. eyes are above the nose), is sufficient to recognize most objects. By contrast, second-order relational information, which is needed for face recognition and recognition of individuals within categories of expertise, is reserved for visually homogeneous categories where slight differences in configuration must be used to distinguish between individuals (e.g. a slight change in the distance between the eyes and the nose). Diamond and Carey [Diamond and Carey, 1986] suggest that experience allows people to develop a fine-tuned prototype and to become sensitive to second-order differences between that prototype and new members of that category (e.g. new faces).

One implication of the Diamond and Carey study is that the inversion effect (a large reduction in same/different performance on inverted faces, compared to inverted objects) is based on a relative reliance on second-order relational information, and that perhaps this characteristic distinguishes face/expert-level processing from regular object recognition. Farah et al. [Farah et al., 1995] found that encouraging part-based processing eliminated the inversion effect, whereas allowing/encouraging non-part-based processing resulted in a robust inversion effect. Thus Farah et al. conclude that the inversion effect, in faces and other types of stimuli, is associated with holistic pattern perception.

However, subjects are also quite sensitive to changes in the features themselves – substitutions of different eyes or mouths can make the face look quite different. The Thatcher illusion [Thompson, 1980] suggests that parts are processed somewhat independently, and only loosely connected to the representation of the whole face. Recently, a study by Mondloch et al. that varied these different aspects of a face (configuration, feature changes, and changes to contour of the face) found differing levels of sensitivity to the type of manipulation in a same/different paradigm. While the manipulations were not performed parametrically (no equating of the diffi-

culty of discrimination was performed), but in a rather ad hoc manner, the results are consistent across subjects. Hence this is a crucial set of data to account for with our model.

In the following, we describe Mondloch et al.'s experiments and our attempts to account for their data. We find that our model must be augmented with a representation of the parts of the face in order to account for most of the data. Finally, we discuss plans for future work.

Mondloch's Stimuli and Experiments

Mondloch et al. began with a single face (called Jane) and modified it to create twelve new versions (called Jane's Sisters). These were divided to three sets of stimuli: a configural set, a featural set, and a contour set (Figure 1). The four faces in the configural set were created by moving the eyes and/or the mouth. The four faces in the featural set were created by replacing Jane's eyes, nose and mouth with those of four different females. The four faces in the contour set were created by pasting the internal portion of Jane's face within the outer contour of four different females. The control stimuli were called "cousins" and consisted of three different female faces (Figure 2).

Figure 1: Jane is shown as the left-most face in each panel, along with her "sisters" from the configural set (panel A), the featural set (panel B), and the external contour set (panel C). (from [Mondloch et al., 2002])

Figure 2: The control stimuli: the cousin set. (from [Mondloch et al., 2002])

These stimuli were presented to 6, 8 and 10-year-old children as well as adults in a series of same-different trials. One face appeared for 200ms. After a 300ms interval, the second one appeared until the participant responded. There were also trials in which upside down versions of these faces were presented.

In this work, we concentrate on modeling the adult data, and hence focus on the black bars in (Figure 3). The results (Figure 3) showed that when stimuli were presented upright, the relative accuracy for adults in each set of stimuli was $cousin > featural > configural > contour$. This is interesting because it suggests that, at least for this stimulus set, subjects were more sensitive to individual feature differences than to configural changes. When the face images were presented upside down, however, the order was $featural > contour > configural$, and there was an inversion effect, i.e. the accuracy rate decreased. Note that the configural set, for which inverted accuracy was the worst, showed a larger inversion effect (measured by the mean accuracy of upright trials minus that of inverted trials) than the featural set.

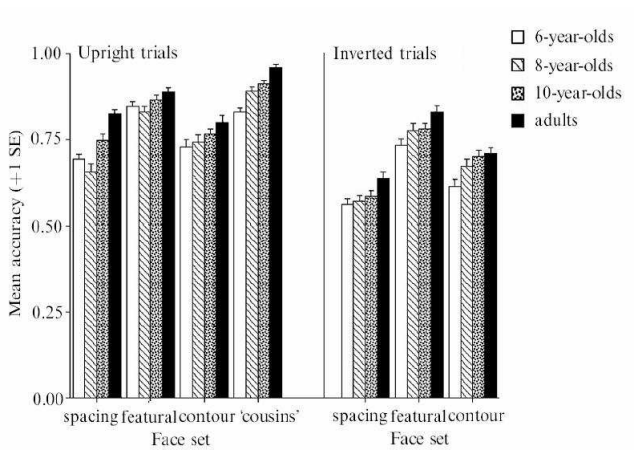


Figure 3: Mean accuracy for each face set and each age group when stimuli were presented upright (left panel) and inverted (right panel). (from [Mondloch et al., 2002])

A Computational Model of Face Recognition

Our model is a three level neural network that has been used in previous work (Figure 4). The model takes manually aligned face images as input. The images are first filtered by 2D Gabor wavelet filters, which are a good model of simple cell receptive fields in cat striate cortex [Jones and Palmer, 1987]. PCA (principal component analysis) is then used to extract a set of features from the high dimensional data. In the last stage, a simple back propagation network is used to assign a name to each face. We now describe each of the components of the model in more detail.

The Training Set

The FERET database is a large database of facial images, which is now standard for face recognition from still images[Phillips et al., 1998]. We used 662 face images (545 upright images of 117 individuals and 117 inverted images of 20 individuals (that were also included in the upright images)) in the training. The inverted faces were used in order to give a reasonable representation

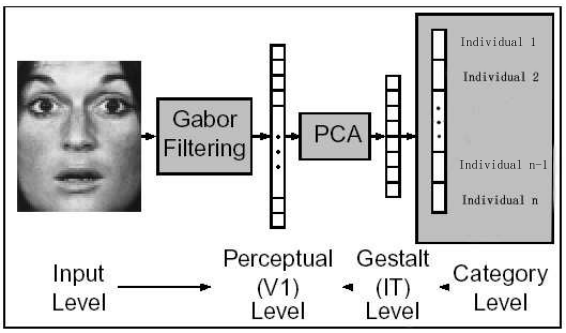


Figure 4: Object recognition model (from [Dailey et al., 2002])

of upside down faces in the PCA layer of the network. In [Dailey et al., 2002], where the task was to learn facial expressions, images were aligned so that eyes and mouth went to designated coordinates. This alignment removed the configural information which is crucial for our work because we are trying to understand how configural processing and featural processing interact with each other in the face recognition task. To avoid this negative effect, we required that the relative spacing between the parts of the face remain the same. The face images were rotated, scaled and translated so that the sum of square distance between the target coordinates and those of the transferred features (eyes and mouth locations) was minimized (Figure 5). Thus, a triangle represented by the eyes and mouth is scaled and moved to fit closely to a reference location, but the triangle is not warped. This way of alignment keeps configural information without affecting holistic processing. The aligned images were 192 pixels by 128 pixels.

Figure 5: Two examples of face image normalization. The faces were cropped with the eyes and the mouth as close as possible to the target position while keeping the shape of the triangle among these features the same.

Perceptual Layer

Research suggests that the receptive fields of the striate neurons are restricted to small regions of space, responding to narrow ranges of stimulus orientation and spatial frequency[Jones and Palmer, 1987]. DeValois et al[DeValois and DeValois, 1988] mapped the receptive fields of V1 cells and found evidence for multiple lobes of excitation and inhibition. Two-D Gabor filters [Daugman, 1985](Figure 6) have been found to fit the 2D spatial response profile of simple cells quite well[Jones and Palmer, 1987]. In this processing step the image was filtered with a rigid 23 by 15 grid of overlapping 2-D Gabor filters[Daugman, 1985] in quadrature pairs at five scales and eight orientations [Dailey et al., 2002](Figure 7). We thus obtained $23 \times 15 \times 5 \times 8 = 13,800$ filter responses in this layer, which is termed the *perceptual* layer [Dailey et al., 2002].

Figure 6: A Gabor function is constructed by multiplying a Gaussian function by sinusoidal function[Daugman, 1985]. We use five scales and eight orientations.

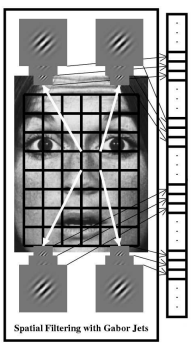


Figure 7: An image filtered with a rigid 23 by 15 grid of overlapping 2-D Gabor filters in quadrature pairs at five scales and eight orientations (from [Dailey et al., 2000])

Gestalt layer

In this stage we perform a PCA of the Gabor filter responses. This is a biologically plausible means of dimensionality reduction[Dailey et al., 2002], since it can be learned in a Hebbian manner. PCA extracts a small set of informative features from the high dimensional output of the last perceptual stage. The eigenvectors of the covariance matrix of the patterns are computed, and the patterns are then projected onto the eigenvectors associated with the largest eigenvalues. At this stage, we produce a 50-element PCA representation from the 13,800 Gabor vectors. Before being fed to the final classifier, the principal component projections are shifted and scaled so that they have 0 mean and unit standard deviation, known as z-scoring (or whitening).

Categorization layer

The classification portion of the model is a two-layer back-propagation neural network. 20 hidden units are used. A scaled tanh [LeCun et al., 1998] activation function is used at the hidden layer and the softmax activation function $y_i = e^{a_i} / \sum_k e^{a_k}$ was used at the output layer. The network is trained with the cross entropy error function [Bishop, 1995] to identify the faces using localist outputs. A learning rate of 0.05 and a momentum of 0.5 were used in the results reported here. 10 percent of the images are selected randomly as a test set and another 10 percent as a holdout set [Dailey et al., 2000]. The network achieves 85-90 percent accuracy within 50

epochs. This is remarkable given that for faces in the test set, there were only 2-3 images in the training set on average. This classification rate was decent enough to show that our model represented face images well.

Modeling Mondloch et al.

Training and Learning

For the following experiments, we simply trained the network on all 662 images, since we are only interested in obtaining a good face representation at the hidden layer. Training was stopped at the 50th epoch based on the above pilot experiment, as we assumed the network had achieved "adult" level identity recognition expertise at this point. After training, the preprocessed Jane stimuli images were presented to the network.

Modelling Discrimination

Hidden unit activations were recorded as the network's representation of images. In order to model discriminability between two images, we present an image to the network, and record the hidden unit response vector. We do the same with a second image. We model similarity as the correlation between the two representations, and discriminability as one minus similarity. Note that this measure may be computed at *any* layer of the network. We computed the average discriminability between images in each of the stimuli sets (featural, configural, etc., both upright and inverted). The average within each set was taken as the measure of the network's ability to discriminate each set. The average of the discriminabilities was computed over 50 networks which were all trained in the same way, but used different initial random weights.

The results (Figure 10 top graph) showed that our model was too holistic, i.e. the model showed high sensitivity to the configural set. As a first pass at adding featural information to the model, we took a cue from Padgett and Cottrell (1998), who developed a parts-based model for facial expression recognition. They simply had rectangular windows over the eyes and mouth and extracted features from those as input to a classifier. Similarly, Pentland et al. (1994) used "eigenfeatures", PCA of local patches, as input to a face identification classifier. From our grid of Gabor filters, we extracted three sets of Gabor responses that corresponded to the left eye, the right eye and the mouth respectively (Figure 8). A 10 dimensional PCA representation was extracted from each of them. Then we gave both the global and local PCA to the network as input.

We repeated this experiment multiple times, keeping the 30 local feature principal components (PC's) as input to the network, while varying the number of global PC's. The results (Figure 10) show how different combinations of global and local PC's affect the behavior of the network. The graph on the top is the result of the original model (50 global PC's with no local PC's). The graph second from the top is the result of 50 global PC's plus 30 local PC's. The remaining graphs show the effects of progressive reduction in the number of global PC's from 30 to 0 in steps of 10, while holding the number of local PC's constant at 30. When the number of

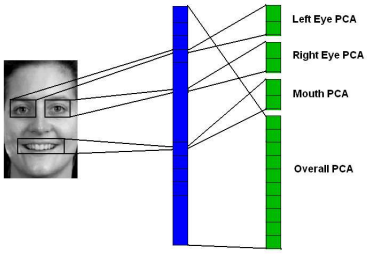


Figure 8: We extracted local PCA representations for the eyes and the mouth. The responses of Gabor filters from patches around the eyes and mouth were extracted and PCA was done on them separately.

global PC's is decreased below 20, the discriminability of the feature set began to exceed that of the configural set in the upright image trials.

Note that the local feature PC's did help the model pay more attention to features because the discriminability of the feature set has increased. Also, when the number of the global PC's was reduced, the discriminability of the feature, configural, and cousin sets increased. The discriminability of the cousin set started around 0.35 when 50 global components with 30 local components were used and ends up at around 0.45 when no global components were used. We can observe a gradual increase in discriminability over the sequence of the graphs from top to bottom. This gradual increase is also seen for the configural set and the feature set, which each grew from around 0.2 to 0.3. Further, the qualitative pattern for the inverted faces is reproduced in almost every variation.

Discriminability at processing stages

Where do these effects come from? Recall our definition of discriminability: one minus similarity, where similarity is equal to the correlation between representations. Hence, we can assess similarity and discriminability at each stage of processing, i.e., original images, aligned images, Gabor filter, PCA, z-score PCA. Note that for preprocessing stages, we are only comparing discriminability between a small number of images (Jane and her sisters), because these stages are identical for all 50 networks.

The order of discriminability for all combinations of local and global PC's and for both image orientations is the same for the first three stages. The order of the sets does not change until the PCA and z-score PCA stages. Figure 9 shows the discriminability of each set of different combinations of global PC's and local PC's at the PCA level and the z-score PCA level for upright images. When there are no local PC's (i.e., the original model), the configural set exceeds the feature set. When there are 30 local PC's and 50 global: the order is correct ($cousin > feature > configural > contour$) at the PCA stage, though the differences are very small. These differences are enlarged at the z-scored PCA stage. As reductions in the number of global PC's leave proportionally more local PC's, we observe the same correct

ordering and progressively larger differences between the sets at these last two stages. Also there is a trend towards increased discriminability for cousin, featural and contour sets.

A change in set order can also be observed at the PCA and z-score PCA stages for the inverted image results (not shown in figures here). The configural set shows a larger inversion effect than the feature set, which is consistent with human data. We also observe an increasing gap between the featural set and the configural set ($featural > configural$) when the local PC's are introduced and as their proportion is subsequently increased (as the number of global PC's is reduced). However, the contour set is always less discriminable or at most as discriminable as the spacing set, which is the wrong order – contour should be more discriminable than spacing in inverted images. The correct ordering shows up in the hidden layer for all networks except the ones with no global PCA or no local PCA (see Figure 10), suggesting that both are needed.

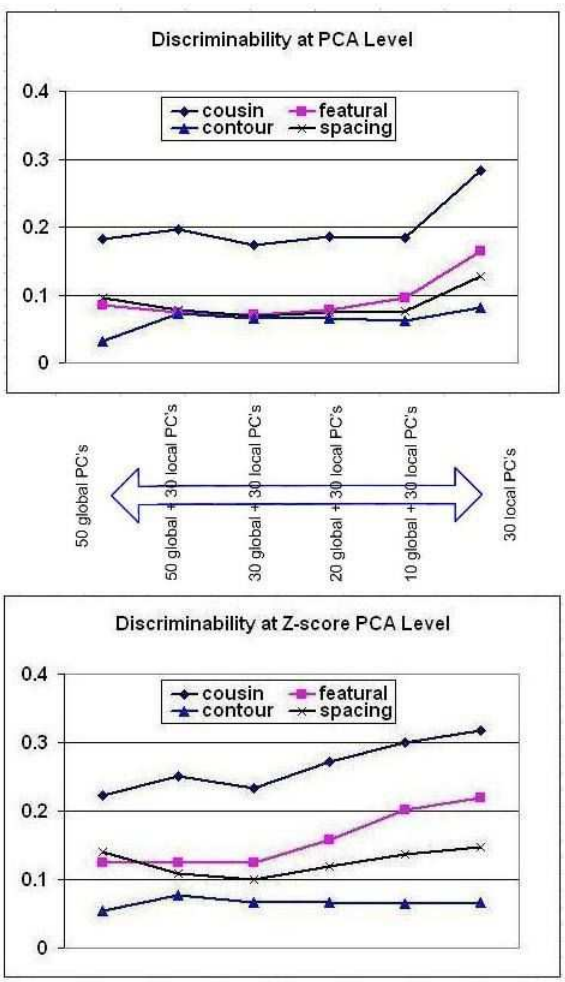


Figure 9: The discriminability of different combination of global PC's and local PC's at the PCA and the z-scored PCA level.

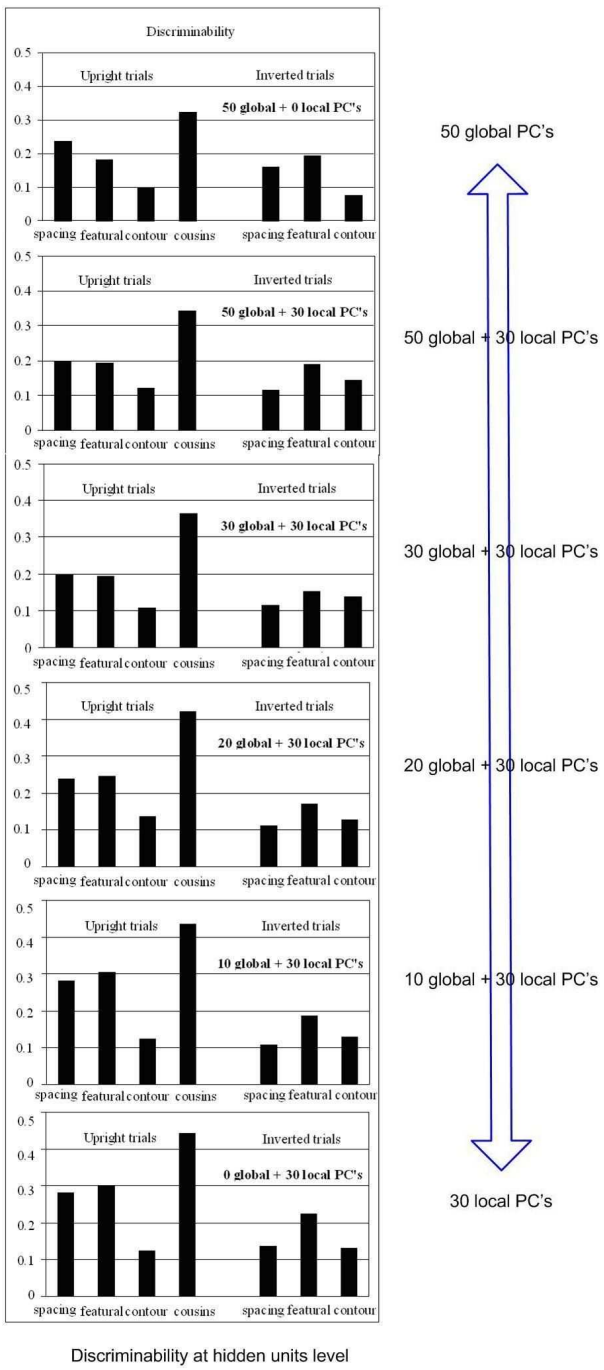


Figure 10: The discriminability of different combination of global PC's and local PC's at the hidden layer.

Discussion

While our standard model has accounted for a fair amount of data over the years, this particular set of data required substantial modifications. We found that our original model was too holistic, in that it was more sensitive to configural changes versus featural changes. This is not surprising given the way the model is constructed. Global PCA of the Gabor representation should act similarly to global PCA of grayscale images. This representation is known to develop ghostly-looking, whole face templates that we have called holons, and others have termed eigenfaces. These representations have proved to be very useful in modeling holistic processing effects. For example, when two halves of different faces are aligned, it is more difficult for the model to identify the top half of a face due to interference from the bottom half, even if the input from the bottom half is severely attenuated to simulate attention to the top [Cottrell et al., 2002]. This is due to the bottom half of the face matching giving a partial match to the templates corresponding to the other person's face.

Adding a parts-based representation, here implemented as a local feature PCA, turned out to be helpful in making the model more sensitive to features. This type of representation can be thought of as a schema for each part. It could be developed through attending to parts of the face, where the parts become well-represented via foveation. As proportionally more of this representation was used, the network's upright discriminability profile qualitatively matched the human subjects results.

Our model successfully showed inversion effects on the configural set and the featural set. This effect on the configural set was especially large, which is consistent with human behavior. The order for inverted trials qualitatively matched the human subjects results when both global and local components were used. While the model showed a strong inversion effect on the configural set, the model did not show any inversion effect on the contour set. This suggests that our model used the information mostly, if not entirely, from the inside of the face instead of the contour. Infants, on the other hand, are known to use the contour of the face before they are able to use the inside of the face for recognizing their mothers. In the future, we intend to add a developmental component to our model, in order to model this "outside-in" progression.

Acknowledgement

We thank Carrie Joyce and Matthew N. Dailey for previous discussions, Gary's Unbelievable Research Unit (GURU) for valuable comments, Daphne Maurer for Jane's data sets and anonymous reviewers for helpful suggestions. This research was supported by NIMH grant MH57075 to GWC.

References

[Bishop, 1995] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

[Cottrell et al., 2002] Cottrell, G. W., Branson, K. M., and Calder, A. J. (2002). Do expression and identity need separate representations? In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, Mahwah, New Jersey. The Cognitive Science Society.

[Dailey and Cottrell, 1999] Dailey, M. N. and Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, 12:1053–1073.

[Dailey et al., 2000] Dailey, M. N., Cottrell, G. W., and Adolphs, R. (2000). A six-unit network is all you need to discover happiness. In *TwentySecond Annual Conference of the Cognitive Science Society*.

[Dailey et al., 2002] Dailey, M. N., Cottrell, G. W., Padgett, C., and Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, 14(8):1158–1173.

[Daugman, 1985] Daugman, J. G. (1985). Uncertainty relation for resolution in space, spacial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of American A*, 2:1160–1169.

[DeValois and DeValois, 1988] DeValois, R. L. and DeValois, K. K. (1988). *Spatial Vision*. Oxford University Press.

[Diamond and Carey, 1986] Diamond, R. and Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimantal Psychology: General*, 115(2):107–117.

[Farah et al., 1995] Farah, M., Levinson, K., and Klein, K. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33:661–674.

[Jones and Palmer, 1987] Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258.

[Joyce and Cottrell, 2004] Joyce, C. and Cottrell, G. W. (2004). Solving the visual expertise mystery. In *Proceedings of the Neural Computation and Psychology Workshop 8*, Progress in Neural Processing. World Scientific, London, UK.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). Efficient backprop. In *Neural Networks—Tricks of the Trade, Springer Lecture Notes in Computer Sciences*, volume 1524, pages 5–50.

[Mondloch et al., 2002] Mondloch, C. J., Grand, R. L., and Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. *Perception*, 31:553–566.

[Padgett and Cottrell, 1998] Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorial perception of facial expressions. In *Proceedings of the Twentieth Annual Cognitive Science Conference*.

[Pentland et al., 1994] Pentland, A. P., Moghaddam, B., and Starner, T. (1994). View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 84–91, Los Alamitos. IEEE.

[Phillips et al., 1998] Phillips, J., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306.

[Thompson, 1980] Thompson, P. (1980). A new illusion. *Perception*, 9:483–484.

Member Abstracts

Expanding the linguistic coverage of a spoken dialogue system by mining human-human dialogue for new sentences with familiar meanings

Gregory S. Aist (gaist@cs.rochester.edu)[1]
James Allen (james@cs.rochester.edu)[1, 2]
[1]Computer Science Department, University of Rochester
P.O. Box 270226, Rochester, NY 14627 USA

Lucian Galescu (lgalescu@ihmc.us)[2]
[2]Institute for Human and Machine Cognition
40 South Alcaniz Street, Pensacola, FL 32502 USA

How can a system grasp linguistic variety?

Computer systems that interact with people using natural spoken dialogue offer many possibilities for effective and efficient interaction. But, a dialogue system for a new domain requires a great deal of expert attention to either collecting data for a new domain or designing a model of the language that people use when solving problems in that domain. Consider the extension of a checkers-playing dialogue system to other games such as Go: game-specific phrases and terms would need to be added – both "official" and colloquial versions. Prior work here includes identifying new lexical items, new word-sequence correlations (Galescu, Ringger, & Allen 1998), or new phrase patterns using existing words or word classes such as *color* or *animal* (Bulyko, Ostendorf, & Stolcke 2003). In this paper, we take a slightly different approach by focusing on semantics. Given a specification (as a sample dialogue), we want to identify alternate ways of talking about the same things. These alternates may not necessarily use any of the words that their counterparts in the script use. In fact, the more dissimilar such new phrases are in terms of surface features, the more helpful they would be if added to the vocabulary and syntax that the system can understand. Such a technique should prove useful when expanding the linguistic coverage of a dialogue system to cover what people say in practice. We describe these techniques in the context of an equipment purchasing system – part of the multisite CALO project, an intelligent personal assistant.

Start with a script

Any dialogue system typically starts with some data collection or user interviews or some other technique designed to give a basic idea of what the users of the system will eventually say. The requirements for the system can then be easily expressed in terms of a dialogue script. The initial script for our domain of computer purchasing began:

System: Hello, this is CALO. User: Hello CALO.
S: Hello. U: I would like you to buy a computer for me.
S: Ok; what kind of computer would you like?

Add human-human dialogues

As part of the dialogue system effort, Rochester collected a set of approximately 40 human-human dialogues carried out by ~20 people playing the role of "buyer" and ~4 people playing the role of "agent". In order to collect dialogue aimed at the computer purchasing domain, we provided a short scenario to role-play when buying the first computer - "You are a small business owner looking to buy a computer", etc. The second scenario was self-directed, that is, "Now get a computer for yourself." These dialogues were recorded and transcribed. Those utterances that closely matched lines from the script we used directly as additional training data. Those utterances about capabilities beyond the scope of the initial system – such as warranty purchases – we reserved for later use. That left a wide range of utterances concerning concepts that are present in the script, but using different words and syntax. We wanted those.

Mine them for semantic matches to script lines

We used Latent Semantic Analysis (http://lsa.colorado.edu) to extract utterances from the human-human dialogues that were similar to each line in the script. For each utterance in the dialogues, we calculated its similarity to each line in the script, and assigned it to the line with the highest similarity. We then hand-filtered the resulting data to yield new ways of saying lines in the script, such as:

On a similar topic, but system initiative rather than user:
Script/User: I would like you to buy a computer for me.
Dialogue/*System: Hello would you like to buy a computer*

With similar meaning, but very different words and syntax:
Script/*S: Ok, I'll start looking.*
Dialogue/S: I'll go ahead and uh save this.

Less similar in meaning, but still reasonable alternatives.
For example, after "*Ok; what kind of computer would you like?"*:
Script/U: A lightweight laptop computer with 500 mb of ram.
Dialogue/U: Pentium processor.

From the utterances paired with the 28 lines in the script, 41 were identified to be useful matches. CALO developers at Rochester and IHMC used these to extend the language understanding of the CALO dialogue system.

References

Bulyko, I., Ostendorf, M., & Stolcke, A. (2003). Getting more mileage from Web text sources for conversational speech language modeling using class-dependent mixtures. *Proceedings of the North American Association for Computational Linguistics*. Edmonton, May 2003.

Galescu, L., Ringger, E., & Allen, J. (1998). Rapid language model development for new task domains. *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain, May 1998.

A COMPUTATIONAL FRAMEWORK FOR THE STUDY OF COLLABORATIVE LEARNING

Rick Alterman (alterman@cs.brandeis.edu)
Svetlena Taneva (svet@cs.brandeis.edu)
Computer Science Department, Volen Center for Complex Systems, Brandeis University
415 South Street., Waltham, MA, 02454 USA

There are a large number of research questions on collaborative learning that cannot easily be answered by simply collecting and comparing quantitative data on the performance of individual and collaborative learners. More often than not, critical information is available only if access to a detailed recording of the students' collaborative work is available. For certain research endeavors in the collaborative learning domain, we need to know how the students organized their task, what roles the students played, and how they participated. Videotaping is an example of a technology that has been used to investigate collaborations within the workplace. But there are some difficulties with using this approach for the studying of collaborative learning, not the least of which is the extensive time-cost of collecting and transcribing video data. We argue that groupware applications are an ideal platform for experimental investigations of collaborative learning.

At Brandeis we have been developing principles, tools, and methods for cognitively engineering groupware systems that support online collaboration. One part of this project is to develop a toolkit that enables the rapid development of groupware applications that can be used as experimental platforms. A key component of the groupware systems that are generated is that a complete transcript of all the online activity is automatically captured in a form that is replayable by an analyst using a replay device (that is created as the system is developed). Students at Brandeis have successfully used this toolkit in a HCI class to produce a working groupware system; each team of students had 28 days to write the code. Another part of our research project is to develop discourse analysis techniques for modeling the online collaborative work of users and the cognitive load it entails for each of the participants. These techniques have been taught in a class on Computational Cognitive Science.

In this poster, we will present some of the details of an experimental study of collaborative learning that we are currently conducting. Some of the questions about collaborative learning that we want to investigate are:

- How does the amount and type of participation affect individual learning?
- What do the participants talk about (i.e. which aspects of the activity do they spend the most effort on)?
- How do the participants organize their collaboration?
- How closely do the participants work together?

Corresponding to each one of these questions are significant hypotheses about the role of participation and/or explanation in collaborative learning tasks.

Our study compares the performance of individuals and pairs of students (with little or no prior programming experience) as they learn to draw figures using JScheme. As a part of this study, we constructed a platform (GrewpTool) for collaborative programming that has been used to support several kinds of classroom related activities. In our experimental study, GrewpTool collects, in a replayable transcript, the representational work of individual subjects and all of the communication between paired subjects. The participants in our experiment complete both a pre and a post test, whose score difference indicates how much they learn.

Our study has produced an enormous amount of data for analysis. The replay of an individual session of collaboration is one of the tools available for analyzing the interactional data. Given this tool, extracting specific and accurate answers to questions about participation is feasible, but the data is not easily codable and the task is labor intensive.

We have developed some automatic methods for analyzing the interaction that can be used to guide the ethnographic analysis of the subjects' online behavior. Each of these representations is relevant to answering questions of the sort listed above. We will show automatically generated representations that depict how close each pair worked together, how they organized their collaboration, the type and amount of each subject's participation, and the content of their conversation. Each of these representations can be created because all of the subjects' participation is mediated by the computer and therefore automatically recorded and transcribed. Given this analysis of the data, it is possible to more selectively engage in the labor-intensive task of analyzing the replay of each session. Given these kinds of representations of data it is easier to explore the effects of explanation, participation, and organization on learning.

Acknowledgments

This work was supported by the National Science Foundation under Grant EIA-0082393, and by the Office of Naval Research under Grant N000-14-02-1-0131.

Moral Cognition: A Dual-Process Model

Eric C. Anderson (ecanderson@hampshire.edu)
School of Cognitive Science, Hampshire College
893 West St., Amherst, MA 01002 USA

Introduction

For centuries philosophers have debated the roles that intuition and reasoning play in making moral judgments. Hume proposed that people possess a 'moral sense' that allows them to distinguish between right and wrong, while Kant argued that reason is, or should be, the basis for making moral judgments.

In the last 40 years psychologists have addressed this question using scientific methods. Kohlberg (1986) emphasized moral reasoning in his cognitive developmental stage model which posits that people progress through a series of six universal stages of moral reasoning. Haidt (2001) recently proposed a dual-process model of moral cognition that rejects reason as the primary cause of moral judgments and posits intuition as more influential. Intuition is characterized as fast, automatic processing that is not available to introspection. Reasoning, on the other hand, is characterized as slow, deliberate processing. While reasoning, people are aware of progressing through a series of steps to generate a judgment (Haidt, 2001; Kahneman, 2003).

We propose a modified version of the social intuitionist model. This dual-process 'interactionist' model proposes both intuition and reasoning processes are used, but their use depends on the situation being judged. Specifically, the model proposes that reasoning processes are used when people have no intuitions or when they have conflicting intuitions. However, when one dominant intuition is generated, reasoning processes are not engaged.

Methods

To asses the interactionist model, an experiment was conducted in which 60 participants read moral dilemmas used by Greene et al. (2001) and made judgments about them in two time conditions. In the fast time condition, participants made judgments immediately without reasoning about the dilemma, while in the slow time condition, participants had as much time as needed to make judgments. If reasoning processes are important, then disrupting that process should decrease people's ability to make judgments. Participants were tested on three different types of dilemmas: dilemmas that produced no intuitions; dilemmas that produced conflicting intuitions; and dilemmas that produced one intuition.

Findings

The findings support the interactionist position by suggesting that participants used intuition and reasoning differently depending on the dilemma being judged. In situations that were predicted to require reasoning (no-intuition dilemmas and strong-conflicting intuition dilemmas), participants did poorly in the fast time condition (Figure 1). However in the dilemmas when reasoning was hypothesized to be causally inert (no-conflicting intuition dilemmas), there was no difference in participants' performances between the time conditions. These findings call into question both strong intuitionist and strong rationalist positions but support the dual-process interactionist model.

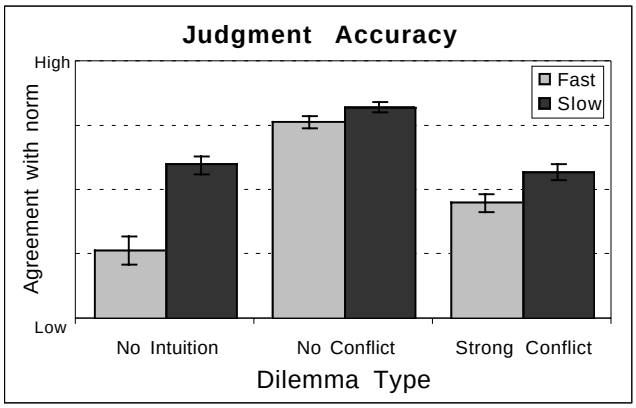


Figure 1: Judgment accuracy based on agreement with the empirically established moral norm.

Acknowledgments

Thank you to Neil Stillings, Laura Sizer, Ernest Alleva, and Phillip Kelleher. This study was funded by the Culture, Brain, and Development program at Hampshire College.

References

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science,* 293, 2105-2108.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review,* 108 (4) 814-834.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58 (9) 697-720.

Kohlberg, L. (1986). A current statement on some theoretical issues. In S. Modgil & C. Modgil (Eds.) *Lawrence Kohlberg: Consensus and controversy*. Philadelphia & London: Falmer Press.

Empirical Results for the Use of Meta-language in Dialog Management

Michael L. Anderson (anderson@cs.umd.edu)
Institute for Advanced Computer Studies, University of Maryland
College Park, MD 20472

Bryant Lee (blee3@wam.umd.edu)
University of Maryland
College Park, MD 20742

Introduction and Background

As is well known, dialog partners manage the uncertainty inherent in conversation by continually providing and eliciting feedback, monitoring their own comprehension and the apparent comprehension of their dialog partner, and initiating repairs as needed (see e.g., Cahn & Brennan, 1999; Clark & Brennan, 1991). Given the nature of such monitoring and repair, one might reasonably hypothesize that a good portion of the utterances involved in dialog management employ meta-language. But while there has been a great deal of work on the specific topic of dialog management, and it is widely (if often tacitly) accepted that meta-language is frequently involved, there has been no work specifically investigating and quantifying the role of meta-language in dialog management. Thus, this small study investigated the correlation between meta-language and dialog management utterances in three dialog files of the British National Corpus (BNC).

Approach and Methods

The three BNC files used in this study, KRF, KRG, and KRH, are transcripts of a series of *Ideas in Action* radio programs, some of which are interviews. Because interviews are more structured than informal conversation, they involve explicit dialog management, and are therefore a good place to start an investigation into the relation between meta-language and dialog management. Focusing exclusively on the interviews in these three files gives 5900 lines to study.

These three files had been previously annotated for meta-language, using the annotation scheme and methods reported in (Anderson, *et al*., 2004).

To identify and annotate the dialog management utterances, we were guided by an analogy with the TRAINS domain and dialogs (Gross, Allen & Traum, 1993). In the TRAINS domain, the base-level actions are moving trains between cities, and the assigned task is to plan and manage these moves through cooperative dialog. In our case, we defined the interview itself as the task domain, the base-level actions as utterances, and the task as planning and managing these base-level actions, i.e. planning and managing the interview itself. As in TRAINS, this management is accomplished through dialog. The utterances involved in planning and managing the interview were identified and annotated according to Dialog Act Markup in Several Layers (DAMSL) (Allen & Core, 1977).

We are still analyzing the results of this annotation for specific correlations between meta-language and the different DAMSL information levels and functions. However, we report some preliminary results, below, for the overall relation between dialog management and meta-language.

Results

Of the 5900 lines annotated, 581 were dialog management utterances, and 1020 included meta-language. 312 lines were both dialog management and meta-language.

Table 1: Meta-language and dialog management results

	Meta	-Meta	Totals
DM	312	269	581
-DM	708	4611	5319
Totals	1020	4880	5900
$X^2 = 597.56$ $p << 0.001$ $\Phi = 0.318247$			

Thus, 53.7% of dialog management utterances involved meta-language. To the best of our knowledge, this is the first quantitative confirmation of the tacitly held assumption that meta-language is frequently involved in dialog management. Detailed results can be found at http://www.cs.umd.edu/projects/metalanguage

Acknowledgments

This work was supported in part by a grant from the ONR.

References

Allen, J. & Core, M. (1977). DAMSL: Dialog Annotation Markup in Several Layers. Technical report, University of Rochester.

Anderson, M., Fister, A., Lee, B., Tardia, L. & Wang, D. (2004). On the types and frequency of meta-language in conversation: A preliminary report. *Proceedings of the 14th Annual Meeting of the Society for Text and Discourse*.

Cahn, J. E. & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, (pp. 25–33).

Clark, H. & Brennan, S. E. (1991). Grounding in communication. In: J. Levine, L. B. Resnik & S. D. Teasley (Eds.) *Perspectives on Socially Shared Cognition*.

Gross, D., Allen, J. F. & Traum, D. R. (1993). The TRAINS 91 Dialogues, TRAINS Technical Note 92-1, Computer Science Dept., University of Rochester.

Working Memory and Virtual Endoscopy Simulation

Pehr Andersson[1] (pehr.andersson@psy.umu.se)
Leif Hedman[1, 2] (leif.hedman@psy.umu.se and leif.hedman@cfss.ki.se)
Lars Enochsson[2] (lars.enochsson@cfss.ki.se)
Pär Ström[2] (per.strom@cfss.ki.se)
Ann Kjellin[2] (ann.kjellin@cfss.ki.se)
Bo Westman[2] (bo.westman@cfss.ki.se)
Li Felländer-Tsai[2] (li.tsai@cfss.ki.se)
[1] Department of Psychology, Umeå University, SE-901 87 Umeå, Sweden, [2] Center for Surgical Sciences,
Center for Advanced Medical Simulation,
Karolinska Institutet at Huddinge University Hospital, SE-141 86 Stockholm, Sweden

Introduction

We report on a study that investigates the relationship between visual working memory and verbal working memory and a performance measure in endoscopic instrument navigation in GI Mentor II (a simulator for gastroscopic surgery). Baddeley's (1998) three-component model of the working memory contains a central executive and two subsidiary slave systems – the phonological loop and the visuo-spatial sketchpad. Both the visual working memory, rehearsal processed based, and the verbal working memory falls within the phonological loop category – when using digit span tasks. We hypothesize that the visual working memory test scores will correlate with the simulator performance measure. We also predict that tasks involving a higher degree of central executive functions will be more discriminate in correlation with the performance measure. We have added the verbal working memory task to distinguish and rule out verbal working memory as a major factor for predicting endoscopic simulator performance – and thus define and elucidate the visual working memory phonological loop function for predicting endoscopic simulator performance.

Method

22 medical students (novices, 11 women and 11 men), ranging in age between 22 and 40 years at Karolinska Institutet, Huddinge University Hospital in Sweden, participated in the study. All participants completed a one hour session in the MIST-VR simulator and the GI Mentor II simulator. We will only present findings from the GI Mentor II simulator. The selected GI Mentor performance score measures the efficiency of screening (ES). The working memory (WM) tasks were taken from the WAIS III test battery (Wechsler, 2003): Forward digit span task (FDS), backward digit span task (BDS) and an alphanumerical task (ANT) – with the forward digit span being the least demanding on the central executive function and with the alphanumerical task being the most demanding.

Results & Discussion

There were significant Pearson's r correlations found between the visual working memory test scores and the simulator performance score and between the verbal visual working memory test score ANT and the simulator performance score.

Table 1. Pearson's *r* correlations between performance score (ES) in the GI Mentor II simulator, Visual Working Memory scores and Verbal Working Memory Scores.

Visual WM task	Verbal WM task
Fds *r*=.607, p=.031	Fds *r*=.306, p=.166
Bds *r*=.570, p=.043	Bds *r*=.324, p=.152
Ant *r*=.617, p=.029	Ant *r*=.639, p=.013
ES	ES

Our findings suggest that visual working memory correlates with the simulation performance measure. The verbal working memory is dependent on more advanced central executive functions to discriminate significant differences for the performance tasks while the visual working memory tasks show a more uniform result regardless of central executive effort. An extension of this study is currently exploring these findings further

Acknowledgment

The first and second authors were supported by grant No. 220-155600 from EU Goal 1 North of Sweden.

References

Baddeley, A: (1998) Working memory. *Life Sciences*., 321, pp. 167-173.

Wechsler, D. (2003) *Wais III. Manual*. San Antonio, TX: The Psychological Corporation.

Learning to Categorize in the Context of Item Triples

Janet K. Andrews (andrewsj@vassar.edu) & Kenneth R. Livingston (livingst@vassar.edu)
Department of Psychology and Program in Cognitive Science, Vassar College
124 Raymond Avenue, Poughkeepsie, New York 12604 USA

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, University of Binghamton
P. O. Box 6000, Binghamton, New York 13902 USA

Background

We evaluate the relative contributions of two mechanisms of category learning: 1) *abstraction* across examples of the same category and; 2) *differentiation* between examples of different categories. A novel "triples" paradigm is introduced in which each classification target is presented with two different context items. Learners are informed of the structure of the triples so they may take advantage of knowledge about the relative category status of the items. We use feature-based categories with perceptually subtle variation among examples. The study is designed to advance a naturalistic yet controlled basis for the study of category learning by using multiply-instantiated feature values (Markman & Maddox, 2003) and three-way rather than binary classification decisions.

In a control condition, items were presented one at a time. Learning was also tested under five experimental conditions based on the following triple structures: aAA (both context items match the category of the target 'a'), aAB (one matching and one mismatching context item), aBB (both context items mismatch the target, but the context items match one another), aBC (both context items mismatch the target and the context items also mismatch one another), and aXX (no systematic structure).

One possible learning strategy is to locate common features between items known to belong to the same category and perform abstraction -- in which case the aAA group should have an advantage. Another potential strategy is to identify contrasts between items known to belong to all three categories -- in which case the aBC structure should be most beneficial. The aAB structure is least informative under either strategy because the learner is unable to know for certain whether any pair within the triple are in the same or different categories. The aBB group benefits from weaker forms of both abstraction and differentiation on each trial.

Method

The stimuli consisted of organism-like patterns created in Adobe Photoshop that varied systematically along three dimensions: body-aspect ratio, flagella length, and stripe width. Each dimension had eight possible values. Three categories called Gex, Kij, and Zof were defined using the higher or lower four dimension values (e.g., Zofs had rounder bodies, longer flagella, and wider stripes) for a total of 192 possible items. Each category was distinct from the other two in terms of exactly one dimension.

Each of sixty-six college students was randomly assigned to one of the six conditions described above (single item control condition, aAA, aAB, aBB, aBC, or aXX) and tested in two phases, a training phase of 144 trials with feedback given on target classification responses, and a test phase without feedback and using all 192 stimuli. Except for the single target control condition, stimuli were always presented in the triples context, and the structure of the triple was carefully explained to participants at the outset.

Results

The manipulation significantly affected accuracy and speed of performance during the first (training) phase of the experiment ($F(5,60) = 3.423$, $p = .0087$), with the aBC condition yielding the best performance (83% correct overall, with chance performance of 33%). As expected the least learning took place in the aAB group (55%). To our surprise, the aAA group was also quite low (63%) and did not differ significantly from the aAB group. Performance was intermediate in the aBB (74%), single item control (73%), and aXX (69%) conditions, which did not differ significantly. Examining performance over the course of 12 blocks of 12 training trials, the aBC condition was most accurate for every single block. Overall accuracy in phase two ranged from 61% (aAB) to 78% (aBC), but there was no main effect of condition at that point.

These results indicate that, at least in the early stages of learning in this context, between-category differentiation is more important than within-category abstraction. Additional experiments are underway to explore whether removing the information given about the triples structure, or highlighting it more dramatically, will alter the outcome.

Acknowledgments

Our thanks to the Undergraduate Research Summer Institute at Vassar College, and to Jessica Cicchino, Emma Myers, and especially Peter Alfaro for assistance with this study.

References

Markman, A. B., & Maddox, W. T. (2003). Classification of exemplars with single and multiple feature manifestations: The effect of relevant dimension variation and category structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 107-117.

Semantic Complexity and Language Production: Simple vs. Complex Verbs

Kathleen T. Ashenfelter (Targowski.1@nd.edu)
Department of Psychology, University of Notre Dame
Notre Dame, IN 46656

Kathleen M. Eberhard (Eberhard.1@nd.edu)
Department of Psychology, University of Notre Dame
Notre Dame, IN 46656

Introduction

Because language is a discrete combinatorial system in which smaller representations at one level combine to form larger representations at another level, verbs can differ with respect to the number of smaller representations that comprise them. The semantic and morphological levels are most closely related to a verb's meaning, and it has been shown that lemmas that share overlapping semantic features compete for production when encoded in the same local context (e.g., Breedin, Saffran, & Schwartz, 1998). This study investigated the competition between complex and simple verbs using a speech error elicitation task intended to induce contextual errors where one verb replaces another in a perseveration, anticipation, or complete exchange (e.g., Baars. 1992). Here, complex verbs contained all of the same semantic features as a simple verb in addition to additional features at the semantic and morphological levels. Two experiments tested the hypothesis that more complex verbs would replace their simpler counterparts more often than vice versa in contextual errors due to being associated with a greater number of activated semantic features at the point of lemma selection.

Experiment 1

Experiment 1 investigated the antonymic contrast between verbs where one is semantically and/or morphologically marked. For instance, in the semantic + morphological condition, the verb UNTIE has all of the features of TIE plus a negation feature at the semantic level and an additional morpheme at the morphological level. In the semantic only condition, the verb DECODE contains all of the semantic features of ENCODE, but does not consist of an extra morpheme at the morphological level. It was predicted that complex verbs in both conditions would replace their simpler counterparts (denoted *simple* → *complex*) in contextual errors more often than vice versa. An asymmetry was also predicted such that the additional morpheme and semantic features of the complex verb in the semantic + morphological condition would result in an even greater number of *simple* → *complex* speech errors than in the semantic only condition.

The results of Experiment 1 are listed in Table 1. As predicted, the effect of error type was significant, as was the interaction between error type and complexity condition. There is evidence that the additional morpheme in the semantic + morphological condition gave the complex verb lemmas an additional advantage over their simpler counterparts.

Experiment 2

This experiment examined the contrast between simple and complex verbs that differed with respect to an added feature of manner specification as well as morphological aspect features. The lemma representations of complex verbs like JOG contain all of the semantic features of the lemma representation of a simple verb like GO plus a specification of manner. In addition, the morphological features of progressive aspect were added to half of the complex verbs (i.e. "is jogging"). Again, more *simple* → *complex* errors were predicted than the reverse, and the added morphological features were expected to enhance this effect.

The results of Experiment 2 are listed in Table 1. The effect of error type was significant both conditions, as was the effect of complexity condition. The interaction was not significant, indicating that the additional morphological aspect features in the semantic + morphological condition did not contribute to the complex verb's lemma activation beyond that of the additional manner feature.

Table 1. Results for Experiments by Subject (p-values)

Effect	Experiment 1	Experiment 2
Error Type	<.001	<.001
Complexity Cond.	>.05	.04
Interaction	.02	>.05

References

Baars, B. J. (1992). A dozen competing-plans techniques for inducing predictable slips in speech and action. In B. J. Baars (Ed.), *Experimental Slips and Human.* New York: Plenum Press.

Breedin, S. D., Saffran, E. M., & Schwartz, M. F. (1998). Semantic factors in verb retrieval: An effect of complexity. *Brain and Language, 63(1)*, 1-31.

Analogies in the Wild: Generated Analogies as Assertions of Categorization

Leslie J. Atkins (latkins@umd.edu)
Department of Physics & Department of Education and Curriculum, University of Maryland
2226 Benjamin Building, University of Maryland, College Park, MD 20742 USA

Analogies in the wild

The research described here is an attempt to understand analogies that are spontaneously generated by students in science classrooms and in science discussions. Most analogy research and associated models of involve the *interpretation* or *application* and not the *generation* of analogies. Analysis of student discourse presented in this poster shows that analogies generated "in the wild" have features that are neither elicited nor explained by research on analogy interpretation. It is the main thesis of this poster that generated analogies are best understood as assertions of categorization in which the base is a prototypical member of an (often) ad hoc category. Categorization research, perhaps because of its focus on the categories that participants and cultures generate, can account for the following phenomena present in generated analogies: multiple analogies, the choice of base, and the variable representation of the base. Furthermore, the ontology of mind implied by a categorization model is consistent with other findings from cognitive science, linguistics and education research.

Features of generated analogies

Far from what transfer studies would suggest, analogies are frequent in discussions about physical phenomena in science classrooms. In one fifth grade class, when discussing whether or not water will spill from a falling cup, students generate multiple analogies: it is like swinging a toy in a basket, throwing a bucket of water, an astronaut in a space shuttle, or tossing a container of dice. I argue that these analogies serve to assert and negotiate a category, and that this assertion is strengthened by multiple analogies.

The choice of the base in these analogies is consistent with categorization as well. While students may have experiences whose features and structure are similar to the topic at hand, the choice of base is often structurally similar and perceptually dissimilar. If one assumes that analogies are assertions of categorization, then these findings may be accounted for by arguing that the choice of base is the category prototype. In categories, prototypes are the first category members to be elicited; they are used to reason about the category as a whole, and are artifacts of cognitive models. Discourse analysis of analogies finds features of prototypes to be features of the base of generated analogies.

The representation of the base is generally taken for granted in models of analogy. However, there is evidence in generated analogies that the representation of the base is variable and can shift depending on the cognitive model a student applies. When reasoning about the relationship between light and heat, students draw an analogy between light and money. This base changes representation during the discussion from one of wealth (in which \$1.00 can unproblematically change to 4 × \$0.25) to one of currency (in which a dollar never "turns into" four quarters). Consistent with categorization research (Lakoff, 1987), this shift in representation is indicative of the change in cognitive model that is applied.

Ontology of mind

Concepts have long been treated as mental representations that are accessed and acted on by computational processes. This assumption of concepts as stable representations and its implications on the ontology of concepts in the mind has been called into question in several fields, including psycholinguistics, categorization, and education. Despite these concerns, the most widely accepted and used models of analogy ascribe representations to concepts and treat these as fixed— perhaps an artifact of the nature of the analogy studies. A categorization model of analogies, in particular the relationship between categories and cognitive models, addresses these concerns and accounts for analogies using a manifold ontology of mind.

Past research

The claim of analogies as assertions of categorization is not new to the conversation. However, past studies on analogy as categorization have been in vitro studies on the interpretation of analogies. Such scenarios limit the ability to observe variability in analogical reasoning, providing an incomplete picture of the nature of analogy. When viewing analogies that are created by students, the similarities between analogies and categorization become apparent.

Acknowledgments

Thank you to the incredible teachers in whose classes these conversations took place: Ms. Alison Alevy, Mr. Rand Harrington, and the faculty from the Governor's School of North Carolina. Support provided by NSF grant 9986846.

References

Gibbs, R.W. (1992). Categorization and metaphor understanding. *Psychological Review 99*(3).

Glucksberg, S., and Keysar, B., (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97(1).

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What categories reveal about the mind*. Chicago: The University of Chicago Press.

A Bi-Polar Theory of Nominal and Clause Structure

Jerry T. Ball (Jerry.Ball@mesa.afmc.af.mil)
Human Effectiveness Directorate, Air Force Research Laboratory
6030 S. Kent Street, Mesa, AZ 85212 USA

It is argued that the basic structure of nominals and clauses is bi-polar—consisting of a *referential* pole and a *relational* pole. The locus of the referential pole is the *specifier*. The locus of the relational pole is the *head*. For nominals, a determiner functioning as an *object specifier* is the typical referential pole. The determiner functions to ground the nominal in a situation model (Kintsch, 1998). For clauses, an auxiliary verb functioning as a *predicate specifier* is the typical referential pole which grounds the clause in the situation model. For nominals, a noun functioning as the head is the typical relational pole (albeit a non-relation). However, some relations may also head nominals ("kick" in "the kick"). The reason a relation can head a nominal is because the object specifier determines the referential type of the expression, not the head. The object specifier coerces the relation, causing it to be viewed objectively. For clauses, a main verb functioning as the head (or *predicate*) is the typical relational pole. However, most adjectives ("he is *sad*"), prepositions ("he is *out*"), indefinite nominals ("he is *a man*") and some adverbs ("he is *there*") can also function as heads of clauses. Again, the predicate specifier determines the referential type of the clause, not the head. The referential and relational poles may be combined in a single lexical item. For nominals, pronouns, proper nouns, demonstratives and some quantifiers may combine the referential and relational poles. For clauses, tensed verbs combine the two poles. The words which occur between the specifier and the head are typically attracted to one pole or the other. *Modifiers* are usually attracted to the relational pole where they combine with the head. *Referential Modifiers* which encode referential meaning may also be attracted to the referential pole. For clauses, the negative particle tends to combine with the referential pole as is suggested by the clitic forms "isn't", "didn't", and "hasn't" and the requirement for do-insertion ("he does not run" vs. "he runs"). Adverbial modifiers tend to combine with the relational pole. For nominals, ordinal quantifiers tend to combine with the referential pole, whereas cardinal quantifiers tend to combine with the relational pole ("the←first ten→books"). Adverbs, which typically function to modify relations, usually combine with a relational modifier and not the head in nominals ("very" combines with "old" in "very old man").

The bi-polar structure of nominals and clauses does not consider *complements* which are an element of relational meaning. The combination of a relational head with its complements interacts with the encoding of referential meaning in interesting ways. In nominals, the complements of relational heads ("kick" in "the kick") are suppressed by the referential function of the object specifier. Expression of the complements requires introduction of relational modifiers ("of the ball" and "by the man" in "the kicking of the ball by the man"). In tensed clauses, the complements are expressed normally, but in non-finite clauses, expression of the subject *argument* (argument and complement are used synonymously) is suppressed, and in passive clauses, the subject argument is expressed, but corresponds to the object in the active construction, with the subject argument of the active construction being left unexpressed.

The bi-polar theory resolves problems that have plagued uni-polar theories like X-Bar Theory (Chomsky, 1970) and Dependency Grammar (Hudson, 2000). The shift to functional "heads" in X-Bar Theory leads McCawley to lament "…all sorts of things…get represented as heads of things they aren't heads of" (in Cheng and Sybesma, 1998). For example, in "the dog" treating "the" as the head of a DP taking the NP complement "dog"—when "dog" by itself isn't even an NP. Likewise, Hudson's strongly endocentric version of dependency grammar leads him to suggest that "the" is a pronoun that just happens to take a complement.

The bi-polar theory outlined above is called **Double R Theory** (Referential and Relational Theory). Double R Theory is focused on the grammatical encoding and integration of referential and relational meaning within the broader scope of Cognitive Linguistics (Langacker 1987, 1991; Talmy 2000; Lakoff, 1987). Adding a specifier as the locus of referential meaning is an extension of Langacker's (1991) conception of nominals and clauses with the specifier functioning as the locus of Langacker's *grounding predication*. Details of Double R Theory are available at www.DoubleRTheory.com.

Cheng, L, and Sybesma, R. (1998). Interview with James McCawley, University of Chicago. *Glot International* 3:5, May 1998.

Chomsky, N. (1970). "Remarks on Nominalization." In R. Jacobs & P. Rosembaum, eds., *Readings in English Transformational Grammar*. Ginn, Waltham, MA.

Hudson, R. (2000). "Grammar without functional categories." In R. Borsley ed, *The Nature and Function of Syntactic Categories*. New York: Adademic Press.

Kintsch, W. (1998). *Comprehension, a Paradigm for Cognition*. NY: Cambridge University Press

Lakoff, G. (1987). *Women, Fire and Dangerous Things.* Chicago: The University of Chicago Press.

Langacker, R. (1987). Foundations of Cognitive Grammar, Volume 1, Theoretical Prerequisites. Stanford, CA: Stanford University Press.

Langacker, R. (1991). Foundations of Cognitive Grammar, Volume 2, Descriptive Application. Stanford, CA: Stanford University Press.

Talmy, L. (2000). Toward a Cognitive Semantics, Vols I and II. Cambridge, MA: The MIT Press

Age Differences in the Effective Monitoring and Regulating of Source Memory

Sameer Bawa (sammybawa@virginia.edu)
Department of Psychology, University of Virginia
102 Gilmer Hall, PO Box 400400, Charlottesville, VA 22904-4400

Chad S. Dodson (cd8c@virginia.edu)
Department of Psychology, University of Virginia
102 Gilmer Hall, PO Box 400400, Charlottesville, VA 22904-4400

Introduction

A number of studies show age differences in *source memory* (see e.g., #1, 2, 5). However, few studies examine age differences in *monitoring accuracy* for such tasks. The current study uses a paradigm inspired by Koriat and Goldsmith (1996) and Kelley and Sahakyan (2003) to examine age differences in monitoring accuracy for a source memory task when overall accuracy is matched between younger and older adults. This study also examines how differences in monitoring accuracy bear on the ability to regulate overall accuracy when the option to withhold responses is given.

Method and Results

Older adults (OA), younger adults (YA), and older adults who were matched on source accuracy with younger adults (O-M) were presented sentences spoken by one of two speakers: one male and one female. At test, participants were asked to identify the source of each item as male, female, or as new. Following each response, participants were asked to give a confidence judgment and were given the choice of either submitting or withholding their answer. Source accuracy is shown in Table 1. OA were worse than YA in making source judgments, while no differences were found between O-M and YA. However, while YA were able to significantly improve their source accuracy by withholding responses, OA and O-M were not.

Table 1: Source accuracy in forced and free testing conditions

	OA	YA	O-M
FORCED	0.66	0.79	0.76
FREE	0.68	0.85	0.78

Table 2: Monitoring Accuracy

	OA	YA	O-M
Calibration	0.17	0.10	0.15
Gamma	0.03	0.58	0.45

Monitoring accuracy data are shown in Table 2. OA and O-M were significantly worse at monitoring the accuracy of their responses compared to YA. We hypothesized that older adults' diminished monitoring ability was due to *misrecollections*. We tested this hypothesis by completing an analysis in which old items for which high confidence source misattributions occurred were removed. The results are shown in Table 3. Removing misrecollections resulted in no monitoring differences between YA and O-M.

Table 3: Adjusted Monitoring Accuracy

	YA	O-M
Calibration	0.11	0.09
Gamma	0.66	0.67
# Misrecollected	2.61	6.44

Discussion

OA were found to be less accurate in making source memory judgments than YA. Furthermore, even when older participants were matched with YA on source accuracy, they showed an impaired ability to accurately monitor their responses. This impairment was due to a large number of misrecollections, and when these items were removed from the analysis, no differences were found in monitoring accuracy. OA and O-M were also unable to effectively withhold responses to improve their source accuracy scores, compared to YA. This was due to two factors: older participants' impaired monitoring ability and their use of a lax criterion in making submit-withhold decisions.

References

1. Craik, F. I. M., Anderson, N. D., Kerr, S. A., & Li, K. Z. H. (1995). Memory changes in normal ageing. A. Baddeley & B. Wilson (Eds.), *Handbook of memory disorders* (pp. 211-241).
2. Henkel, L. A., Johnson, M. K., De Leonardis, D.M. (1998). Aging and source monitoring: Cognitive processes and neuropsychological correlates. *Journal of Experimental Psychology: General,127(3),* 251-268.
3. Kelley, C. M., Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory & Language, 48(4),* 704-721.
4. Koriat, A., & Goldsmith, M. (1996c). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.
5. Rahhal, T. A., May, C. P., Hasher, L. Truth and character: Sources that older adults can remember. *Psychological Science, 13(2),* 101-105.

Children's Semantic Representations of a Science Term

Rachel Best (r.best@mail.psyc.memphis.edu)
202 Psychology Building, University of Memphis, Memphis, TN 38152

Julie E. Dockrell (j.dockrell@ioe.ac.uk)
Institute of Education, University of London, 20 Bedford Way, London, WC1H OAL

Danielle S. McNamara (d.mcnamara@mail.psyc.memphis.edu)
202 Psychology Building, University of Memphis, Memphis, TN 38152

Background

School-age children acquire hundreds of new words each year, many of which are acquired incidentally, from uses in discourse contexts. Although children are adept word learners, lexical acquisition from oral language is not necessarily inevitable. Children have particular difficulties in the acquisition of terms that involve complex semantic representations, such as science terms (Braisby, Dockrell, & Best, 1999). Learning science terms poses particular challenges for acquisition because they are 'conceptually complex' and can be understood at various levels of abstraction (Meyerson et al, 1991). Our study investigated the kinds of knowledge children acquire about a science term during the process of lexical acquisition.

Understanding the process of lexical acquisition requires a thorough assessment of the nature of children's word-related knowledge (Beck & McKeown, 1991). While previous studies of word learning have focused on what drives children's acquisition of word meanings, less is known about the nature of children's lexical representations, particularly those relating to complex vocabulary. When children acquire a new word, they must identify the sound in the speech stream to encode a phonological representation and then establish a mapping between the word and concept. Ultimately a detailed semantic representation is developed for the new term.

Because of this multifaceted representation, ascertaining the nature of vocabulary knowledge requires multiple measures (Beck & McKeown, 1991). These measures should include both production and comprehension, moving beyond the conventional multiple-choice comprehension task. Indeed, recent research has indicated that tapping into children's knowledge of conceptually difficult concepts may necessitate creative methods, such as drawing-based assessments (Gross & Teubal, 2001).

Assessing knowledge across a range of tasks does not, on its own, provide information about the maturity of children's lexical representations. Although seldom used in lexical acquisition research, comparison with adults' performance allows us to identify knowledge gaps.

The Study

The present study examined children's representations of a science term following a fortuitous exposure to the term. Thirty children's (mean age = 6.7 years) knowledge of the term eclipse was examined before and after a partial solar eclipse that was visible throughout Europe in the summer of 1999. There was considerable media interest at the time, but no general formal educational instruction occurred because children were on summer break.

Our study assessed the nature of understandings that children acquired about the term eclipse and a control term, comet (which was not related to an eclipse) at three points in time (baseline test, two-week post-test, and five-month post-test), using a range of assessment tasks (multiple-choice comprehension, picture-naming, drawing, and eclipse 'making' task). Also, children's knowledge was compared to 15 adult controls during the baseline test and two-week post-test. According to the two-week post-test and five-month post-test, children acquired extensive knowledge about eclipses, but not comets. The majority of children successfully named and drew eclipses and 'made' an eclipse using models of the sun, moon, and earth. Also, children's eclipse knowledge more closely approximated adult-level understandings at the two-week post-test than at the baseline test. Overall, the study offered an important insight into the nature of children's lexical knowledge when words are acquired. The study also identified effective methods for tapping into children's lexical knowledge.

References

Braisby, N. R., Dockrell, J. E., & Best, R. M. (1999). Children's acquisition of Science Terms: Does Fast Mapping Work? *Proceedings for the 8th Conference of the International Association for the Study of Child Language*, 1066-1087. Somerville, MA: Cascadilla Press.

Beck, I. L., & McKeown, M. (1991). Conditions of vocabulary acquisition. In R. Barr, M. L. Karmil, P. Mosenthal, & D. D. Pearson (eds.), *Handbook of Reading Research* (Vol. 1), Longman Publishing Group.

Gross, J., & Teubal, E. (2001). *Microscope use in scientific problem solving by kindergarteners.* Paper presented at the IXth European Conference for Research and Learning, Fribourg-Switzerland, 28th August – 1st September.

Meyerson, M. J., Ford, M. S., Jones, W. P., & Ward, A. W. (1991). Science vocabulary knowledge of third and fifth grade students. *Science Education*, *75*, 419-428.

Cue Onset Asynchrony in Task Switching

Svetlana Evt. Bialkova (s.bialkova@nici.kun.nl)

Ab de Haan (a.dehaan@nici.kun.nl)

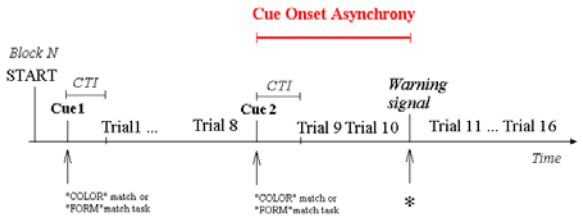
Herbert J. Schriefers (schriefers@nici.kun.nl)

Nijmnegen Institute for Cognition and Information
Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

The Overlapping Cues Paradigm (OCP) and Cue Onset Asynchrony (COA) are introduced as an experimental tool to investigate the dynamics of control mechanisms involved in task switching in a situation with "competition" between two concurrent task goals. We report three experiments focusing on the questions (1) what are the consequences for task performance when two sequentially implemented and overlapping goals are in "conflict"? (2) In such overlapping goal situations, what are the consequences of differences in time pressure (externally paced Cue Target Interval)? (3) How are task switch costs affected by stimulus driven factors (Convergent vs. Divergent trials)?

Experimental paradigm

In the experiments there were two tasks, detection of a form-match or a color-match between a colored geometric figure functioning as a reference and an array of four figures. Which of the two tasks the participant had to perform was indicated by corresponding cues, either the word "Form" or the word "Color" (see figure). Two cues, separated by 8 trials, were presented within each block of 16 trials in two possible combinations: non-conflict (Cue1=Cue2) or conflict (Cue1≠Cue2). Two trials after Cue2, a star was presented as warning signal, which forced a task switch in the conflict condition, but not in the non-conflict condition. The Cue Onset Asynchrony refers to the distance between Cue2 and the Warning-signal.



Manipulated factors

Within the experiments we manipulated the following factors: 1. Task type with levels "Color" match and "Form" match; 2. Cue type with levels non-conflict (Cue1=Cue2) and conflict (Cue1≠Cue2); 3. Stimulus convergency with levels Convergent stimuli (the two different tasks require the same response) and Divergent stimuli (the two tasks require different responses); 4. Cue Target Interval duration (CTI): self-paced (Experiment1); 200ms (Experiment 2) and 900ms (Experiment 3).

Results and discussion

The results show (1) Slower performance for the conflict than for the non-conflict condition on trial 9 (after Cue2) and on trial 11(after warning signal) associated with **top-down control** during COA in order to suppress a conflict if Cue1≠Cue2; (2) For non-conflict condition and self-paced CTI, better performance on trial 9 than on trial 8, and on trial 11 than on trial 10: elimination of restart costs presumably based on **forward facilitation;** (3) On trial 9, faster and more accurate performance for convergent than for divergent condition, because of stimulus driven, **bottom-up control.** And a consequence: On trial 10 better performance if trial 9 was divergent rather than convergent, presumably associated with **backward inhibition;** (4) Different patterns of performance for self-paced and externally paced CTI, associated with different control strategies for the conflict and the non-conflict condition for different CTI.

Issues for further research

These findings reconcile some opposite previous views regarding control mechanisms in task switching and provide a perspective for new investigations on executive control. In the next step we intend to implement two main modifications: a spatiotemporal manipulation of COA, concerning the variation of the relative position at which Cue2 and Warning signal are presented, and second, a spatiotemporal manipulation of Convergency, concerning the variation of the relative position at which Convergent stimuli are presented.

References

Norman, D. A. & Shallice, T. (2000). Attention to action: Willed and automatic control of behavior. In M. Gazzaniga (Eds.), *Cognitive Neuroscience* (pp. 376-390), Massachusetts: Blackwell Publishers Ltd.

Scientific Reasoning in Day-to-Day Research

Janet Bond-Robinson (jrobinso@ku.edu)
Amy Preece Stucky (apreece@ku.edu)
University of Kansas, 1251 Wescoe Hall Drive,
2010 Malott Hall, Lawrence, KS 66045 USA

Introduction

Klahr and Simon (1999) identified four approaches to scientific studies of science emerging in recent decades: (a) Historical accounts of scientific advances, (b) psychological experiments of non-scientists on structured and ill-structured problems, (c) observations of researchers' daily work in science, and (d) computational modeling of scientific discovery processes. Our study fits as (c) observations of daily work in organic synthesis laboratories as others have done in biomechanical engineering (Nersessian, et al, 2002) and molecular biology (Dunbar, 1995). We expect to develop a grounded theory (Glaser & Strauss, 1967) of scientific reasoning within a community of practice (COP).

Theoretical Framework & Methodology

Cognitive apprenticeship is situated learning within a proficient COP through each participant's immersion with frequent opportunities for practice, reflection and discussion while pursuing goals (Lave & Wenger, 1991). When Dunbar studied four different laboratories, all four COPs practicing molecular biology reasoned very similarly, i.e., similar experimental heuristics, mental representations, and problem solving heuristics, and differed only in their own combinations of these features. He noted that researchers interacted with the COP's domain knowledge and fellow researchers to reduce reasoning errors. Logic in scientific reasoning requires *substantial leaps* from the data to infer conclusions (Toulmin, 1977). Toulmin explains that each field (COP) has different things to reason about, different consequences to gauge, and thus, different criteria for justifying inferred conclusions. Thus, apprentices must learn COP-specific standards of justifiable reasoning.

Video data collected included 80 hours of researchers working in the lab, gathering and interpreting data, interacting with mentors, and attending group meetings. Semi-structured interviews, field notes, and laboratory notebook pages supplemented video data. All COP data were analyzed for norms, practices and reasoning.

Results & Conclusions

We asked how scientific reasoning, is instantiated when apprentice researchers pursue their daily work towards Ph.D. "certification" as scientists. This organic COP synthesizes compounds for potential in treatment of diseases, e.g., HIV. The research director determines norms of distributed work from success in funding proposals**;** each project proceeds from a different foundational molecule, however uses similar techniques, equipment, and instruments to perform chemical reactions. Long series of reactions and what makes them work (a mechanical system) lead to a molecule engineered to possess specific and valuable properties.

Problems punctuate researchers' progress. We define a problem as a *difficulty* when the issue shows a basic lack of understanding of the process or inability to get the mechanical system working whereas an *anomaly* is an unexpected and therefore, problematic, piece of evidence. Experience with COP problems inspires integration of explicit declarative knowledge of chemical properties and mechanisms with functional procedural knowledge, whose product is often tacit expertise.

Scientific reasoning is instantiated as "street smarts" developed in a specific research COP where reasoning: (a) Is guided by expectations of the organic synthesis COP's norms and standards (constraints) while researchers do valued COP work. (b) Leads to and develops further apprentices' learning in what to notice, understand, and take advantage of in terms of physical, human, and disciplinary COP resources (affordances). (c) Determines causal interactions of relevant variables in a mechanical system causing difficulties. (d) Is *learning how* to interpret the COP's typical kinds of evidentiary formats in feedback because evidence is often evident only to COP members. (e) Recognizes anomalies in feedback. (f) Deciphers and explains anomalies.

Acknowledgements

National Science Foundation Grant REC-0093319

References

Dunbar, K. (1995). How scientists really reason: In R. J. Sternberg & J. E. Davidson (Eds.), *The Nature of Insight* Cambridge, MA: MIT Press.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory:* Chicago: Aldine.

Klahr, D., & Simon, H. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin, 125*(5), 524-543.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: University Press.

Nersessian, N. J., Newstetter, W. C., Kurz-Milcke, E., & Davies, J. (2002). *A Mixed-method Approach to Studying Distributed Cognition in Evolving Environments.* Proceedings of International Conference on Learning Sciences, Seattle.

Toulmin, S. (l977) *Uses of argument* (updated ed.). Cambridge, UK. Cambridge University Press.

Quantity and quality: A model of how linguistic input drives lexical and cognitive development

Arielle Borovsky (aborovsk@cogsci.ucsd.edu)
Jeff Elman (elman@cogsci.ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Drive La Jolla, CA 92093-0515

In early childhood, words are acquired at a very fast pace. Yet there is also much variation in the overall number of words children learn. While these differences have often been attributed to learning abilities of the child, there is also considerable evidence that this variation can be attributed to the amount of language exposure (Huttenlocher et al, 1991). At the same time, there is some debate about whether grammatically simple or complex child-directed speech (CDS) aids in lexical development. On one hand, some results suggest the number of isolated words in CDS predicts words known (Siskind, 2001), but on the other, there are findings that longer CDS utterances are correlated with improved vocabulary scores (Hoff & Naigles, 2002).

Gopnik & Meltzoff (1997) offer one explanation of how differences in speed of lexical learning might result from cognitive changes that are influenced by linguistic input. Here, a child's ability to learn new words improves with the realization that objects can be classified through grouping new words into known categories. Evidence that the vocabulary spurt coincides with the ability to sort objects into multiple categories lends support to this hypothesis. However, these data are derived only from indirect measures of categorization abilities in children, and it remains unclear what variables might promote the emergence of category knowledge that could facilitate word learning.

In this study, we use simple recurrent networks (SRNs) to explore the relationship between language input (in amount of input, the frequency distribution across categories, and grammatical complexity) on category formation, and the relationship between category formation and rate of acquisition of new words.

Methods

Input: 6 corpora, ranging in size from 20-1000 sentences were developed. 52 nouns were assigned to one of four categories: *animals* (15), *humans* (15), *food* (12) and *objects*(10). 33 verbs were assigned to *eating*, *motion*, *perception*, *action*, *communication*, *change of state* categories. Sentences were either transitive (NVN) or intransitive (NV). One additional corpus of 1000 sentences was developed with additional constructions of NVNN, NVNV and NVNVN.

Simulations: Each of the 7 corpora were used to train 10 SRNs on a next-word prediction task.

Analysis*:* Every 20,000 sweeps, each network was probed to determine (a) rate of new noun learning and (b) internal category structure developed by each network. Category structure was measured by average precision (Keibel & Elman, in prep), which quantifies the similarity of hidden unit values of all nouns in a category. New word learning was tested by training each network on new words in sentences. Learning was measured by the activation value of the word's node when tested in five unseen sentence contexts.

Results and Discussion

Networks that had been exposed to larger vocabulary had a strong effect in facilitating rate of acquisition of new words. Additionally, networks trained with simpler grammatical constructions learned new words faster than ones with complex input. The measure of category formation, AP, correlated well with these results: we found that category coherence values were higher in networks that learned words faster. That is, networks that had been trained with more sentences and simpler syntax showed evidence of better category formation.

These results support the idea that category knowledge plays an important role in lexical acquisition; new words are learned more quickly with better developed category structures. These data also suggest that both the amount and nature of the input (e.g., grammatical complexity; frequency distribution across specific items) play a role in the induction of lexical categories.

Acknowledgements

We are grateful to Holger Keibel for his invaluable advice on measurement of category development.

References

Brent, M. and Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:B33B44

Gopnik, A & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, Mass.: Bradford, MIT Press.

Hoff, E. and Naigles, L. (2002). How children use input to acquire a lexicon? *Child Development, 73,*(2), 418-433.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. Developmental Psychology, 27, 236-248.

Keibel, J. H. & Elman, J. (in prep). From words to categories: Distributional regularities in child-directed speech.

Retention of Contextually Biased Interpretations of Conceptual Combinations

Heather Bortfeld (bortfeld@neo.tamu.edu)
Randy E. Sappington (rsappington@neo.tamu.edu)
Steven M. Smith (stevesmith@neo.tamu.edu)
Rachel M. Hull (mrhull1@juno.com)
Department of Psychology
Texas A&M University
College Station, TX 77843-4325

Novel expressions frequently emerge during everyday language use. For example, a demographic group of middle class mothers who spend time chauffeuring their children to soccer matches and other activities has come to be referred to as "soccer moms." Such terms are introduced to provide labels for meanings that may not have been previously characterized in an efficient manner. These are examples of what Gerrig (1989) has referred to as *sense creation.* The present study concerns the retention of such newly created meanings.

In the series of experiments reported here, participants were asked to read a series of vignettes. All of these vignettes were designed to bias a rare interpretation, as established through out-of-context pre-testing. In the first experiment, participants were randomly assigned to one of three groups. One group was allowed to define the target conceptual combination immediately after reading the vignette, another group read all of the vignettes, then defined the target combinations, and a third group read the vignettes, then completed a 30-minute filler task prior to defining the target combinations. The second experiment was similar to the first in that a series of time delays were introduced to examine the time course of contextual bias. In this experiment, however, type of instruction (explicit reference to a subsequent memory test versus no such mention) was manipulated, in addition to length of delay (in this case, immediate, within one hour, or after two days).

Results from both experiments indicate that, while not permanent, contextual bias has a powerful and relatively long-lasting influence on the way people interpret novel noun-noun combinations. The present study provides evidence that the interpretation of conceptual combinations cannot be completely understood by considering the relationship of the two nouns in the pair or by the relationship between the head and the modifier noun, but must also consider the powerful effects of disambiguating discourse contexts. In our study, it was shown the discourse contexts provided a more powerful influence on interpretations of noun-noun pairs than did the similarity of the nouns considered out of context. Meanings (or interpretations) that were rarely given to a pair of nouns seen out of context were strongly biased by accompanying discourse contexts. Furthermore, these contextualized interpretations persisted over time, with effects observed up to a two day retention interval even though they were only exposed a single time to the brief discourse contexts. Although discourse contexts had a strong effect on interpretations of noun-noun pairs, the number of out-of-context dominant interpretations increased following longer retention intervals. In sum, these findings indicate that interpretations of noun-noun combinations can be strongly affected both by discourse contexts as well as the relationship between the two nouns in each pair.

References

Gerrig, R. J. (1989). The time course of sense creation. *Memory & Cognition, 17 (2),* 194-207.

Finding the Change:
The Role of Working Memory and Spatial Ability in Change Blindness Detection

Gary L. Bradshaw (glb2@ra.msstate.edu),
Courtney Bell (cmb18@msstate.edu),
J. Martin Giesen (jmg1@ra.msstate.edu)
Psychology Department, Mississippi State University
P.O. Box 6161, Mississippi State, MS 39762

Introduction

Change blindness, an inability to spot changes in a visual scene, occurs when normal motion transients are masked by factors such as blank screens, "cuts" from one camera to another, and the like (Simons, 2000). Rensink, O'Regan, & Clark (1997) employed a "flicker" technique to induce change blindness: two versions of an image are presented in alternation, but a blank visual screen is shown between each image. Under these circumstances subjects may take many seconds to notice even large changes in the image, especially if the change occurs in a background element.

Change blindness reveals important limitations in our ability to process visual scene information. Several explanations have been advanced to explain why change blindness occurrs. To date, no single explanation has gained broad acceptance (Simons, 2000).

In the current report, we consider a complementary question: given that change blindness occurs, what factors enable people to overcome it? Our methodology is to exploit the natural individual differences that appear between individuals in their ability to detect changes.

Method

Our methodology was to administer a broad variety of individual differences tests to a large set of subjects then perform correlational and regression analyses to determine the ability factors that predict change blindness detection.

Following the administration of a demographic questionnaire (not discussed further), subjects completed a battery of tests: Integrating details (Alderton, 1989); shape memory (Ekstrom, French, Harman, & Dermen, 1976); identical pictures (Ekstrom et al., 1976); perceptual speed (Guilford & Zimmerman, 1947); a change blindness test; and a measure of operations span (Hambrick & Engle, 2002).

The change blindness task included 20 trials. On each trial two different versions of a photograph were shown repeatedly in sequence, with a blank gray screen appearing between each pairing of the images. Subjects knew a change appeared in each trial and were allowed to view the images until they detected the change. After detection one version of the image reappeared with a set of 5 regions identified, and subjects selected one region to indicate where the change took place.

85 subjects completed the battery of tests during a 1-hour session for class credit. Data from 8 subjects were discarded due to computer errors.

Results

The correlational analysis showed that several ability tests correlated significantly with the *accuracy* of change blindness detection (Table 1), while other factors correlated with the *latency* of change detection .

Table 1: Correlations with Change Blindness Accuracy

	Int. Details	Shape Mem..	Indent Pictures	Percept. Speed	Op. Span
Change Blind. Accuracy	.448**	.406**	.240*	.472**	.493**

* sig at $\alpha < .05$; ** sig at $\alpha < .01$

A stepwise regression analysis revealed that only three factors independently predicted accuracy on the change blindness task: operation span, perceptual speed, and shape memory. The heavy involvement of operation span indicates an important role of working memory in successfully detecting changes in an image. Curiously, measures that are designed to measure spatial ability, such as integrating details, did not show an independent effect. A different set of factors correlated with the latency of change blindness detection, suggesting different mechanisms are involved.

References

Alderton, D.L. (1989). *Development and Evaluation of Integrating Details: A Complex Spatial Problem Solving Test.* NPRDC Technical Report 89-6.

Ekstrom, R.B., French, J.W., Harman, H.H., & Dermen, D. (1976). *Manual for the kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Guilford, J.P., & Zimmerman, W.S. (1947). *Guilford-Zimmerman aptitude survey.* Orange, CA: Sheridan Psychological Service.

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology, 44,* 339-387.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368-373.

Simons, D.J. (2000). Currrent approaches to change blindness. *Visual Cognition, 7,* 1-15.

Thurstone, L.L. (1949). *SRA Primary abilities*. Chicago: Science Research Associates.

Building Mental Models of Multimedia Procedures: Implications for Memory Structure and Content

Tad T. Brunyé & Holly A. Taylor
Department of Psychology, Tufts University
490 Boston Ave., Boston, MA 02155

David N. Rapp
Department of Educational Psychology, University of Minnesota
206A Burton Hall, 178 Pillsbury Dr. SE, Minneapolis, MN 55455

The present experiments examined memory for procedural instructions following three presentation formats. In three experiments participants learned procedures for assembling toys either with instructions presented in text-only, picture-only, or multimedia formats. Testing examined recall, serial order knowledge, and source knowledge. In Experiment 1A, multimedia learning produced faster and more accurate serial order determinations and greater recall, but more source monitoring errors, compared to the other formats. Experiment 1B demonstrated that additional multimedia exposure following initial learning can further influence memory. Experiment 2 examined working memory processes during multimedia learning by attempting to selectively interfere with visuo-spatial and articulatory resources. Contrary to Baddeley's (1992) working memory model, verbally- and spatially-based divided attention tasks failed to selectively interfere with individual slave-system processing, suggesting central executive involvement in the sequential 2-back concurrent tasks. These results provide empirical support for the underlying nature and potential benefits of mental representations following multimedia experiences.

Experiment 1

Participants were presented with a total of 18 5-step Kinder Egg™ toy assembly sequences in picture-only (6), text-only (6), or multimedia (6) format. Half of the participants performed a verbal concurrent task. Subsequent testing included order verification (O.V.), instructions recall (I.R.), and source monitoring tasks. Order of testing for the O.V. and I.R. was reversed in Experiment 1B to test the effects of subsequent multimedia exposure on recall performance.

Results

Congruent with past research (e.g., Mayer & Anderson, 1991), multimedia produced the highest accuracy rates on both the order verification and recall tasks. Interestingly, multimedia also produced the highest source monitoring error rates, with a tendency for participants to inaccurately recall multimedia presentations as picture-only. No evidence for selective verbal interference was found. Additionally, later exposure to pictures following text-only presentations increased recall accuracy.

Experiment 2

Experiment 2 replicated Experiment 1B with the addition of a spatially-based divided attention task.

Results

In line with Experiment 1, dependent measures revealed significant multimedia effects. Verbal and spatial concurrent tasks did not selectively interfere with articulatory or visuo-spatial working memory, respectively.

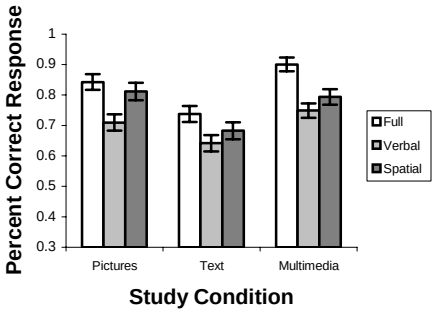


Figure 1: Percent correct response as a function of presentation condition and attention group.

Conclusions

The combination of pictures and text consistently produced higher accuracy rates on tests examining memory structure and content in comparison to pictures or text alone. However, learners are more prone to source monitoring errors after learning with multimedia in comparison to the other two presentation formats. Further research should examine selective working memory interference, with a particular emphasis on multimedia processing.

References

Baddeley, A.D. (1992). Working memory. *Science, 255,* 556-559.

Mayer, R.E. & Anderson, R.B. (1991). Animations Need Narrations: An Experimental Test of a Dual-Coding Hypothesis. *Journal of Educational Psychology, 83,* 484-490.

Behavioral and Electrophysiological Evidence for Configural Processing in Fingerprint Experts

Thomas A. Busey
Indiana University, Bloomington

John R. Vanderkolk
Indiana State Police Laboratory, Fort Wayne, Indiana

Holistic processing has been studied in faces, and may represent one process underlying the development of expertise. In this work we examine how fingerprint examiners show evidence of configural processing when viewing fingerprint fragments. We find evidence for configural processing when fingerprints are presented in noise. Converging evidence was demonstrated in electrophysiological recordings. A particular component of the EEG/ERP known as the N170 is reliably delayed for inverted faces. This component was also delayed for inverted fingerprints, but only for fingerprint experts. The results constrain models of perceptual expertise and provide converging evidence that the delayed N170 component reflects the absence of configural processing.

Iconic Gesture Production in Controlled Referential Domains

Ellen Campana (ecampana@bcs.rochester.edu)
Department of Brain and Cognitive Science, University of Rochester
Meliora Hall, University of Rochester, Rochester, NY 14627 USA

Laura Silverman (lauras@psych.rochester.edu)
Department of Clinical and Social Sciences in Psychology, University of Rochester
Meliora Hall, University of Rochester, Rochester, NY 14627 USA

Introduction

Research on naturally occurring gestures during face-to-face communication has often taken second-seat to the study of spoken communication. This might be due, in part, to widespread beliefs that gesture is too unconstrained to study in a controlled way. Studies in the past have often confounded message selection and gesture/speech form. For example, in one often-used paradigm, participants describe a video they've recently watched to a partner. In these studies the participants 1) select which episodes they describe, 2) decide how to describe the episodes (including word choice and the choice to use gesture), and 3) decide how much information is "enough" to say in order to get the point across to the listener. This confounding of parameters increases noise and adds to the perception that gesture is too unconstrained to study in a controlled way. This study demonstrates that gestures can be predicted precisely when the referential domains are controlled, and it offers a tool for future studies of gesture production.

Method

Participants were 11 individuals from the University of Rochester community. They received either course credit or $7.50 in compensation for their time. Participants were seated at a table, across from a partner with a laptop. The participant was shown a card (figure 1) and instructed to "get their partner to click on the target quadrant by doing anything [they wanted], short of getting out of [their] seat and pointing directly to the target". There were no further restrictions on communication. The partner's screen was identical to the participant's card, except that the quadrant locations did not necessarily correspond and the target quadrant was not indicated in any way. The trial continued until the partner clicked on one of the quadrants.

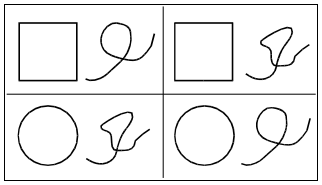


Figure 1: Target quadrant (top left hand corner) might be described as "the square and the squiggle" (with gesture)

Critically, for each trial there were 4 quadrants that the participant could potentially refer to, so the participants had to consider the contents of these quadrants when deciding what to say to their partner (Grice, 1975). Each quadrant contained two "features", which were usually separate objects. In each quadrant one feature would be easy to describe verbally and the other feature would be more difficult to describe verbally. There were a total of four features on each screen – two nameable features that never occurred in the same quadrant, and two less nameable features that never occurred in the same quadrant.

Results

We examined the gestures produced by the participants for each trial, focusing on the participants' first attempts to describe the target quadrant. Participants did produce co-occurring speech and gesture when describing the target quadrant. They were more likely to gesture for the feature that was more difficult to describe (83%) than for the feature that was easy to describe (21%). This pattern held for each individual subject, and for each individual item. We then tabulated the types of gestures that were produced in each case – gestures that were produced while the participant was describing the easy to describe entity were likely (75%) to be beats (McNeil, 1992) or tokens (Lidell 1994), specifically gestures that indicated that the co-occurring speech introduced a new entity to the discourse. They were less likely to be iconic gestures (25%). This pattern was reversed for the gestures that occurred while the participant was describing the feature that was more difficult to describe. In this case gestures were more likely (95%) to be iconic.

Acknowledgments

The authors would like to thank Michael K. Tanenhaus, Loisa Benetto, and Rebecca Webb for theoretical contributions.

References

Grice, H. P. (1975). "Logic and conversation." In Cole, P. and Morgan, *Syntax and semantics: Speech acts.* Volume 3. New York: Academic. 1975, 41–58.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press: Chicago.

Liddel, S., K. (1994) "Tokens and Surrogates." In Ahlgren, I., Bergman, B., Brennan, M. (eds): Perspectives on sign language structure: Papers from the Fifth International Symposium on Sign Language Research. Vol. 1; Held in Salamanca, Spain, 25-30 May 1992. Durham : isla (1994)

Influences of Knowledge on Eye Fixations While Interpreting Weather Maps

Matt Canham (canham@psych.ucsb.edu)
Mary Hegarty (hegarty@psych.ucsb.edu)
Department of Psychology, University of California, Santa Barbara
Santa Barbara, CA 93106-9660 USA

Theories of gaze control in scene perception predict that viewers will fixate on aspects of visual scenes that are either visually salient or informative. Thus far, research has focused mainly on the bottom-up processes that direct a viewer's gaze to salient aspects of a display. However, Semantic knowledge also plays a major role in directing attention toward the most relevant aspects. The current research applies this paradigm to graphical weather map interpretation.

Several studies involving experts from many domains have demonstrated that experts focus on the most relevant information of a display, and because of this they are able to process complex visual information related to their domain much faster than novices.

Although these effects have been well documented, it is still not well understood how these processes develop, or how much knowledge and experience are required before these processing differences begin to appear.

This study examined whether and how the process of focusing on the relevant graphical aspects (while ignoring the irrelevant), changes with a brief period of instruction. The main hypothesis tested was that after instruction in meteorological principles, novice participants would spend more time fixating on relevant aspects of a weather map, and less time fixating on irrelevant aspects.

Methods

Novice participants (N = 16), were shown a series of weather map displays, in which they were asked to determine whether an arrow shown within a target circle, see Figure 1, was showing the correct direction of wind in a target area (true), or the incorrect wind direction (false) .

Participants made judgments on an initial block of 30 trials, were then provided with training on the principles of surface air movement, and finally made judgments on a second block of 30 trials.

Throughout the experimental trials, participant's eye movements were tracked using an SMI EyeLink head mounted eyetracking system.

Eye fixations were analyzed using pre-determined regions of interest, which were assumed to have high or low relevance for successful task completion. These areas included the closest pressure system (relevant) and the temperature scale (irrelevant).

Results and Discussion

Participants showed a significant improvement in performance from before, to after instruction $F(1, 15) = 32.297$, $p < .001$, $\eta 2 = .683$, demonstrating that they learned how to make better judgments about surface wind direction. There was also support for the main hypothesis that after training novice participants would spend more time fixating on the relevant map aspects, and less time fixating on the irrelevant aspects. Participants spent more time fixating on the closest pressure system (highly relevant to the task) after training compared to before $F(1, 15) = 5.162$, $p = .038$, $\eta 2 = .256$, and spent less time fixating on the temperature scale (irrelevant to the task) $F(1, 15) = 5.162$, $p = .038$, $\eta 2 = .256$, after training compared to before. These results, suggest that participants interacted with the weather maps in a qualitatively different way as a result of training.

This study makes two significant contributions to this field of research. First it demonstrates that minimal instruction can influence novices to behave more like experts. Thus, the eye fixation analysis indicated that, participants were able to search the graphic more efficiently (and more like an expert) after a brief amount of instruction. Second this study demonstrates that semantic knowledge influences eye fixations on graphical displays, and not just on pictorial displays or real-world scenes, which have received much more attention in the literature. This research has implications for training of individuals who must interpret complex graphical displays, and for the design of these displays.

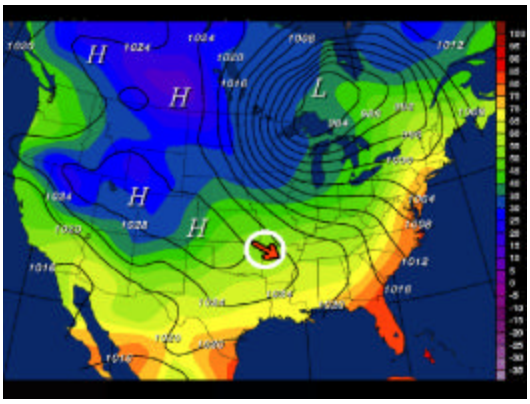


Figure 1. Sample weather map shown to participants. Their task was to judge whether the arrow displayed the correct or incorrect surface wind direction.

Acknowledgement

This research was supported by grants N00014-96-1-0525 and N00014-03-1-0119 from the Office of Naval Research.

Representation of Intentions in Routine Skills

Richard A. Carlson (racarlson@psu.edu)
Daniel N. Cassenti (dnc112@psu.edu)
Lisa M. Stevenson (lms152@psu.edu)
Department of Psychology, The Pennsylvania State University
130 Moore Building, University Park, PA 16802 USA

Intentions can be represented theoretically as active, structured states of working memory. This approach is complementary to standard componential and capacity approaches to understanding executive control. We report the results of several studies examining hypotheses derived from this view of intentions.

Carlson (1997, 2002) described intentions as mental states that instantiate goals and have a schematic structure that specifies desired outcomes, operations for achieving those outcomes, and mental or physical operands. This structure is dynamic, such that instantiating a goal to apply a specific operator evokes a procedural frame to which operands can be assimilated. Instantiation as an intention is one phase of an intention-outcome cycle in which goals are first represented prospectively (Figure 1). In a complex activity, this intention-outcome cycle is embedded in a plan that represents a larger goal structure.

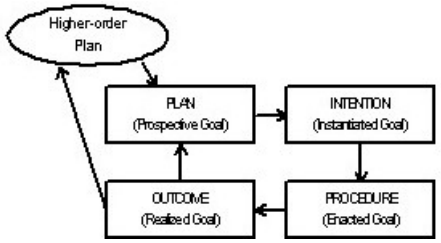


Figure 1: The intention-outcome cycle

Applying this analysis to routine skills with repetitive sequential structures (e.g., counting, running arithmetic) provides a basis for examining several specific hypotheses. The *deictic specification* hypothesis suggests that fluent performance is achieved in part by streamlining intention representations such that their elements are specified deictically rather than semantically, and error monitoring is implicit rather than explicit. The *temporal tuning* hypothesis suggests that instantiating a goal as an intention serves to establish a temporal frame of reference that can be used to coordinate cognitive processing with the perceptual pickup of information.

Deictic Specification

Carlson and Cassenti (in press) examined the deictic specification hypothesis in an event-counting paradigm. Participants counted visual events in a variety of timing conditions. We found support for a model in which events are specified by when they appear rather than by their identity, when temporal regularity makes that possible. This deictic specification allows intention-outcome confusions and promotes implicit error monitoring. Only disruptions to the flow of events (e.g., non-rhythmic trials) seem to trigger error detection, leaving errors in rhythmic trials largely undetected.

Temporal Tuning

Carlson and Stevenson (2002; Stevenson & Carlson, in preparation) examined the temporal tuning hypothesis in a running arithmetic paradigm. We found that preview of at least one upcoming operator seems to be necessary for participants to establish a temporal reference frame. This reference frame is adjusted to the structure of the task, and allows individuals to learn to coordinate self-paced displays with ongoing mental operations. Neither declarative nor procedural knowledge of upcoming operators appear to substitute for preview. Temporal tuning thus depend on environmental support for the specification of operators.

Conclusions

Analyzing intentions as active, structured states of working memory suggests new hypotheses about the control of routine activities. The experimental results reported here support several of these hypotheses, and suggest boundary conditions for them. Other results support hypotheses concerned with information-acquisition strategies and the coordination of information in working memory and in the environment. The present analysis can be related to recent work on the computational modeling of executive control.

References

Carlson, R.A. (1997). *Experienced cognition* . Mahwah, NJ: Erlbaum.

Carlson, R.A. (2002). Conscious intentions in the control of skilled mental activity. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Volume 41), pp. 191-228. San Diego, CA: Academic Press.

Carlson, R.A. & Cassenti, D.N. (accepted for publication). Intentional control of event counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Carlson, R.A. & Stevenson, L.M. (2002). Temporal tuning in the acquisition of cognitive skill. *Psychonomic Bulletin & Review, 9,* 759-765.

Stevenson, L.M. & Carlson, R.A. (in preparation). Constraints on temporal tuning in cognitive skill.

Attention Unites Form and Function in Spatial Language

Laura A. Carlson*, Terry Regier, William Lopez*, and Bryce Corrigan****

*Department of Psychology, 118-D Haggar Hall, University of Notre Dame,, Notre Dame, IN 46556 USA
(lcarlson@nd.edu, wlopez@nd.edu)

**Department of Psychology, University of Chicago, 5848 S. University Avenue, Chicago, IL 60637 USA
(regier@uchicago.edu, b-corrigan@uchicago.edu)

Introduction

Traditionally, the processing of spatial terms has been explained independently of more general cognitive processes, operating upon strictly geometric representations of the objects being spatially related. Challenges to this idea have focused on either process or representation. Research on process has linked spatial language with attention, but has assumed only abstract representations of the objects; research on representation has shown that both geometric and functional information about the objects and their interaction influence spatial language – but the process by which this is accomplished is left largely unspecified. We bring together process and representation, and offer an extension of the Attention Vector-Sum (AVS) model (Regier & Carlson, 2001) in which geometric and functional information is integrated via the process of attention.

Two Assumptions

Carlson-Radvansky, Covey & Lattanzi (1999) observed that spatial terms are defined on the basis of both geometric and functional information. For example, given the instruction: *Place the tube of toothpaste above the toothbrush*, participants were biased to place the toothpaste away from the center toward the bristles of the toothbrush. This functional bias was mediated by the typicality of the relationship between the objects (i.e., a smaller bias with a tube of oil paint). The explanation of the functional bias relies on two critical assumptions: 1) attention can be allocated to a particular functional part of an object (Lin & Murphy, 1997), with a consequent bias to define spatial terms with respect to space around that part; and 2) that the amount of attention allocated to the part is mediated by the typicality of the interaction between the objects.

Empirical support

Empirical support for the first assumption was obtained by manipulating the location of attention within the reference object, and assessing whether there was a bias to define spatial terms around this locus of attention. In Experiment 1, we used an exogenous cueing task to anchor attention at various locations within a rectangle, and then presented a circle as the located object either at the attended location or elsewhere. In Experiment 2, we used a watering can as the reference object, and a plant as the located object; with attention presumably allocated to the spout. In both experiments, response times for verifying that the located object was above/below the reference object were faster when the placement of the located object coincided with attention. Empirical support for the second assumption was obtained by collecting ratings of the functional importance of the parts of the reference objects used by Carlson-Radvansky et. al. (1999) in the context of functionally typical located objects, functionally atypical located objects, or in isolation. Ratings of the functional part were greater in the context of the functionally typical located objects, and were significantly correlated with the linguistic functional bias, suggesting that the typicality of the interaction mediated the strength of the functional information.

Computational support

The Attentional Vector Sum (AVS) model of spatial language involves an attentional beam that is focused on the reference object, and extends outward toward the located object (Regier & Carlson, 2001). There is a vector-sum representation of the direction of the located object relative to the reference object, with vectors anchored at points within the reference object and pointing toward the located object, weighted by the amount of attention paid to the point on the reference object. In order to incorporate functional information about the reference object, the attentional weight in AVS was modified such that functionally important object parts receive greater attention (Lin & Murphy, 1997). With this change, AVS captures the two critical assumptions, and simulations successfully account for the functional bias effect (Carlson-Radvansky et al., 1999) and its dependence on the typicality of the interaction between the objects.

References

Carlson-Radvansky, L. A., Covey, E. S., & Lattanzi, K. M. (1999). "What" effects on "where": Functional influences on spatial relations. *Psychological Science, 10*, 516-521.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception & Performance, 23*, 1153-1169.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General, 130*, 273-298.

Performance *vs.* Learning: Knowing the Right Answers for the Right Reasons

Norma M. Chang (nchang@andrew.cmu.edu)
Kenneth R. Koedinger (koedinger@cmu.edu)
Marsha C. Lovett (lovett@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Introduction

An important dilemma to resolve in instruction is distinguishing between short-term performance and long-term learning in assessing students' progress. Conditions that appear favorable in acquisition are not always as effective at promoting subsequent retention and transfer, due to differences in the processing activities involved in training and at test; in some cases, poor performance in training produced better performance at test (Schmidt & Bjork, 1992). Since global measures of accuracy and speed are insufficient predictors of the effectiveness of training, we have used a modeling approach to draw inferences about the knowledge structures students use to solve problems both at test and in training. We will present the results from one study demonstrating the usefulness of such qualitative measures in predicting learning outcomes from training.

Method

We collected complete sets of data from 47 statistics-naïve undergraduate students in a five-day training study in which they received instruction and guided practice in solving exploratory data analysis problems. The focus of the instruction was to learn when and how to use pie charts, histograms, boxplots, scatterplots, and contingency tables to analyze a set of data. Following each lesson explaining and demonstrating how to use the new representation and method of data analysis, participants worked through a series of practice problems (30 problems in total). One group solved problems in which the problems' surface features were spuriously correlated with their deep structure (*S*-condition), while the other group solved problems whose surface features were varied across all the problem structures (*V*-condition). All participants received problems that were broken down into their individual steps, as well as correct-answer feedback on their solutions. On the final day, participants solved 25 new problems without any scaffolding or feedback. (See Chang, Koedinger, & Lovett, 2003, for a fuller description of a similar procedure.)

Results and Discussion

Consistent with the claim that good performance in training does not guarantee good performance at test, *V*- participants demonstrated a slight disadvantage at training but superior performance at test, in terms of their accuracies and latencies in selecting the appropriate representation type for analyzing the dataset given in the problem. Examining participants' actual answers revealed that *S*-participants' errors were not merely random, but reflected negative transfer from the surface features that had been incorporated into their training.

To assess the extent to which their answers were driven by surface features or by problem structure, we developed a model of participants' knowledge that specified the different possible features they could be using to choose the appropriate statistical display to answer each question. This model was fit to participants' data by adjusting the parameters indicating the degree to which different features were used. The best-fitting models indicated that at test, *S*-condition participants tended to derive their answers from surface features rather than deep structure, whereas *V*-condition participants made greater use of deep structure than surface features.

Analyzing the training data using the same modeling methodology showed that even during the learning phase, *V*-participants demonstrated stronger knowledge of deep structure, whereas *S*-participants exhibited a stronger influence from surface features. The contrast between the apparent performance of *S*- and *V*-participants according to the two different methods of assessment underscores the importance of measuring the target skills that students are intended to learn. Examining the accuracy data alone would suggest that the *V*-participants were performing more poorly than the *S*-participants, with average scores about half a standard deviation lower. However, examining the reasons why participants chose the answers they did reveals more sophisticated understanding in the *V*-condition. Revising our assessments of students' learning to reflect their knowledge representation, rather than relying merely on accuracy scores, may better inform instructional design by distinguishing more clearly between learning and performance.

Acknowledgments

This research was supported by a Department of Education Jacob K. Javits fellowship and NSF grant no. 0087632.

References

Chang, N.M., Koedinger, K.R., & Lovett, M.C. (2003). Learning spurious correlations instead of deeper relations. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207-217.

Visual Cognition in Microanatomy

Julia H. Chariker (julia.chariker@louisville.edu) **John R. Pani (jrpani@louisville.edu)**
Department of Psychological and Brain Sciences, University of Louisville
Louisville, KY 40292

Ronald D. Fell (rfell@louisville.edu)
Department of Biology
University of Lousiville
Louisville, KY 40292

Much of a person's knowledgeable interaction with the world concerns the use of visual information. One useful place to expand our knowledge of visual cognition is in the study of what we will call visual symbol systems. These consist of visual domains that are used as sources of information about additional domains – target domains. Examples include the use of telescopes to study astronomy, aerial photographs, and microscopy.

We report here two studies aimed at understanding the use of microscopy in histology. Histology is the microanatomy of biological tissues. It is a core course in both the biology and medical curricula, and it is essential to the study and practice of pathology.

While the target domain in histology consists of three-dimensional structures, the information domain consists of thin sections through the interiors of these structures that have been treated with a variety of stains. The outcome is that: 1) Histology includes a visual information domain and an anatomical target domain that are both very large and complex. 2) The target domain and the information domain are related by a spatial transformation (taking thin slices) that does not generally preserve structure and appearance. 3) There is a one-to-many mapping from the target domain to the information domain. A single type of structure can have a wide variety of looks in a microscope. 4) There is a many-to-one mapping from the target domain to the information domain. Different structures often look alike.

An interview study was conducted with 5 pre-medical or pre-dental graduates of a college histology course. Participants viewed four different microscope slides. In a first phase, participants thought out loud. In a second phase, a structured interview followed up on statements from the verbal protocol. The view through the microscope and everything that was said was recorded with a digital video camera. The four slides varied in their complexity, their familiarity, and in whether the stain was a common one. The audio recordings were transcribed to written form.

This task was clearly challenging for the students. A correct identification of the whole tissue was made 12 times during the verbal protocol out of the possible 20 identifications. One of the slides was identified by everyone -- one was identified by four of the five people. A slide with an unfamiliar stain was identified by just two people, and a tissue that had not previously been seen in a slide was identified by just one person.

Two formal coding systems were developed to help guide exploration of the cognitive processes involved in the interpretation of histological slides. One system was used to characterize the content of the language used to talk about the slides. The second system was used to characterize the manner in which participants worked toward the goal of tissue identification.

The coding of the language revealed that nearly sixty percent of all propositions used by the participants referred to structures on the slide. Fourteen percent of all propositions were associated with reasoning. Almost sixteen percent of all propositions were expressions of prior knowledge.

A second coding system captured high-level goal-directed cognitive processes. First, a master list was composed of the types of elementary cognitive process used across participants to work toward the goal of identification. Second, for each participant and each slide, progress toward identification was diagrammed using the listed processes.

Across all participants and slides, 70 instances of 13 cognitive processes on the master list were recorded. There were 23 attempts at recognition that did not appear to include hypothesis generation. Thirteen of these consisted of a participant listing structures on the slide and then immediately inferring a whole tissue. There were 39 examples of hypothesis testing across all participants and all slides. Interestingly, disconfirming evidence was used more often than confirming evidence.

This investigation demonstrated that identification of histological structures in a microscope is an extremely challenging task, and individual differences among the students are large. In addition, identification of histological structures in a microscope is remarkable for the degree to which it forces an integration of visual knowledge, general (anatomical) knowledge, and reasoning into a single cognitive system. This includes the use of holistic visual information, analytical knowledge about the diagnostic structures in slides, and general knowledge of anatomy. These forms of representation combine to allow recognition and immediate inference when that is available and extensive processes of reasoning when they are needed.

You Write Better When You Get Feedback From Multiple Peers Than an Expert

Kwangsu Cho (kwangsu@pitt.edu) Christian D. Schunn (schunn@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara St., Pittsburgh, PA 15260 USA

In colleges and universities content classes outside composition classes are providing *near-total neglect* of writing. This unfortunate situation appears to be caused by instructors' workload in generating feedback on student writing. As a result, students do not often practice writing. Therefore, it seems a natural choice to replace instructor or expert reviews with reciprocal peer reviews to remedy the problem. Fortunately, peer reviews seem to allow various advantages beyond the obvious fact that they help instructors spend more time on pedagogically desirable activities by reducing instructors' workload. However, reciprocal peer reviews may be fundamentally limited in that student peers are subject-matter novices in their disciplines and inexperienced in reviewing writing in their disciplines. To improve these issues, Cho and Schunn (2003) developed a web-based reciprocal peer review system called *SWoRD* (refer to the procedure section). The goal of this paper is to show the effectiveness of the SWoRD approaches.

Method

Participants. Participants included 28 students and a domain expert in a 12-week summer class at the University of Pittsburgh, USA. The students had an average of 3.4 college years (*SD* = 1.0). They as writers worked for their class credits. They individually wrote first drafts and final drafts on a topic '*informal science learning*'. They as reviewers also reviewed six peers' first and final drafts. The domain expert was a Ph.D. on the writing topic and had taught similar courses for the past eight years. She was not the instructor of the class but reviewed all of the drafts.

Design. Based on basic writing skill test scores, the students were matched into blocks and then randomly assigned to one of three different conditions: an expert feedback condition (SE), a single peer feedback condition (SP), and a multi-peer feedback condition (MP). The writers in SE received feedback and grades on their drafts only from the expert. Those in SP received them from a single best peer. Those in MP received them from six peers. Also, to get rid of reviewer's status effect, the writers and reviewers were blind to each other. The writers were told that they would not receive writing grades by their instructors, but by their reviewers. All procedures were undergone without marking any identity information.

Procedure. The general procedure of the experiment followed the built-in processes in SWoRD with some modifications for experimental purposes. All of the remaining procedure was managed online by SWoRD. After the writers turned in their first drafts, individual reviewers received a set of six drafts that were randomly selected by SWoRD. They individually generated written comments on six peer drafts and evaluated their qualities on 7-point rating scale (1:Disastrous to 7:Excellent). The same period, the expert reviewed all of the drafts. Then, the writers received selected feedback based on their feedback condition, revised their writing over a week period. Then, writers turned in their final drafts, which were reviewed by the same reviewers. Then, the writers back-reviewed their reviewers' feedback on a five-point rating scale in terms of how helpful it was/would be in revising their first drafts. The results of the back-review were not delivered to the reviewers unlike the SWoRD normal procedure. As a final cycle, the writers received the second round of feedback and back-reviewed the feedback.

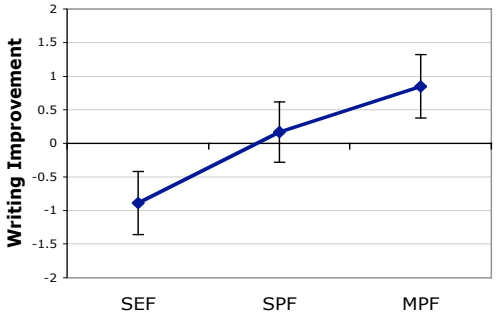


Figure 1: Writing quality improvement

Results

Based on the expert's blind evaluations on all of the papers, a two-way mixed ANOVA on the improvement of writing quality found a significant difference between the feedback conditions F (2, 25) = 3.50, $p < .046$ as in Figure 1. *Tukey* pairwiswe comparison found only the difference between SE and MP signficant, $p. = .015$. Thus, this result supported the SWoRD approaches in that student writers benefited from getting multiple peer feedback and rewriting practice.

Acknowledgments

Funded by grants from the University of Pittsburgh Provost office and the Andrew Mellon Foundation.

References

Cho, K., & Schunn, C. D. (2003). Scaffolded writing and rewriting in the discipline. Available at http://ladybug.lrdc.pitt.edu/sword

Insight Problem Solving as Goal-Derived, Ad-Hoc Categorization

Evangelia G. Chrysikou (lila@temple.edu)
Robert W. Weisberg (weisberg@temple.edu)
Department of Psychology, Temple University
1701 N. 13th St., Weiss Hall 6th Floor, Philadelphia, PA 19122-6085 USA

Problem Solving and Categorization

Research on categorization suggests that, when facing a goal, people construct *goal-derived categories* through *conceptual combination*. These categories may be either well-established or ad hoc, constructed "on-the-spot" from elements from well-established or taxonomic categories. Under pressure or uncertainty people tend to employ taxonomic categories, that reflect world models (i.e., *primary categorizations*) and are denoted by lexemes (e.g., *cups*), as opposed to goal-derived categories, that are unstable, dependent on context, and are denoted by phrases (e.g., *things to pack when going on vacation*; Barsalou, 1983, 1991; Murphy & Ross, 1994).

Insight problem solving is an instance in which the solver, being in a state of uncertainty regarding the solution, is likely to form solution strategies based on primary categorizations. *Insight* is an abrupt and unanticipated shift in the solution path that leads the solver to success. Previous studies regard insight as the result of ordinary cognitive processes (e.g., Perkins, 1981; Weisberg & Alba, 1981) or as indicative of a special way of thinking, characterized by a *representational shift*, a restructuring of the elements of the problem (e.g., Knoblich, Ohlsson, & Raney, 2001). Research, thus far, has neither examined: (i) how solvers interpret insight problems *before* they proceed to a solution, or (ii) how category construction and categorical induction during problem solving are involved in the planning and evaluation of strategies to achieve the intended goal.

This study examined the effects of training to construct goal-derived categories on solving insight problems. We hypothesized that participants who received training in considering secondary, goal-derived categories, in addition to primary, taxonomic categories of items, would exhibit better performance on insight problem solving.

Method

Thirty-six undergraduates were randomly assigned to one of three conditions: (i) Control (n = 12), (ii) Alternative Categories Task [ACT] (n = 12), and (iii) Alternative Categories Task with critical items [ACT-C] (n = 12). The Control condition was administered a word association task and then received six insight problems. The ACT condition was given the Alternative Categories Task and then received the six problems. The ACT-C condition received a version of the Alternative Categories Task, which included six items, each critical for the solution to the problems that followed. The six insight problems were *Charlie, Fake Coin, Candle, Two-Strings, Ten Coins, and Nine-Dots*. Participants in the ACT and ACT-C conditions also received a hint concerning the relevance of the categorization task to the problem-solving task. Participants were tested individually. Sessions were videotaped with subjects' consent. Participants were given specific instructions to *think aloud* during the experimental tasks (Perkins, 1981).

Results and Discussion

A contrast-based ANOVA on solution rates and times revealed that the ACT and ACT-C conditions outperformed the Control condition, with the ACT-C condition exhibiting the highest performance. Results suggest that the construction of goal-derived, ad hoc categories and the ways these categories are used to guide participants' inferences may predict problem solving. The primary aim of this study was to consider problem solving as an instance of goal-derived categorization. Our findings may offer a new perspective on the mechanisms underlying insight. In addition, although much previous research on categorization focuses on the taxonomic organization of isolated items, this study examined categorization in an ecologically valid and dynamic problem-solving task.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*, 211-227.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 1-64). San Diego, CA: Academic Press.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29*, 1000-1009.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27,* 148-193.

Perkins, D. N. (1981). *The mind's best work.* Harvard.

Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of "fixation" in the solution of several insight" problems. *Journal of Experimental Psychology: General*, *110*, 169-192.

The effect of stimulus shape on orientation discrimination of windmill pattern

Ji-Won Chun (occipital@catholic.ac.kr)
Department of Psychology, The Catholic University of Korea
43-1 Yokkok 2-dong Wonmi-gu, Puchon City, Kyonggi-do 420-743, Korea

Jong-Ho Nam (texton@catholic.ac.kr)
Department of Psychology, The Catholic University of Korea
43-1 Yokkok 2-dong Wonmi-gu, Puchon City, Kyonggi-do 420-743, Korea

Instructions

The motion aftereffect (MAE) is a powerful of motion in the visual image caused by prior exposure to motion in the opposite direction (Anstis et al, 1998). The study with use of MAE provides information of motion direction property.

This study was conducted to investigate an effect that shape information of stimulus affects orientation discrimination of windmill pattern with use of MAE. In according to the prior study of visual pathway, shape information of objects is processed to separates from motion information (Lenny, Trevarthen, 1990). Recently, however, there were reports on the interaction between two visual pathways. Kim (2001) proposed that the interaction chromatic and luminance modulation. Nisida (2001) proposed that our brain may process different sensory modalities and attributes in an integrative fashion on the unified spatiotemporal coordinates.

Experiment 1: the effect of stimulus property

In the first experiment, we investigated the effect of stimulus shape in case of adapted stimulus is same with test stimulus, by measuring the orientation discrimination of MAE on windmill pattern. We measured the perceived orientation on test stimulus and the duration time of MAE in different shapes of stimulus.

Stimulus and Method

There were two stimuli in experimental 1. One stimulus was a shape of circle and the other stimulus was a shape of ring.

An adapted stimulus was same with test stimulus. An adapted stimulus was presented on monitor during 15sec. And then test stimulus was presented. Observers experience the MAE. They pressed space bar on keyboard when MAE finished. It was measuring the duration time of MAE. After duration time was measured, observers reported the perceived orientation discrimination. Four observers, native to the purpose of the study and with normal or corrected-to-normal vision were participated.

Results

All observers reported that the orientation of MAE by the shape of circle was different with the orientation of MAE by the shape of ring. The orientation discrimination of MAE by the shape of circle was higher than the shape of ring, which was significant ($F(1,3) = 53.53$, $p < .005$). However MAE duration difference was not significant ($F(1,3) = .422$, $p < .562$). The results suggest that motion information is affected by shape property of object.

Experiment 2: the effect of adapted stimulus

In the second experiment, we investigated only the effect of adapted stimulus in case of an adapted stimulus was different with a test stimulus. We measured the perceived orientation on test stimulus and the duration time of MAE for different shapes of adapted stimulus.

Stimulus and Method

There were three stimuli in experimental 2. An adapted stimulus was circle or a ring and a test stimulus was adding shape. A test stimulus was same each condition. A procedure was same experiment 1.

Results

All observers reported that the orientation discrimination of circle shape MAE was higher than ring shape MAE, which was significant that the difference of orientation discrimination ($F(1,3) = 11.457$, $p < .043$). But it was not difference for duration time within each observer. The results suggest that the shape of object affect the motion information. And it implies to attribute integration.

References

George Mather et al (1998). The Motion Aftereffect, *Trends in Cognitive Sciences, 2, 111-117.*

Jeounghoon Kim (2001). Nonlinear contribution of chromatic information to identifying an object motion. *The Korea Journal of Experimental and Cognitive Psychology,13,173-192.*

Shin'ya Nisida (1999). Influence of motion signals on the perceived position of spatial pattern. *Nature, 379, 610-613.*

Gregory Francis, Hyungjun Kim (2001). Perceived motion in orientational afterimages: direction and speed, *Vision Research, 41, 161-172.*

Language-Specific Grammatical Attention in Second Language Proficiency

Wai Men Noel Chung (wmn_chun@alcor.concordia.ca)
Norman Segalowitz (norman.segalowitz@concordia.ca)
Department of Psychology & Centre for the Study of Learning and Performance, Concordia University
7141 Sherbrooke Street West, Montréal, QC H4B 1R6 Canada

Introduction

The present study investigated whether attention control for grammatical elements plays a role in second language (L2) proficiency. The meanings of grammatical elements (e.g., grammatical morphemes, inflections, and word order patterns) derive from how they relate various message elements to each other. Unlike nouns, adjectives and other content words, the referents of grammatical elements cannot be "experienced directly in our perceptual, sensorimotor, and practical dealings with the world" (Slobin, 1996, p. 91). For instance, in *The teacher was reading a new book*, the elements (*the/a, was, -ing*) refer to definiteness, time, and how the action unfolded. These meanings are not directly available to perception in the same way as are those of *teacher, read, new*, and *book*. Because not all languages use grammatical elements in the same way, L2 learners may experience particular difficulty in their use (Slobin, 1996).

Chung and Segalowitz (2003), using *a non-matching to sample task*, found that L2 proficiency correlated positively with performance in a task of L2 attention control for grammatical elements (pronouns, prepositions, copula forms, and conjunctions). A potential confound, however, was that subjects (Ss) may have used meta-linguistic knowledge about grammatical categories to perform the task, knowledge that may possibly be correlated with L2 proficiency. The present study attempted to replicate that study by removing the metalinguistic confound. In the grammatical condition, only spatial prepositions were used, divided into four subsets (e.g., *above/over/…*; *below/under/…*; *far/beyond/…*; and *close/near/…*). Two control conditions used non-grammatical words unrelated to language structure: concrete words, subsets of "animal" (*cat/dog/…*; *ant/bee/…*; *trout/salmon/…*; *sparrow/eagle/…*) and abstract words, subsets of "qualities" (*happy/glad/…*; *smart/clever/…*; *polite/honest/…*; and *beautiful/pretty/…*).

Method

Bilingual undergraduates (n=32; First language (L1) = English; L2=French) performed the following tasks.

Proficiency was operationalized as efficiency of accessing word meaning in a *lexical categorization task*. In separate L1 and L2 blocks, bilinguals were required to panel press to indicate whether a word referred to a living or non-living object (136 trials in each language). Intra-individual variation in reaction time (based on the coefficient of variation—CV) was the measure of processing efficiency (Segalowitz & Segalowitz, 1993). L2-specific measures were obtained by partialling out L1 from L2 measures.

Attention control was operationalized as efficiency of attention shift judgments in a *non-matching-to-sample task*. In a Non-Match condition, Ss saw a sample word at the bottom of the screen and 4 display words across the top. They had to press one of 4 buttons to indicate the position of a word belonging to a different subcategory than the sample. L1 and L2 versions of the task were created to measure attention control for grammatical (GRAM), concrete (CONC), and abstract stimuli (ABST) (40 experimental trials each). In a Match condition, Ss had to select a stimulus that matched the sample. CVs provided the measure of processing efficiency. Attention control indices were computed by partialling out Match CVs from Non-Match CVs. L2-specific measures were obtained by partialling out L1 from L2 attention indices.

Results

The data were submitted to hierarchical multiple regression with L2-specific proficiency as the dependent measure. In Step 1, measures of L2-specific attention control for abstract (ABST) and concrete (CONC) stimuli were entered. In Step 2, measures of attention control for grammatical stimuli (GRAM) were entered. For the 16 most proficient Ss, in Step 1 (CONC, ABST), R^2= .110 (*n.s.*), and in Step 2 (GRAM), R^2 change = .428 (p = .005). For the 16 least proficient Ss, total R^2 = .020 *(n.s.*).

Discussion

In more highly proficient bilinguals, efficiency of L2 attention control for grammatical elements accounted for 42% of unique variance of L2 proficiency, after controlling for non-grammatical attention. Because all L2 measures had been residualized against L1, the results reflect a language-specific form of attention, not general processing abilities. This replicates Chung and Segalowitz (2003), without the potential metalinguistic confound.

References

Chung, W.M.N., & Segalowitz, N. (2003). A language-specific form of attention that underlies L2-proficiency. Cognitive Science Society Conference. Boston.

Segalowitz, N., & Segalowitz, S. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14*, 369-385.

Slobin, D. (1996). From "thought and language" to "thinking for speaking". In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity*. Cambridge, U.K.: Cambridge University Press.

Managing Multiple Tasks: Reducing the Resumption Time of the Primary Task

Jonathan D. Clifford (cliffo44@msu.edu)
Erik M. Altmann (ema@msu.edu)
Department of Psychology
Michigan State University
East Lansing, MI 48824 USA

The instinctive way in which people use notes to augment their memory suggests one way in which they might manage the disruptive effects of multi-tasking. In this study we investigated whether mental notes and/or physical notes taken before an interruption would reduce time to resume the interrupted task afterwards. Imagine that you are writing email to a colleague when the telephone rings. The *interruption lag* – the time from when the ringing starts until you pick up the phone – is an opportunity to make notes that might help shorten the *resumption lag* – the time from when the phone call ends until you resume the preexisting cognitive state required to compose the email message. A physical note would be some contextual information recorded on a physical medium, whereas a "mental note" would be such contextual information encoded in memory. The current study indicates that resumption lag is indeed reduced by having cues available during the interruption lag (to facilitate mental note taking), but is *increased* by requiring participants to take physical notes.

Participants

Participants were 48 undergraduate psychology students.

Task and Materials

Two tasks were used in the experiment. The *tank task* was a complex resource-allocation task that involved planning simulated missions to defeat targets using tanks (Brock and Trafton, 1999; Trafton, Altmann, Brock, and Mintz, 2003). The *radar task* was a simulated tactical assessment task that involved classifying "tracks" on a radar screen (Ballas, Kieras et al. 1999; Brock, Ballas et al. 2002; Brock, Stroup et al. 2002; as cited in Trafton, Altmann et al. 2003).

Design and Procedure

Participants performed the tank task for three blocks of 20 minutes each. At 12 random points during each block, a visual alert would appear indicating that the secondary task was about to start. The interruption lag following this alert lasted six seconds, during which input to the tank-task interface was frozen (meaning that no actions were possible). After the interruption lag, the tank task display was replaced by the radar task display. The radar task lasted 30 to 45 seconds, after which the tank task display was immediately restored.

There were two between-participants factors: Cue or No Cue, and Record or No Record. The Cue/No Cue variable probed mental note taking, on the assumption that mental notes are easier to make when cues from the interrupted task are perceptually available. In the Cued condition, the tank task display was preserved throughout the interruption lag, whereas in the No Cue condition the tank task display was erased at the start of the interruption lag, so that participants saw a blank screen for six seconds until the start of the radar task. In the Record condition, participants were instructed to use the interruption lag to record data on a prepared form positioned next to the keyboard.

Measures

The resumption lag was computed as the interval from the moment the tank task interface was restored following the interruption to the first mouse click or key press a participant make to resume the primary task.

Results

Each participant's 36 individual resumption lags were extracted from the log files, and the medians entered into an analysis of variance (ANOVA). There was a significant increase in resumption lag for participants in the No Cue condition, $F(1,44)=6.551$, $p=.014$. In the No Record condition the resumption lag was significantly lower than the participants in the Record condition, $F(1,44)=8.332$, $p=.006$. There was no significant interaction between the Cue and Record manipulations, $F(1,44)=1.238$, $p=.272$. The first finding was that visual cues available in the brief transitional period before an interruption speeded resumption of the primary task afterwards. The second finding was that the act of writing contextual information on a form hindered the resumption of the primary task.

Acknowledgements

This work was funded by ONR grant N00014-03-1-0063.

References

Brock, D., Trafton, J.G. (1999). Cognitive representation of common ground in user interfaces. *User Modeling: Proceedings of the Seventh International Conference*. J. Kay. New York, NY, Springer-Wien.

Trafton, J. G., Altmann, E. M., Brock, D. P., Mintz, F. E. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal" *International Journal of Human-Computer Studies 58,* 583-603.

Part-Set Cuing: A Connectionist Approach to Strategy Disruption

Edward T. Cokely (cokely@psy.fsu.edu)
Department of Psychology, Florida State University
Tallahassee, FL 32306

Roy W. Roring (roring@psy.fsu.edu)
Department of Psychology, Florida State University
Tallahassee, FL 32306

Part-Set Cuing

In a part-set cuing paradigm, when part of a previously studied list of words is provided as a memory aid, a reliable and robust impairment of the non-cued list items results. Since its discovery (Slamecka, 1968; as cited in Nickerson, 1984), this paradoxical phenomenon has been characterized as a persisting enigma in memory research (Nickerson, 1984), of both theoretical and practical concern. One leading informal model, the strategy disruption interpretation (Basden & Basden, 1995), suggests that the part-set cuing impairment results because one's retrieval strategy is changed and differs from the original encoding strategy, following the presentation of cues. The strategy disruption account is thoroughly supported by empirical evidence; however, it has been criticized as theoretically vague and poorly defined. In contrast, the other leading account, a formal model using SAM (Raaijmaker & Shiffrin, 1981), while precise, has been criticized as overly defined, theoretically inconsistent, and unable to account for the full range of findings (Roediger & Neely, 1982). In an attempt to more precisely identify and extend the strategy disruption interpretation, we examine and compare both neural network simulations and human experiments in a part-set cuing paradigm.

The Neural Network

The artificial neural network used was a fully-connected, auto-associative, three-layer perceptron, using a backpropagation algorithm with a learning rate set to 0.1. The network used 15 input and 15 output nodes with a bias, 10 hidden, and 10 context units. The context layer used a 1 to 1 association from hidden units to context units and was fully connected from context to hidden units.

Experiment 1: Human Results

A within-participant (N=24) design was used and counterbalanced for list-order, list-cue-order, and randomized part-set cuing. A typical and robust part-set cuing impairment was observed for cued (M=.35) verses non-cued (M=.40) items, $F(1,23) = 4.99$, $p < .05$.

Experiment 2: Simulation Results

A within-simulated-participant (N=24) part-set cuing design was used and counterbalanced for list-order and list-cue-order, with randomized part-set cues. Output vector error served as the dependent variable and was summed and analyzed for cued and non-cued states. A typical and robust part-set cuing impairment was observed, $F(1, 23) = 24.00$, $p < .05$, without evidence of catastrophic interference.

Conclusion & Discussion

The neural network was consistent with the observed human performance, providing a good fit across a number of analyses. The findings suggest that the neural network formalism is consistent with and may serve as an extension of the Basden and Basden strategy disruption account of part-set cuing. That is, following cuing, different study and activation patterns disrupt the subsequent process of recall. This disruption is caused by a change in the availability and accessibility of cued items, altering the retrieval process and thus the retrieval strategy.

Although the experimental evidence is from a small set, results suggest that the neural network can provide an increasingly precise mechanistic account of part-set cuing impairment that is consistent with the leading informal theoretical account. Future simulations should attempt to replicate key findings including part-set cuing facilitation and category-cuing impairment.

References

Basden, D. R., & Basden, B. H. (1995). Some tests of the strategy disruption interpretation of part-list cuing inhibition. *Journal of Experimental Psycholgoy: Learning, Memory, and Cognition, 21,* 1656-1669.

Nickerson, R. S. (1984). Retrieval inhibition from part-set cuing: A persisting enigma in memory research. *Memory and Cognition, 12,* 531-552.

Raaijmakers, G.W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93-134.

Roediger III, H. L., & Neely, J. H. (1982). Retrieval blocks in episodic and semantic memory. *Canadian Journal of Psychology, 36,* 213-242.

Setting of Goals in Museum Websites: General and Specific Influence of Previous Knowledge

Javier Corredor (jac11@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara St., Pittsburgh, PA 15260 USA

Introduction

Both general and domain-specific knowledge influence human performance in tasks as different as scientific discovery (Shunn,1999), social science reasoning (Voss, Tyler & Yengo, 1983) and web-based search (Hsieh-Yee, 1993). These types of knowledge are also crucial for learning in open-ended, ill-structured situations to the extent that domain-specific knowledge and general metacognitive skills are critical to the acquisition of complex knowledge (Lawless & Kulikowich,1996). It is also the case that learners who have either general or domain-specific skills seem to be able to learn in new, open online situations but the absence of either knowledge resource requires that the environment be well-structured (Steinberg, 1989).

Museum websites are open tasks designed to fit the expectations and backgrounds of multiples audiences. Goal-setting for either a web or physical visit is a critical process for the success of both visits and learning in museums because goals determine visitors' paths and learning. Clear goals can make the difference between a superficial drifting visit and a meaningful learning experience. Previous research in real museums shows that visitors develop specific sets of goals to drive their interaction with the exhibitions. Visitors attend to three elements: the content knowledge provided by the exhibitions, the navigational clues provided by the museum environment and the goals they have (Leinhardt, Tittle, & Knutson 2002). Here we explore how visitors to museum websites, with different backgrounds, set goals, make navigational decisions, and attend to the exhibition content in the service of learning.

Method

Eight graduate students were asked to think aloud while they surfed freely through two museum websites of different domains for 20 minutes (anthropology and natural sciences). Half (4) of the participants had robust domain-specific backgrounds in anthropology and half were social science graduate students in other domains. Within each half, two had more than four years of graduate study, while the remainder had less than a year. They were instructed to explore each of two web sites freely and to think aloud while doing so. No specific goals or tasks were given to the visitors. The data were the pages that they visited, the order of the visits, and the comments made while visiting. The online pages were classified as: content pages that presented domain-specific information (e.g. exhibits, articles), and navigational pages that presented information about what could be found in the museum (e.g. link pages). Comments made by participants were coded as to whether or not the visitor was attending or searching for navigational support, were setting goals, or if they were elaborating on content.

Results

For this ill-structured task of visiting a web-based museum, it appears that having high levels of general knowledge (high experience) has the greatest impact on surfing and reasoning behavior. The evidence for this is that high experience visitors *elaborated* more deeply on *conten*t pages than did non-experienced visitors. However, visitors with high content knowledge combined with high experience produced more immediate *clear-cut goals* and used fewer moves to meet them than all other groups. Low-experience higher knowledge visitors seemed to support the establishment of goals but not the actions of elaboration, thus they did not benefit as much from meeting goals; while low knowledge-low experience visitors showed a more random, rapid "click" behavior. We take this to mean that high experience positions individuals learn more through elaborating while high content and experience combined positions individuals learn more about the content domain offered by the museum because they are more effective goal setters in that context.

References

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novices and experienced searchers. *Journal of the american society for information science*, 45, 161-174.

Lawless, K., & Kulikowich, J. (1996). Understanding hypertext navigation through cluster analysis. *Journal Educational Computing Research*, 14, 385-399.

Leinhardt, G., Tittle, C., & Knutson, C. (2002). Talking to oneself: diaries of museum visits. *Learning conversations in museums.* Mahwah, MJ: LEA.

Shunn, C. D. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science.* 13, 337-370.

Steinberg, E. (1989). Cognition and learner control: a literature review, 1977-1988. *Journal of Computer-Based Instruction*. 16, 117-121.

Voss, J. F., Tyler, S.W., & Yengo, L. A. (1983). Individual differences in the solving of social sciences problems. In R. F. Dillon & R. R. Schmeck (Eds.), *Individual differences in cognition* (Vol. 1, pp 205-232). New York: Academic.

A Contingent Response Analysis of Negative Feedback

Andrew Corrigan-Halpern (ahalpe1@uic.edu)
University of Illinois at Chicago, Chicago, IL

Introduction

Negative feedback is information provided by a teacher or other instructional agent given to correct the errors a learner has committed. One might expect that corrective feedback is effective because it helps learners alter their performance. In fact, it has been suggested that negative feedback should be given immediately, so that it can more easily be tied to the cognitive structures responsible for the error.

It is then expected that negative feedback decreases the chance of committing the same error in future situations when the error could occur (i.e. due to the *corrective effect*). While this position makes intuitive sense, it has been largely untested. Studies that compare a negative feedback group to control often use global performance measures and have not considered why feedback is effective. Feedback effects, when obtained, could be due to the corrective effect, or they might be due to other factors.

To consider the question more carefully, we tracked individual responses made in a learning experiment involving multiple trials. We provided feedback for some errors, but allowed others to go uncorrected. It was then possible to consider whether feedback facilitated the correction of errors.

Method

Task

A letter extrapolation task was used, similar to those used by Kotovsky and Simon (1973) and Restle (1970). To make letter extrapolation into a task with multiple opportunities to receive feedback, we presented the given sequence via several short presentations and asked for a response after each one. The subjects viewed the given sequence for 20 seconds, and then attempted to extrapolate it. They were asked to reproduce as much of it as they could, guessing the letters for which they were uncertain. They received feedback on their extrapolation as described below. Then the next trial (20 second study period, plus extrapolation attempt) began. The subjects went through 8 such trials.

The sequence used was [MKNPPNKMNLOQQOLN]. This pattern is composed of the four-letter chunk 'MKNP', which is then reversed to form the chunk 'PNKM'. The Chunks 'NLOQ' and 'QOLN' are translation of the other two chunks.

Feedback and Design

Two negative feedback conditions were used. In the *local* condition, feedback was given for each letter response. In the *global* condition, feedback was given for the 4-letter chunks below. Subjects were told that they would not receive feedback for all errors. They were instructed that a 'none' message would appear below some responses. This message appeared below 25% of all errors, as well as below all correct responses. When the subject received this message, they received no useful information. The study was completed on a Macintosh computer using the Pyscope software. Feedback was given after all responses were made and remained on the screen for 45 seconds.

Results

We have previously reported that subjects in the local condition outperform those in the global condition. (Corrigan-Halpern & Ohlsson, 2002). The current goal is to better understand the source of this effect.

Subjects in the global condition were significantly better at correcting errors after receiving feedback, $F(45,1) = 4.59$, $p < .05$. After receiving negative feedback, subjects in the local condition corrected errors 26% of the time, compared to the global condition where correction occurred 32% of the time.

Subjects in the local group were more likely to correct errors after 'none' messages, $F(37,1) = 68.46$, $p < .001$. After receiving the 'none' feedback, subjects in the local condition corrected errors 92% of the time, compared to 28% of the time for the global group.

Subjects in the local condition were more likely to maintain correct responses, $F(43,1) = 5.02$, $p<.05$. Subject in the local condition reproduced a correct response 71% of the time, compared to 53% for the global condition.

Discussion

Despite the fact that the local feedback condition resulted in superior performance, this effect could not be attributed to the corrective effect. Subjects in the local condition perform well because they are able to correct errors made for responses where feedback was *not* provided. This result suggests that negative feedback achieves its effect indirectly or in a more cumulative fashion.

References

Corrigan-Halpern, A., & Ohlsson, S. (2002). Feedback effects in the acquisition of a hierarchical skill. In W. D. Gray, & C. D Schunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 226-231. Mahwah, NJ: Laurence Erlbaum Associates.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology, 4(3)*, 399-424.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77(6)*, 481-495.

Anaphora and Indefinite Noun-Phrases

Maria Luiza Cunha Lima (marialuizacl@yahoo.com)
and
Edson Françozo (edson@unicamp.br)
LAFAPE, State University of Campinas (UNICAMP)
C.P. 6045 Campinas, SP (CEP 13084-971) BRASIL

Introduction

Indefinite NPs are usually taken to introduce new referents and, thus, are not deemed capable of acting as anaphors. Recent research, however, has attested the occurrence of indefinite anaphoric expressions (Schwarz, 2000), which occur (cf. Cunha Lima, 2004) (i) when the anaphor expresses part-whole relations, including partitive and specifying relations; and (ii) when the sentence or phrase containing the indefinite NP does not enclose a finite VP expressing an event which is different from the one in relation to which the antecedent was introduced.

Consider:

(1) O gato caçou um rato na cozinha. Um rato grande e gordo. *(The cat chased a mouse in the kitchen. A big, fat mouse)*

(2) O gato caçou um rato na cozinha. Um rato saiu pela porta dos fundos. *(The cat chased a mouse in the kitchen. A mouse left by the back door)*

(3) O gato caçou um rato na cozinha. O rato saiu pela porta dos fundos. *(The cat chased a mouse in the kitchen. The mouse left by the back door)*.

In (1) there is no doubt that the second occurrence of *a mouse* refers to the very same mouse mentioned previously. In (2), however, the second occurrence of *a mouse* is not co-referential with the first – it introduces an unmentioned referent in the discourse. Contrast this with (3): now, *the mouse* is old information.

One way to explain the difference between (2) and (3) above is to postulate that the verb following an indefinite NP forces its re-interpretation as not co-referential with the previously focused entity. If this is so, we can predict that processing the verb following an indefinite NP will be costlier than processing the verb following definite NP.

Method

Thirty-six students (native speakers of Brazilian Portuguese) at the State University of Campinas took part in the experiment. Twenty-four pairs of sentences ("texts") were constructed. In a self-paced reading experiment [1], the stimulus texts were chunked as follows: "Meu gato / caçou / um rato / na cozinha. / Um rato (1) / saiu (2) / pela porta (3) / traseira (4)"; and responses were recorded in points (1)-(4).

[1] The experiment was run using the DMDX software, developed at Monash University and at the University of Arizona by K.I.Forster and J.C.Forster

Results and Discussion

Reading times for the verb position (see Table 1) was significantly slower following indefinite than definite NPs. That is, following indefinites, verbs took longer to read.

Table 1: Mean reading times (ms) for tensed sentences

	1	2	3	4
Definite	484,13	387,86	684,08	757,20
Indefinite	519,68	445,67	723,92	795,83
Difference	-35,55	-57,81*	-39,84	-38,63

*F_1=(1,99)7.0379, p=0.009 and F_2=(1,123) 3.9192, p=0.049.

This result is consistent with the prediction that verbs following indefinite NPs are costlier than verbs following definite NPs. The source of such cost may be in the mechanism which bridges referring exp ressions to discourse (Almor, 1999). Recent data (Nadig et al., 2003) indicate that children, in a truth-value judgment test, tend to bridge indefinites to previously mentioned entities; similarly, adults also bridge indefinites to previously focused referents in a forced choice task. It seems that, at least in the case of children, bridging is driven by attention, rather than by type of referring expression. One can hypothesize that, in the present study, referring expressions, either definite or indefinite, were bridged to given/focused referents; when the verb incrementally makes its contribution, the need for re-interpretation becomes apparent – and exerts its tolls.

Acknowledgments

Research funded by CNPq and FAPESP (01/00136-2).

References

Almor, A. (1999) Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review, 106*, 748-765.

Cunha Lima, M.L. (2004). *Anáfora, indefinido e construção textual da referência.*. Doctoral dissertation, Department of LInguistics, State University of Campinas, Brazil.

Nadig, A., Sedivy, J. & Sobel, D. (2003) Bridging definite and indefinite referring expressions to discourse: A developmental view. *Proceedings of the The Sixteenth Annual CUNY Conference on Human Sentence Processing* (p.96). Boston, MA: Northwestern University.

Schwarz, M. (2000) Indirekte Anaphern in Texten: Studien zur domägen bundenen REferenz und Kohärenz in Deustchen. Tübingen: Max Niemeyer Verlag.

The Effects of Self-Explanations of Correct and Incorrect Solutions on Algebra Problem-Solving Performance

Laura A. Curry (lac@ichp.edu)
Department of Psychology, University of Florida
Box 112250, Gainesville, FL 32611

Various strategies such as self-explanation (Chi, 2000), collaborative problem solving (Ellis, Klahr, & Siegler, 1993), scaffolding (Vygotsky, 1978), reciprocal teaching (Brown & Palinscar, 1989), and learning from worked-out examples (Mwangi & Sweller, 1998), have been used successfully to facilitate learning and understanding. Psychologists are particularly interested in the cognitive processes underlying and affected by these methods, the varying effectiveness of each across different domains, and the mechanisms that are associated with the learning that results from the utilization of each. Although these techniques are different in form, each one encourages the student to engage in learning during which knowledge is actively processed, and mental models and schema are constructed and reconstructed. The goal of this study was to extend our knowledge of the mechanisms by which students acquire knowledge and the strategies that could be used to facilitate these processes.

The effects of feedback and self-explanation have been examined under various conditions, and within various domains (Chi, de Leeuw, Chiu, & LaVancher, 1994; Mwangi & Sweller, 1998; Tudge, Winterhoff, & Hogan, 1996). Because both have shown to have advantageous effects under many circumstances, they were used together in this study of algebra problem solving. To extend prior research, both the self-explanation of correct and incorrect solutions was elicited and compared to the condition in which only the correct answer was self-explained. It was hypothesized that students who received feedback and were asked to explain both correct and incorrect solutions would demonstrate the most improvement in solving algebra word problems.

Method

Participants included 80 college students (60 females, mean age = 19.73 years, *SD* = 2.05), including 50 Caucasians, 12 African Americans, 10 Hispanics, 6 Asians, and 2 "Others".

An algebra pretest consisting of 14 multiple-choice compare word problems (4 simple-direct, 5 simple-indirect, and 5 complex) was used to assess algebra problem-solving abilities. Participants then participated in a directed practice session during which they were randomly assigned to one of four experimental conditions (No feedback/"Explain own" (Control), Ambiguous feedback/"Explain own and alternative", Feedback/Explain correct, and Feedback/Explain correct and incorrect"). Students were asked to provide algebraic equations for each of 10 problems, and to explain why they thought these equations were correct (or incorrect). Finally, an algebra post-test, identical in form to the pretest, was administered.

Results & Discussion

Pre- to post-test improvements in performance for students in each of the experimental conditions exceeded those for students in the control condition. Results indicated that feedback and self-explanation conditions positively affected post-test performance. Students who self explained both correct and incorrect solutions outperformed all others, and students in the control group had the smallest increase in performance between pre- and post-tests.

This study extends our knowledge of the strategies that could be used to facilitate the processes by which students learn, and offers insights that could prove valuable to educators in selecting task appropriate instructional techniques.

References

Brown, A. L., & Palinscar, A. S. (1989). Guided, cooperative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Hillsdale, NJ: Erlbaum.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology, Vol. 5: Educational design and cognitive science*. Mahwah, NJ: Erlbaum.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

Ellis, S., Klahr, D., & Siegler, R. S. (1993). Effects of feedback and collaboration on changes in children's use of mathematical rules. *Paper presented at the Society for Research in Child Development.* New Orleans.

Mwangi, W., & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. *Cognition and Instruction*, *16,* 173-199.

Tudge, J. R., Winterhoff, P. A., & Hogan, D. M. (1996). The cognitive consequences of collaborative problem solving with and without feedback. *Child Development, 67,* 2892-2909.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Towards a Model of the Prime-Retention Effect

Eddy J. Davelaar (e.davelaar@bbk.ac.uk)
School of Psychology, Birkbeck, University of London
Malet Street, WC1E 7HX, London, UK
http://www.geocities.com/ejdavelaar

The semantic priming effect has been central in many debates on the structure and dynamics of semantic memory. One view states that concepts in the memory system are interconnected via links of variable strength through which activation spreads. On presentation of a word (prime), its semantic representation will increase in activation, which then spreads via the semantic connections to related concepts. When a second word is presented, whose semantic representation is close to the prime word, shorter lexical decision or naming latencies are observed; the basic semantic priming effect.

This simple semantic network (SSN) model predicts that the more activation is given to the prime, the larger the priming effect. However, it was found that when the prime word had to be retained in short-term memory (STM) while a lexical decision was being made, the priming effect was absent (Davelaar, 2004). This *prime-retention effect* is modulated by semantic distance and the stimulus-onset asynchrony (SOA) between the prime and the target, with weak associates being more affected than strong associates and especially at long SOAs. In order to account for this finding, an extension to Dagenbach and Carr's (1994) Center-Surround hypothesis was proposed. Here, a computational structure is sketched that implements this hypothesis and is easy to incorporate in a recent connectionist models of priming (Huber & O'Reilly, 2003; Plaut & Booth, 2000).

Proposed Model Architecture

The single-layer SSN is changed into a two-layer network and augmented with two types of inhibitory connections that correspond to known anatomical connections in the human cortex (local and global inhibition). The architecture is depicted in Figure 1 and is presented as having a columnar structure, where every column represents a separate concept. Within each column, the input unit (bottom unit) sends activation to the corresponding output unit (top unit) and to the local inhibitory unit (filled circle). The output and inhibitory units are reciprocally connected, which lead to the output unit exhibiting adaptation; with increase in stimulus duration the activation increases from baseline to a maximum and then drops to an intermediate level (cf. Huber & O'Reilly, 2003). Between columns, input units are connected to output and inhibitory units of related concepts, where the connection strength reflects the semantic distance. This inter-columnar organisation implements a semantic on-center/off-surround receptive field. All output units feed into a common global pool of inhibitory units that feeds back to the output units (depicted by the circle-headed arrow). This dynamically enforces a limitation on the maximum number of concepts that can be active simultaneously (cf. Davelaar, et. al., in press). In order to model the prime-retention effect, the output unit of the prime word has a self-recurrent connection (dashed arrow) whose strength affects the probability that the prime remains in STM (conceptualised as sustained activation).

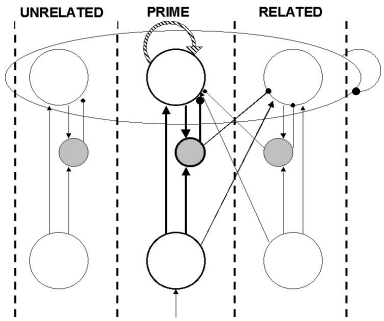


Figure 1. Architecture of a proposed model that captures the prime-retention effect.

Acknowledgments

The support from the Economic and Social Research Council (RES-000-22-0655) is greatly acknowledged.

References

Dagenbach, D., & Carr, T. H. (1994). Inhibitory processes in perceptual recognition: evidence for a center-surround attentional mechanism. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory Processes in Attention, Memory, and Language*. San Diego, CA: Academic Press.

Davelaar, E. J. (2004). Semantic inhibition due to short-term retention of prime words: the prime-retention effect and a controlled center-surround hypothesis. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.*

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (in press). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review.*

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: temporal segregation through synaptic depression. *Cognitive Science, 27*, 403-430.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107*, 786-823.

Reexamining the "Distinctiveness Effect": Poorer Recognition of Distinctive Face Silhouettes

Nicolas Davidenko (ndaviden@psych.stanford.edu)
Michael Ramscar (michael@psych.stanford.edu)
Department of Psychology
450 Serra Mall, Stanford, CA 94305 USA

The distinctiveness effect in face processing

A recognition advantage for distinctive faces has been widely reported (e.g., Valentine, 1991). In such studies, distinctive faces produce more hits and fewer false alarms than typical faces. Although the finding is robust, the mechanism for this advantage has not been carefully explored. The choice of distractors in these studies does not guarantee equivalent target-distractor distances for typical and distinctive faces. In fact, because typical faces lie in a denser, more central region of face space (Valentine, 1991), they will be on the whole more similar to the distractor set than distinctive faces. The location of distractors may thus be sufficient to explain the distinctiveness advantage. In fact, theories of perceptual learning would predict a processing *disadvantage* for distinctive faces that we have less experience with. To control for the effect of unevenly spaced distractors, we constructed a parameterized face space and created equally spaced targets and distractors.

Parameterized face silhouettes

Forty-eight face profiles from the FERET database were reduced to two-toned silhouettes (Figure 1 A and B). The position of 18 key points was recorded for each silhouette from which a 32-dimensional set of principal components (PCs) was computed to fully describe the shape of each silhouette, up to rotation and dilation (Figure 1 C).

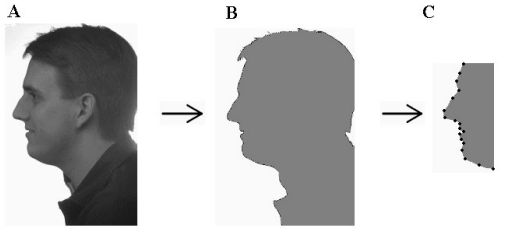


Figure 1. Silhouette parameterization

Experiment 1

From this parameterization, 100 typical and distinctive silhouettes were constructed by setting two of the first 10 PC values to +/- 1 (for typical faces) or +/- 3 (for distinctive faces) standard deviations from the mean. This resulted in distinctive faces being farther from the origin of face space (see Figure 2A), a measure that correlated highly with rated distinctiveness. Distractors were constructed for each face by varying two orthogonal PC values to +/- 1 and +/- 2. In a 3AFC recognition task, 16 Stanford undergraduates observed the randomly presented faces, each followed by a 2-second mask and a choice of three faces (the target and two distractors). Performance was coded as percent identification of the target face. Mean performance was 61% for typical and 56% for distinctive faces, a significant disadvantage for distinctive faces ($p<.05$). To control for the possibility of biased online learning of the typical region of face space, we conducted a second experiment where the size and density of the two regions were matched.

Experiment 2

The design was the same as above except that the set of distinctive faces was defined as a translation in face space from the set of typical faces. Each distinctive face corresponded to a typical face translated by a fixed number of units on a set of eight orthogonal PCs. To control for item effects, the direction of translation was reversed in two between-participant conditions (see Figure 2 B and C).

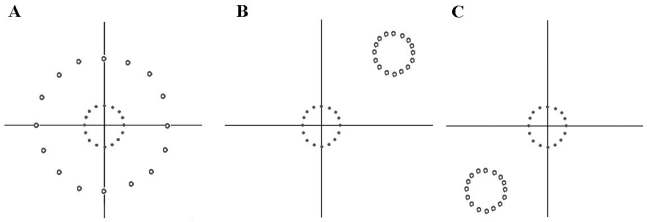


Figure 2. Relative sizes of typical (dots) and distinctive (rings) face regions in Experiments 1 (A) and 2 (B and C)

In conditions 1 (N=16) and 2 (N=14), typical faces were correctly identified more often than distinctive faces (62% vs. 57% and 64% vs. 59% respectively; $p<.05$ in each case).

Discussion

By using parameterized silhouettes, we were able to construct distractors that were equally spaced from their respective targets, across typical and distinctive faces. In two experiments, we found that when controlling for distractor distance, the advantage associated with distinctive faces *reverses*. This "reverse distinctiveness effect" is consistent with the notion that people have less experience with distinctive regions of face space.

References

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43A*, 161-204.

Lexical decision and semantic categorization: age of acquisition effects with students and elderly participants

Simon De Deyne (simon.dedeyne@psy.kuleuven.ac.be)
Department of Psychology, University of Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Gert Storms (gert.storms@psy.kuleuven.ac.be)
Department of Psychology, University of Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Introduction

Age of acquisition (AoA) effects have often been found in a large variety of tasks involving word processing. One of the ongoing discussions remains the confounding of AoA effects with word frequency (e.g. Morrison & Ellis, 1995). In this study we removed possible frequency confounds by comparing AoA and word familiarity differences with young and older participants. A lexical decision experiment was conducted to test if reaction time (RT) differences of both age groups might be explained by differences in AoA of the words.

Previous research (Brysbaert, Van Wijnendaele, & De Deyne, 2000) has shown that AoA effects might also have a semantic origin, apart from proposed origins at the word output or word form access level. To further test this hypothesis we adapted the lexical decision procedure to become a semantic categorization task.

Method

Norms

Subjective ratings of AoA and word familiarity ratings were gathered for 309 Dutch words from students (18 to 23 years old) and older persons (52 to 56 years old). The rated stimuli consisted of early acquired nouns (e.g. apple) and late acquired nouns (e.g. radar), of which some are acquired only recently by the older participants (e.g. modem) and were used for selecting stimuli for the RT experiments. Results showed that not only AoA differed significantly, but also familiarity differed between both age samples.

Lexical Decision

A lexical decision experiment with 108 Dutch words was conducted with young (n = 22, 18 to 23 years old) and older participants (n = 20, 52 to 56 years old). Due to the significant differences between the rated familiarity of words for both groups a factorial design was undesirable and correlational designs were used.

Semantic Categorization

Subjects from both age groups were required to make a categorization between manmade concepts and natural concepts for 160 Dutch words. Ages varied in the young group (n = 21) from 18 to 23 years and from 52 to 56 years in the older group (n = 21).

Results and Conclusions

Results from the lexical decision experiment showed that there was an effect of difference in AoA but not familiarity when predicting RT differences for the young and older participants. The results were complete analogues in the semantic categorization experiment. In both experiments no effects were found for AoA and familiarity when predicting error rates.

The main conclusions of this study are threefold. First, the data show that the normation of words depends on age, both for AoA and familiarity. In this respect it is the first study where age-specific norms are used. Second, our study clearly demonstrates, by only manipulating the age between subjects, that AoA is an important factor in lexical decision. Third, we provide further evidence for an interpretation of the AoA effect as a general effect of learning systems. More specifically, besides the proposed effects on word output or word form access level, AoA plays a significant role in processing the meaning of words.

Acknowledgments

This research was sponsored by research grants G.0266.02 of Belgian National Science Foundation and OT/01/15 of The Leuven University Research Council.

References

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-Acquisition of words is a significant variable in semantic tasks. *Acta Psychologica, 104,* 215-216.

Morrison, C.M., & Ellis, A.W. (1995). Roles of Word Frequency and Age of Acquisition in Word Naming and Lexical Decision. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 21,* 116-113.

Cognitive Style and Integration of Verbal and Visual Information

Kyung Soo Do (ksdo@skku.edu)
Hye-ran Hwang (5823305@hanmail.net)
Department of Psychology, Sungkyunkwan University
Seoul 110-745, KOREA

Cognitive style refers to an individual's preferred and habitual approach to organizing and representing information (Riding & Rayner, 1998). One of the most widely used cognitive styles is the distinction between verbalizers and visualizers. If the verbalizer/visualizer distinction were valid, the presentation order of verbal and visual information should affect learning differently. That is, verbalizers should learn better when verbal information is presented prior to the visual information, whereas visualizers would benefit when visual information is presented prior to the verbal information.

Methods

Ninety-eight Sungkyunkwan University students, selected out of 160 students based on their scores on a cognitive style questionnaire (Kirby, Moore, & Schofield, 1988), participated in the experiment. They studied three Korean historic sites using two different versions of instructional material. Forty-eight students, twenty four verbalizers and twenty-four visualizers, studied using text with video clips presented on computer monitors (text condition). Text and video clips were presented on the monitor screen. Fifty students, twenty-five visualizers and twenty-five verbalizers, studied with narrations and video clips on monitor screens (narration condition). Instructions for each site consisted of seven segments, each of which lasted twelve to fourteen seconds. In the visual first condition, video clips for each segment started three seconds prior to the start of text presentation, and the screen for the video clips remained blank after the end of the segment until the presentation of the text segment ended. In the simultaneous condition, both text (or narration) and the video clips of the segment started simultaneously. In the verbal first condition, text (or narration) started to play three seconds prior to the start of the video clips, and the monitor screen for the text remained blank or silent until the video clips of the segment ended. After students finished studying three sites, they answered twelve retention questions, four for each site, for four minutes, and twelve integration questions, four for each site, for eight minutes.

Results and discussion

The number of correct answers for the retention questions and the integration questions were analyzed. The interaction effect of cognitive style and the presentation order in the retention test was significant in the narration condition ($F(2,96)=6.97$, $p<.01$), and marginal in the text condition ($F(2,92)=2.84$, $p<.06$). The pattern of the interaction effect in the integration test was similar to that of the retention test, but failed to reach statistical significance. As was shown in Fig. 1, visualizers got the most benefit when the visual information was presented first, compared to the simultaneous and verbal first condition ($F(1,24)=3.04$, $p<.09$, and $F(1,24)=3.61$, p,.07, respectively). Whereas verbalizers retained more information when the verbal information was presented first, compared to the visual first condition and simultaneous condition ($F(1,24)=8.35$, $p<.01$, $F(1,24)=6.68$, $p<.02$, respectively). The results suggested that the multimedia instructional material would be better if separately prepared for visualizers and verbalizers.

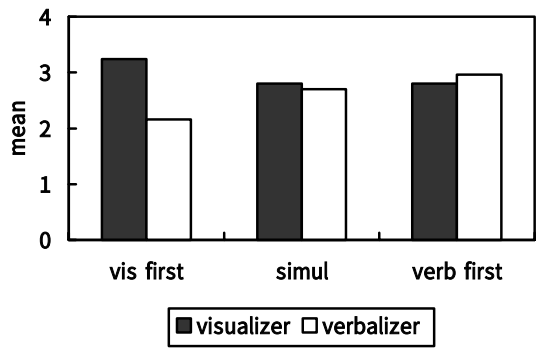


Fig 1. Average mean retention score as a function of cognitive style (visualizer vs verbalizer) and order of visual and verbal information (vis first: visual precedes, simul: visual and verbal simultaneously, verb firs; verbal precedes)

Acknowledgements

This work was supported by the Korea Research Foundation Grant (KRF-2003-041-H00060).

References

Kirby, J. R., Moore, P. J., & Schofield, N. J. (1988). Verbal and visual learning styles. *Contemporary Educational Psychology, 13*, 169-184.

Riding, R., & Rayner, S. (1998). *Cognitive styles and learning strategies: Understanding style differences in learning and behavior.* London: David Fulton.

Segmenting Everyday Actions: an Object Bias?

Rebecca E. Dowell (rebecca.dowell@stanford.edu)

Bridgette A. Martin (martin@psych.stanford.edu)

Barbara Tversky (bt@psych.stanford.edu)
Department of Psychology, Bldg. 420 Jordan Hall
Stanford, CA 94305 USA

Introduction

Recognizing and understanding observed actions is critical to effective social interaction. To do this, observers must segment complex and continuous human behavior into discrete events. Newtson (1973) developed a paradigm for measuring behavior segmentation, in which participants observed films of everyday behavior and used a key to mark off separate events. Zacks, Tversky and Iyer (2001) showed that people segment observed action according to a *hierarchical* structure: larger (coarse) action units were defined by changes in the object being manipulated, whereas smaller (fine) units were defined by changes in actions performed upon the same object. Zacks et al. suggested that this organization reflected a cognitive bias to relate objects to goals and actions to subgoals. Because Zacks et al.'s findings could be due to the organization inherent in their stimuli, the present study examines the possibility of an object bias by asking observers to segment differently organized tasks.

Methods

We filmed two familiar activities: packing a suitcase and washing dishes, according to two different organizations. One version organized larger goals by object changes and subgoals by action changes (*object* films), and the other organized larger goals by action changes and subgoals by object changes (*action* films). Participants viewed both object films or both action films and segmented them according to the Newtson paradigm. In Experiment 1, thirty-two participants viewed the films twice, marking off coarse units on one viewing and fine units on the other. In Experiment 2, sixteen participants segmented the same films into whatever events felt natural.

Results and Discussion

For Experiment 1, linear regression analyses revealed that observers changed their segmentation criteria based on the organization of the films they observed. For *object* films, changes in objects were the best predictor of coarse event boundaries, $F(1,560) = 76.2$, $p < 0.001$, and changes in actions on the same object were the best predictor for fine boundaries, $F(1,560) = 96.2$, $p < 0.001$. For *action* films, this pattern reversed: changes in actions predicted coarse segmentation, $F(1,552) = 125.8$, $p < 0.001$, and changes in objects predicted fine segmentation, $F(1,552) = 27.5$, $p < 0.001$. Although segmentation followed event organization by action or object, there was greater agreement on segment boundaries and greater hierarchical alignment for events organized by objects than events organized by actions (Fig 1). To further test whether objects bias segmentation, participants in Experiment 2 segmented the films into natural units. Object changes were the best predictor of event boundaries for both *object*, ($F(1,560) = 51.068$, $p < 0.001$) and *action* conditions ($F(1,552) = 72.244$, $p < 0.001$). Furthermore, participants in the object condition produced segmentation patterns suggesting that they monitored both coarse and fine levels of action. Participants in the action condition did not show this pattern, suggesting poorer identification of hierarchical structure. Taken together, these data suggest that while observers are adept at uncovering the structure in different task organizations, there is a bias towards object-based segmentation

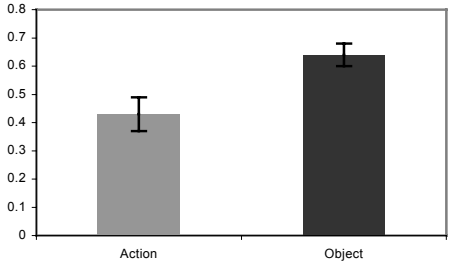


Figure 1: Hierarchical Alignment for *action* and *object* films.

References

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28,* 28-38.

Zacks, J.M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology, 130*, 29-58.

The Creativity of Invented Alien Creatures: The Role of Invariants

Yana Durmysheva (YanaD@brooklyn.cuny.edu)
Aaron Kozbelt (AaronK@brooklyn.cuny.edu)
Department of Psychology, Brooklyn College, CUNY
2900 Bedford Ave., Brooklyn, NY 11210 USA

Purpose

In his study on structured imagination, Ward (1994) found that in general people are not very creative. When asked to imagine alien creatures, participants usually draw beings that resemble common animals and stereotypical science-fiction characters. However, this result does not address why some aliens might be judged as creative, or if creativity can be enhanced, e.g., by certain instructions. Kaplan and Simon (1990) suggested that creative insights may be achieved by actively noticing "invariants," i.e., characteristics that attempted solutions all have in common. Ward (1994) found that most invented creatures had two eyes, four limbs and bilateral symmetry. If these invariants are given to participants in Ward's creature generation task, do they boost creativity? Also, what aspects of the drawing or description of the creature predict its judged creativity?

Method

Creation Task

Sixteen undergraduate students at Brooklyn College participated in the Creation Task, which was similar to that of Ward (1994). The session consisted of six trials (seven minutes each) in which participants were asked to imagine, draw, and describe an alien creature. Participants invented a new creature on each trial. In trials 1-3, participants were given "General" instructions, without any constraints or suggestions. In trials 4-6, they were given "Invariants" instructions and told that their creature should not have 2 eyes, 4 limbs, or bilateral symmetry.

Coding

Coding systems were developed for the drawings and paragraphs. Two raters independently coded each creature. One coding system categorized each drawing in terms of following each of the three invariants. Drawing coding was mostly high (Cohen's kappa = .88 and .79, and .56, for eyes, limbs, and symmetry, respectively). In the paragraph coding system, each proposition in each paragraph was coded into one of 14 categories, e.g., Analogies, Extraordinary abilities, Planet conditions, etc. Inter-rater reliability on coding paragraph categories was also high: r (1342) = .85, p < .0001. All coding was done prior to the Evaluation Task.

Evaluation Task

Ten undergraduates at Brooklyn College participated in the Evaluation Task. Participants were asked to rate the creativity of each of the 96 creatures from the Creation Task, taking into account both the drawing and the paragraph, on a scale from 1 (very low creativity) to 6 (very high creativity). Participants rated the creatures by sorting them into six piles. They could define creativity any way they liked, but they were advised to think about the originality of the drawings and paragraphs.

Results

Performance in the General condition replicated Ward (1994): participants mostly drew creatures with two eyes, four limbs, and bilateral symmetry. In the Invariants condition, almost all participants avoided two eyes and four limbs; fewer followed instructions to avoid bilateral symmetry. In general, there was a strong association between instructions and invariants: for eyes, χ^2 (4 df, N = 96) = 42.5, p < .001, for limbs, χ^2 (4 df, N = 96) = 27.8, p < .001, and for symmetry, χ^2 (4 df, N = 96) = 49.3, p < .001.

Do instructions affect creativity? Evaluations were refined using Rasch analysis (Wright & Masters, 1982), which generates an interval-scale dependent measure of creativity for each creature. A paired t-test, conducted using each participant's average rating for the three creatures in the two conditions, showed no reliable difference, t (15) = 1.06, p = .30. Therefore, the instructions had no discernible effect on the judged creativity of the creatures.

If instructions do not predict creativity, what does? To assess this, a multiple regression was performed using drawing and paragraph coding categories for each creature. The full 17-predictor regression was highly significant, F (17, 78) = 3.66, p < .0001, adjusted-R^2 = .32. Non-significant predictors were dropped, yielding a final 5-predictor model, F (5, 90) = 9.86, p < .0001, adjusted-R^2 = .32. Significant predictors (Betas) were: Activities (.21), Explanations (.26), Extraordinary abilities (.27), Feature description (.19), and Personality characteristics (.32).

Conclusion

Results suggest that constraining participants to avoid common invariants does not enhance the creativity of their productions. However, creativity can be predicted by a several verbal description categories given by participants.

References

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology, 22,* 374-419.

Ward, T. B. (1994). Structured imagination: the role of conceptual structure in exemplar generation. *Cognitive Psychology, 27,* 1-40.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

A Cycle of Learning: Human & Artificial Contextual Vocabulary Acquisition

Karen Ehrlich (ehrlich@cs.fredonia.edu)
Computer Science Department, Fredonia State University, SUNY
234 Fenton Hall, Fredonia, NY 14063, USA

William J. Rapaport (rapaport@cse.buffalo.edu)
Department of Computer Science and Engineering, University at Bufflalo, SUNY
201 Bell Hall, Buffalo NY 14260-2000, USA

Introduction

We are developing a system, Cassie, that defines unknown words from their linguistic context combined with background knowledge [Ehrlich & Rapaport, 1997; Rapaport & Ehrlich, 2000]. This work is of significance for cognitive science, education, computational linguistics and philosophy of mind.

Cassie is built on SNePS, a semantic-network-based knowledge representation and reasoning system developed by Stuart C. Shapiro and the SNePS Research Group [1999], with facilities for parsing and generating English and for belief revision. SNePS has been and is being used for several natural language research projects.

Cassie Learns as Humans Learn

When asked to define of a word, Cassie reports relevant aspects of her experience with that word. Which aspects are chosen depends upon the type and quantity of her exposure to the word, and the contexts in which it occurs, as well as what background knowledge she has.

Our algorithms for selecting information salient to a good definition have been drawn, in large part, from observation of humans with very strong verbal skills. These humans were asked to reason aloud as they read a series of passages containing an unfamiliar word, or a familiar word used in a new sense. It was observed that certain types of information were almost always regarded as salient, while other items would be reported when the preferred data was lacking, but would be dropped from the definition (even if still believed) once the more important information could be observed or inferred.

Though monotonic reasoning is preferred where possible [Ehrlich, 2004], it is at times necessary to withdraw or modify previously held beliefs. Therefore, Cassie has been given an ability to analyze her beliefs, and to select (from among the beliefs supporting a conclusion shown to be erroneous) a belief most likely to be at fault. Cassie then withdraws or modifies the belief according to pre-defined algorithm. This allows her to revise an incorrect definition (as opposed to the more typical case of an incomplete definition).

Cassie's ability to define nouns, verbs, and adjectives has been developed and refined through work on a number of examples, following protocols taken from several readers. Current work includes the further development of a rudimentary discourse analysis for finding the effects of verbs and further development of case-based reasoning.

Humans Learn as Cassie Learns

Meanwhile, the algorithms we have developed and tested on Cassie are being used to formulate an educational curriculum [Rapaport & Kibby, 2002]. It is well-known that the majority of a person's vocabulary is obtained from context, but to date there has been little in the way of instruction in methods of acquiring vocabulary from context. The explicitly formulated algorithms needed to allow Cassie to learn word meanings, however, provide what appear to be a useful set of techniques that can be explicitly taught to secondary school and college students. Investigation is ongoing into how helpful these techniques may be in improving reading comprehension, especially in reading scientific or technical materials that may include a variety of new terms.

References

Ehrlich, K. (2004). Default Reasoning Using Monotonic Logic: Nutter's modest proposal revisited, revised and implemented, *Proceedings of the 15 th Midwest ArtificialIntelligence and Cognitive Science Conference (MAICS2004, Roosevelt University*): 48–54

Ehrlich, K., & Rapaport, W. J. (1997). A Computational Theory of Vocabulary Expansion, *Proceedings of the 19th Annual Conference of the Cognitive Science Society (Stanford University).* (Mahwah, NJ: Lawrence Erlbaum Associates): 205–210.

Rapaport, W. J., & Ehrlich, K. (2000). A Computational Theory of Vocabulary Acquisition, in Iwanska, L. & Shapiro, S. C. (eds.), *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language* (Menlo Park, CA/ Cambridge, MA: AAAI Press/MIT Press).

Rapaport, W. J., & Kibby, M. W. (2002). Contextual Vocabulary Acquisition: A Computational Theory and Educational Curriculum, *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002; Orlando, FL)*

Shapiro, S. C., & the SNePS Implementation Group, (1999). SNePS-2.5 User's Manual, *SNeRG Technical Note* (Buffalo, NY: SUNY Buffalo Department of Computer Science and Engineering).

A Coupled Oscillator Model of Creative Cognition Process for Emergent Systems

Tetsuji Emura (emura@kinjo-u.ac.jp)
College of Human Sciences, Kinjo Gakuin University
Omori 2-1723, Nagoya 463-8521, Japan

Introduction

Psychologists (e.g., Wallas 1926; Guilford 1959; Finke et al. 1992) have clarified the existence of a process controlled by imagination that precedes what is called design, which are deductive logic operations, for the act of creation, in which things that had hitherto not existed are created; a creative act absent this process in mental space is not possible. In other words, a process that could be perceived as what is called insight or idea generation before deductive logic operations substantially controls the creative process. However, research proposing mathematical models for such as the creative cognition process for direct linkage to creativity has seldom been conducted.

When, analyzing the musical work's structures of Brahms, Wagner, for example, a melody is present here, although the melody and harmony are inseparable; there is absolutely no way to first have the melody and then harmonization with it. Moreover, the melody and harmony are allocated to individual instruments for respective sounds and with harmonic progression are changed to be extremely effective as melody; if melody and harmony do not exist simultaneously and if changes in both harmonic progression and timbre in the process of creation do not exist simultaneously in the brain of the composer as a sound image, then creation of a work like this would be close to impossible. That is, harmony, melody, and timbre are in a mode where they are blended into one another and creation must be interpreted to progress with simultaneous processing of these in parallel in the brain of the composer. However, the music theory proposed until now is only static system theory (e.g., Lerdahl & Jackendoff 1983; Forte 1973) and dynamical system theory is not proposed yet.

A Device for Emergent Systems

Two continuous-time autonomous dynamical systems $\mathbf{X_a}$ and $\mathbf{X_b}$ are considered in n-dimensional Euclidean space $\mathbf{R}^n$.

$$\dot{\mathbf{X}}_\mathbf{a} = \mathbf{F}(\mathbf{X}_\mathbf{a}), \quad \dot{\mathbf{X}}_\mathbf{b} = \mathbf{F}(\mathbf{X}_\mathbf{b}) \quad \cdots(1)$$

Here, $\mathbf{F}$ is considered to be the Lorenz system for both with n=3, when considering the $\mathbf{X_a}$ and $\mathbf{X_b}$ in bi-directionally coupled by $0<c_{1,2,3}<1$ are temporal coupling coefficients and $0<d_{1,2,3}<1$ are spatial coupling coefficients, where individual vector components are

$$\mathbf{X}_\mathbf{a} = [x_1, x_2, x_3], \quad \mathbf{X}_\mathbf{b} = [x_4, x_5, x_6] \quad \cdots(2)$$

$$\begin{pmatrix} \dot{x}_{1,4} \\ \dot{x}_{2,5} \\ \dot{x}_{3,6} \end{pmatrix} = \begin{pmatrix} \sigma(x_{2,5} - x_{1,4}) \\ x_{1,4}(r - x_{3,6}) - x_{2,5} \\ x_{1,4}\, x_{2,5} - b\, x_{3,6} \end{pmatrix} + \mathbf{D} \begin{pmatrix} x_4 - x_1 \\ x_5 - x_2 \\ x_6 - x_3 \end{pmatrix} \quad \cdots(3)$$

$$\mathbf{D} = \begin{pmatrix} c_1 & d_2 & d_3 \\ d_1 & c_2 & d_3 \\ d_1 & d_2 & c_3 \end{pmatrix} \quad \text{: excitatory connection}$$

$$\tilde{\mathbf{D}} = \begin{pmatrix} c_1 & d_2 & 1-d_3 \\ 1-d_1 & c_2 & d_3 \\ d_1 & 1-d_2 & c_3 \end{pmatrix} \quad \text{: inhibitory connection}$$

Discussion

The presented Lorenz model having two parameters c and d is a device that has coupled three one-dimensional information codes of $\{x_1$-x_4, x_2-x_5, x_3-$x_6\}$. This device can be used as an emergent device for three channels through control of on-off intermittent chaos as observed in this model with the c and d as parameters. The c and d control on-off intermittent chaos, although they have no direct effect on individual vectors and work as independent parameters without providing internal disturbance. The wandering on the three one-dimensional information coded space in the burst phase with seeking and gathering of valuable information from this, synchronized stabilization on a point in the laminar phase can be modeled as a process that intermittently and irregularly repeats and the phase transition between laminar phase and burst phase simultaneously occur in three dimensions. Figure 1 shows the phase transition from chaos→limit cycles→intermittent chaos → laminar phase with increase of the value of d=d_1=d_2=d_3 in case as constant of c=c_1=c_2=c_3. The figure is plotted in t=0~100000, d=0~1. The d is changing linearly with t, where d=0.00001t.

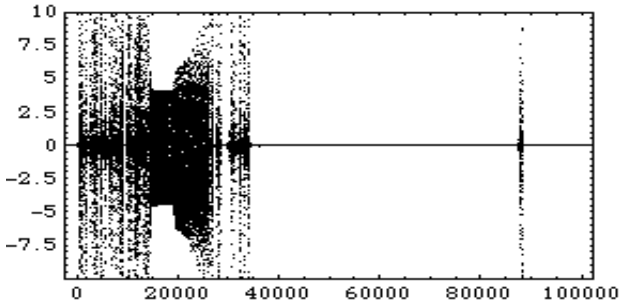


Figure 1. x_1-x_4 versus d (indicated by 10^5), where, σ=10, b=8/3, r=28, c=0.4, inhibitory connection.

A Model of Analogical Retrieval Using Intermediate Features

Mark Alan Finlayson (markaf@mit.edu)
Patrick Henry Winston (phw@mit.edu)
Computer Science and Artificial Intelligence Laboratory, MIT
32 Vassar St., Cambridge, MA 02139 USA

We present a model of analogical recall in people which draws inspiration from recent work in visual classification by Ullman (2002). Our model is intended to unify two bodies of evidence regarding recall in people: on the one hand, we seek to cover a body of evidence that indicates people drawn from a population without regard to task expertise are heavily influenced by surface similarity during retrieval; on the other hand, we also seek to account for the fact that experts are able to achieve analogical recall on a consistent basis. Our model works by breaking the symbolic graph representation of an input situation into sub-graph structures (structures we call *features*), and looking for these features in other situations. By varying the informativeness of the features we use to retrieve situations, we are able to promote or suppress analogical retrieval.

Our model is consistent with previous models of recall (Thagard, Holyoak, Nelson, & Gochfeld, 1990; Forbus, Genter, & Law, 1994) which indicate object similarity, first-order relations, and some small amount of structure dominate recall in normal subjects. These models were primarily intended to account for evidence of the predominance of so-called "mere-appearance" matches in normal recall (Gick & Holyoak, 1980; Rattermann & Gentner, 1987), while still acknowledging some structural effects (Holyoak & Koh, 1987; Wharton et al., 1994)

In contrast to these previous models, however, our model indicates an explanation for certain results in the field of expert problem-solving and retrieval, which has received less attention to date. Evidence drawn from this literature (Chi, Feltovich, & Glaser, 1981; Schoenfeld, 1982; Shneiderman, 1977) indicates that certain sorts of people do consistently achieve analogical recall in particular domains: while these people often fall under the heading "expert," non-experts are also able to attain structural reminding under particular circumstances.

We run our model on a dataset of descriptions of complex political scenarios, and show the predicted switching of preference from mere-appearance to analogical matches when moving from low average feature informativeness to high average feature informativeness. Furthermore our results indicate, as Ullman's did, that features of an intermediate size and complexity provide the most robust recall within analogical category.

Acknowledgments

This work is supported in part by the National Science Foundation under Grant Number 0218861.

References

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Forbus, K. D., Genter, D., & Law, K. (1994). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141-205.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306-355.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, *15*, 332-340.

Rattermann, M. J., & Gentner, D. (1987). Analogy and similarity: Determinants of accessibility and inferential soundness. In *The Ninth Annual Conference of the Cognitive Science Society* (p. 23-35). Lawrence Erlbaum Associates.

Schoenfeld, A. H. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 484-494.

Shneiderman, B. (1977). Measuring computer program quality and comprehension. *International Journal of Man-Machine Studies*, *9*, 465-478.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, *46*, 259-310.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*, 682-687.

Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., & Melz, E. R. (1994). Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, *26*, 64-101.

A Morpheme-Specific Constraint Approach to Vowel Harmony in Korean

Sara Finley (finley@cogsci.jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N Charles St. Baltimore, MD 21218 USA

The problems with previous analyses of Korean vowel harmony are solved when harmony is treated as a consequence of morphological alternation rather than a purely phonological process. This morphological alternation is best accounted for using morpheme-specific correspondence constraints in an Optimality-Theoretic Analysis.

In Korean, vowel harmony occurs in the semantic contrasts of sound symbolic (SS) words. SS words are words whose sound bears some symbolic meaning, and are extremely productive in Korean (Cho, 1994). The majority of SS Korean words alternate between LIGHT and DARK, which have fast and slow connotations, respectively. Alternations between LIGHT and DARK are based on the vowels that can occur in these forms. DARK forms harmonize to contain only DARK vowels: [i, y, ɨ, e, ə, u] (as in [t**e**ŋg**ə**ŋ] 'chopping slowly'). LIGHT forms harmonize to contain only LIGHT vowels: [ɛ, ø, a, o] (as in [t**ɛ**ŋg**a**ŋ] 'chopping quickly').

Phonological analyses of vowel harmony must use one harmonic feature to capture vowel alternations. The problem in Korean is that there is no single harmonic feature. Some alternations involve only a change in height, as in [u] to [o] in [h**u**llə**ŋ**]/[h**o**lla**ŋ**] 'take off clothes'. Some alternations involve only a change in advanced tongue root (ATR), as in [e] to [ɛ] in [t**e**ŋgə**ŋ**]/[t**ɛ**ŋga**ŋ**] 'chopping'. Other alternations involve both changes in aTR and HIGH. Analyses using one harmonic feature ([low] or [ATR] (see Chung, 2000)) cannot completely capture the distinction between DARK and LIGHT without significant restructuring or arbitrary assignment of phonological features. This problem can be solved if this instance of vowel harmony is treated as a morphological process whereby morphemes for DARK and LIGHT bear phonological features that are in correspondence with the output surface form.

The feature associated with DARK is [+ATR] while the features associated with LIGHT are [−ATR] and [−HIGH]. Using both [HIGH] and [ATR] captures the fact that the phonological alternations result in changes in one or both features. The presence of a DARK or LIGHT morpheme triggers morpheme-specific correspondence constraints which restrict the occurrence of vowels in the output. Correspondence between the morpheme and the output is governed by left/right anchoring constraints and output-contiguity. These constraints represent the drive for the morpheme to be in correspondence with all vowels in the output. The interaction of the correspondence constraints with IO-Faithfulness and markedness constraints gives the expected outcome, including an account of high vowels, which do not undergo harmony after the first syllable. The analysis presented also accounts for unexpected behavior, such as the absence of [o] in DARK forms and alternations of [u] and [o] after the first syllable.

Use of correspondence constraints to account for morphologically controlled harmony as opposed to agreement (Bakovic, 2000) is in line with work on featural affixation (Akinlabi, 1994), and is part of a larger project involving the use of morpheme-specific faithfulness constraints to account for morphologically controlled harmony.

Acknowledgments

Thank you to Paul Smolensky, John Alderete and graduate students at UMass and Rutgers for their helpful comments and contributions.

References

Akinlabi, A. (1994). Alignment constraints in ATR harmony. Studies in Linguistic Sciences, 24, 1-18.

Bakovic, E. (2000). Harmony, dissonance and control, Linguistics, Rutgers University.

Cho, M.H. (1994). Vowel harmony in Korean: A grounded phonology approach. PhD Dissertation, Indiana University

When Mats Meow: Phonological Similarity of Labels and Induction in Young Children

Anna V. Fisher (fisher.449@osu.edu)
Department of Psychology & Center for Cognitive Science
Ohio State University
208B Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210 USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210 USA

The ability to make inductive inferences is crucial for humans, and it has long been demonstrated that labels play an important role in induction. However, the mechanism by which labels contribute to induction remained unclear. According to one theoretical position, often referred to as the *naïve theory*, even for young children labels presented as count nouns are special properties: even young children understand that count nouns denote categories, communicating what the things are (Keil, et al, 1998; Gelman & Coley, 1991). According to a recently proposed alternative model SINC (Similarity, Induction and Categorization in Children), children perform induction on the basis of the overall similarity among compared entities, and labels are features contributing to the overall similarity (Sloutsky & Fisher, in press). If labels are features contributing to the overall similarity then not only identical, but phonologically similar labels should contribute to the overall similarity, and therefore to induction. This research was designed to test this prediction of SINC, which, if supported would present challenges to the naïve theory position. Results of two experiments supported the prediction.

Experiment 1: Inductive inference with similar, identical, and different labels

Participants (N = 67, *M* = 4.9 years; *SD* = 0.34) were presented with an induction task in one of the three between subject labeling conditions: identical, similar, and different labels. Children were presented with triads of animal pictures introduced by identical, similar, or different labels, and informed about pseudo-biological properties of two members of each triad. Then children were asked to generalize these properties to the third member of the triad. If labels are category markers as the naïve theory suggests, then identical labels should be fully predictive (thus promoting inferences), while similar labels should be completely non-predictive (thus promoting no inferences). According to the SINC model identical, but also similar labels should promote inductive inferences (i.e., similar labels should be at least partially predictive). Results of Experiment 1 supported predictions of the SINC model: similar labels were found to be partially predictive and likely to promote inductive inferences.

Experiment 2: Label Verification

Results of Experiment 1 could be due to children treating similar novel labels as mispronunciations of identical labels. Experiment 2 was designed to eliminate this potential confound. Participants (N = 29, *M* = 4.8, *SD* = 0.45) were presented with sets of pictures consisting of a Target and four Test stimuli of various degree of similarity to the Target (i.e., identical, very similar, less similar, and dissimilar). On each trial a Target and one of the Test stimuli was labeled with similar labels used in Experiment 1. Children were asked whether the labeled entities had the same name. If children consider similar labels as mispronunciations, then, at least when pictures are identical, they should respond that similar labels were the same. However, the majority of children considered similar labels as different words, and their responses were not affected by picture similarity.

Acknowledgments

This research is supported by a National Science Foundation grant (BCS # 0078945) to Vladimir M. Sloutsky.

References

Gelman, S. A., & Coley, J. (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman, S. & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 146-196). New York: Cambridge University Press.

Keil. F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, *65*, 103-135.

Sloutsky, V. M., & Fisher, A. V. (in press). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General.*

A Critique of the Small Sample Account of Covariation Detection

Wolfgang Gaissmaier (gaissmaier@mpib-berlin.mpg.de)
Lael Schooler (schooler@mpib-berlin.mpg.de)
Jörg Rieskamp (rieskamp@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

The starting point for this project was the finding that people with a low working memory capacity perform better in a covariation detection task (Kareev, Lieberman, & Lev, 1997). In the task people successively encountered envelopes with two different colors and each time had to decide which out of two objects they think they will contain. The explanation for the low capacity advantage is that people with a lower working memory capacity have to rely on smaller samples when they make decisions. This helps them to detect a correlation earlier, because statistically small samples are more likely to indicate a correlation that exceeds the correlation in the population (Kareev, 1995b).

Experiments

We conducted two experiments with an extended version of the original task to test and model the original finding that low capacity people perform better in a covariation task. An implication of the small sample account is that they are also better in detecting a change in the correlational structure of the environment. As in the original experiment, working memory capacity was assessed with a digit span test. The original finding was replicated in the first but not in the second experiment, thus it seems to be a weak and unstable effect. It is worth noting that the probability of replicating a result at the same or a higher level of significance (and in the same direction) is only 50% (Goodman, 1992). Contrary to the predictions by the small sample account there was a high capacity advantage after a change in the first experiment. In the second experiment we did not find any differences between low and high capacity people, neither before nor after a change. Therefore, we focus on the first experiment with regard to modeling.

Modeling

Two different models have been tested, a naïve window model and a reinforcement learning model. Every model was fitted to each individual separately since we wanted to relate capacity to model parameters. The naïve window model that tries to translate the small sample idea directly could not capture the low capacity advantage. But we were able to model it with the reinforcement learning model (Camerer & Ho, 1999) with a decay, a sensitivity and an initial attraction parameter, where we forced the variance in each of the parameters separately by fixing the other two to their means. All three versions were able to capture the low capacity advantage on covariation detection, but only the initial attraction version was related to capacity *and* could predict behavior after a change.

Conclusions

The small sample account is not clearly supported by our data. First, the deduced hypothesis of a low capacity advantage after a change does not hold, we find either no effect or the opposite. Second, the naïve window model and the reinforcement learning model version with the decay parameter which has the strongest connection to memory have to be rejected. Instead, an initial attraction parameter model is successful, indicating a faster learning process of low capacity people in the beginning, but not later on. Still, faster learning can be interpreted as relying on smaller samples. But it is also congruent with the finding of Weir (1964) that children use the simple but most successful payoff maximization strategy (i.e. always choose the more frequent option given a color) earlier in a similar task because they are simply reinforcement driven. Adults, in contrast, develop complex hypothesis and apply complex strategies because they believe that there exists a perfect solution, but they end up worse. As capacity differs between children and adults (Kail, 1984) and plays an important role in hypothesis generation (Dougherty and Hunter, 2003) this could be an explanation for the low capacity advantage.

References

Camerer, C.F., & Ho, T. (1999). Experience-weighted attraction learning in normal-form games. *Econometrica*, *67*, 827-874.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment and individual differences in working memory capacity. *Acta Psychologica*, *113*, 263-282.

Goodman, S.N. (1992). A comment on p-values, replication and evidence, *Statistics in Medicine*, *11*, 875-879.

Kail, R.V. (1984). *The development of memory in children* (2nd ed.). New York: Freeman.

Kareev, Y. (1995b). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*, 263-269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, *126*, 278-287.

Weir, M. (1964). Developmental Changes in Problem-Solving Strategies. *Psychological Review*, *71*, 473-490.

Dynamical Field Theory Predicts a Developmental Reversal in an A-not-B-like Task

Joshua Goldberg (joshgold@cs.indiana.edu)
Department of Computer Science, Indiana University, Bloomington
125 South Woodlawn Ave., Bloomington, IN 47401 USA

The A-not-B Error

Since Jean Piaget's observations of the A-not-B error in his own children, a great deal of scientific effort has been applied to understanding how the error occurs and how children overcome it. The error occurs when a young infant (around 10 months) watches a toy being hidden and then reaches for it. First the child retrieves the toy, hidden a few times at a location A. Then, when the toy is hidden at a new location B and the child must wait a few seconds before retrieving it, she will reliably *perseverate,* reaching to the old location for the toy (often with great frustration.) By about 14 months, infants no longer perseverate in this task.

Dynamical Field Theory And A-not-B

Dynamical Field Theory has been a source of many validated predictions of infant behavior in the A-not-B task (Thelen et al., 2001). These have included manipulations of age, number of practice trials, delays, spacing between target locations, and distinctiveness of hiding boxes. Field theory's lack of dependence on the object concept in conceptualizing the A-not-B error has led to demonstrations that the same patterns of behavior are evident even without a hidden toy, using only light-up buttons for example.

The field model accounts for the dynamics of the A-not-B task by postulating a nonlinearly interactive activation field isometric to the space in front of the infant (Erlhagen & Schöner, 2002). This field, with local excitation and distal inhibition, builds up activation into a peak that indicates where the baby will reach. It is driven by perceptual inputs as well as bias from motor memory of past reaches. Young infants differ from old in that they are less able to maintain a stable reach decision (a peak in the activation field) in the absence of a cue. Therefore, after a short delay, they forget the cue at B and reach to A because of motor memory from practice trials.

A New Task

The A-not-B task does not exhaust the dynamical possibilities that the field model is equipped to handle. Specifically, the A-not-B task does not lead to inhibitory competition between multiple peaks in the activation field. (The competition between a peak in activation and a peak in motor memory is of a different sort with different dynamics.)

A new task we are exploring consists of a cue at A, followed by a delay during which there is a "distractor" cue at B before the baby's turn to reach. We manipulate the duration and timing of the distractor within the delay, as well as the number of training A-trials before this test.

Predictions

Computer simulations of the model allow testing how changes in experimental conditions will affect behavior. More training trials lead to more perseveration to A. A longer distractor more effectively draws the infant to B. A later distractor is more effective because it occurs closer to when the infant may reach.

More striking predictions derive from the differing dynamics of the older versus younger infants. Since young infants cannot maintain a reach decision over a delay, a distractor that is too early is not effective, even if it is long. By the end of the delay, they forget B and are dominated by motor memory at A. Older infants do maintain stable decisions, so a distractor must compete against the cued, stable peak at A. Thus, for older infants, duration is crucial. In the case of a late, short distractor, old infants perseverate more than young—a reversal of the classical A-not-B effect. As illustrated in Figure 1, each condition follows a different dynamical "story," even if the resulting reach is the same.

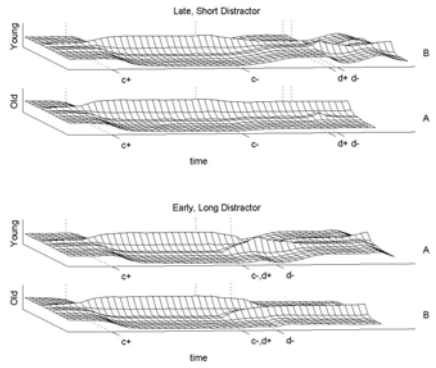


Figure 1: Simulations of four conditions. The first test-trial after 3 A-trials. (c+/c- is cue at A. d+/d- is distractor at B.)

Acknowledgments

This research was funded by National Institute of Health training grant T32 HD 07475 and with National Institute of Child Health & Human Development support. Thanks to Gregor Schöner and Esther Thelen for essential help.

References

Erlhagen, W. & Schöner, G. (2002) Dynamic field theory of movement preparation. *Psychological Review 109*, 545-572.

Thelen, E., Schöner, G., Scheier, C. & Smith, L.B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences 24*, 1-86.

A Simple Model of Encoding and Judgment about Non-Adjacent Dependencies

Pablo Gómez (pgomez1@condor.depaul.edu)
Trisha Hinojosa (tripp@condor.depaul.edu)
Department of Psychology, 2219 N Kenmore
Chicago, IL 60614 USA

Studies with adult and infants have shown that subjects can learn fairly complex probabilistic relationships. Researchers have used statistical learning as a laboratory to explore issues like word segmentation (Saffran, Aslin & Newport, 1996) and the acquisition of grammar (Morgan, Meier & Newport, 1987). Statistical learning has become a scenario for an argument between the two competing views about language acquisition: the view that assumes that humans have some innate ability to acquire grammar (cf., Chomsky, 1965); and the view that claims that statistical learning is based on the same learning mechanisms (e.g., distributed supervised learning) as other domains (see Seidenberg, 1997).

Of particular relevance to our research program are the studies that focus on learning of relationships between non-adjacent speech elements. Newport & Aslin (2004) and R. Gomez (2002) have shown that, only under some special circumstances, participants learn relationships between non-adjacent speech elements (e.g., syllables and words). Here, we present the first version of a model that can account for that data. The model uses a simple encoding process, and a decision mechanism inspired in signal detection theory (Green & Swets, 1966). Our mode supports the notion that very simple mechanisms are enough to explain non-adjacent dependency learning without resorting to special language learning modules.

R. Gomez (2002) has shown that adults and infants can learn non-adjacent regularities when the set size of the intermediate element is large (24 elements), but not when the size set is small (e.g., 2). In a follow up study, she showed that participants could learn non-adjacent dependencies if the intermediate element set size was 1. In these studies, words from an invented language were used in utterances of the form $a_1\ X_{1\ to\ N}\ b_1$, where the dependency was between elements (words) a and b, and the set size of the intermediate element was N.

Newport & Aslin (2004) showed that participants could learn non-adjacent dependencies between letters, but not between syllables. For their experiment with syllables, the stimulus had the form $CV1_1\ CV2_{1\ to\ 4}\ CV3_1$, where the dependency was between the consonant-vowel syllables CV1 and CV3. For the experiment with letters, their experiment had the form $C_1V_{1\ to\ 2}\ C_2V_{3\ to\ 4}\ C_3V_{4\ to\ 5}$, where the dependency was between the consonants C_1, C_2 and C_3.

Description of the Model

The model assumes that subjects use a minimalist approach when they encode the training stimuli. If in their subjective estimation, the adjacent (first order) relationships are informative about the rules to form the artificial language, they will tend not to encode the nonadjacent (second order) relationships.

How informative the first order relationship (say, between the first and second elements in Gomez's studies) is can be determined by a very simple computation:

$$I_{X,b}=p(X_j|a_i)\ (1-p(X_j|a_i)), \qquad (1)$$

where $I_{X,b}$ is a measure of how informative the first order relationship is, and $p(X_j|a_i)$ is the estimated conditional probability of element X_j given element a_i. This measure of informativeness can be thought of as the probability to encode the next order of (non-adjacent) relationship.

The grammaticality judgments are based on familiarity (cf. Signal Detection Theory) at the order of relationship that the learner estimated as informative using Equation 1.

This simple model can account for the u-shaped pattern of accuracy that Gomez found as a function of set size in the intermediate component. In addition, it accounts for the difference between the syllable and letter conditions found by Newport and Aslin; this, because the first order relationship between consonants and vowels had some level of subjective informativeness in the syllable condition.

Acknowledgments

Funded partially by a grant of the URC-DePaul University.

References

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Gomez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431-436.

Green, D. M. and Swets, J. A. Signal detection theory and psychophysics. New York: Wiley.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of intonational and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19*, 498-550.

Newport, E. L., & Aslin, R. N. (2004). Learning at distance. I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48,* 127-162.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1929.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275,* 1599-1603.

From Beetle to Bug: Progression of Error Types in Naming in Alzheimer's Disease

Laura M. Gonnerman[1], Justin M. Aronoff[2], Amit Almor[3], Daniel Kempler[4], & Elaine S. Andersen[2]
[1]Department of Psychology, Lehigh University, Bethlehem, PA 18015
[2]Program in Neuroscience, University of Southern California, Los Angeles, CA 90089-2520
[3]Department of Psychology, University of South Carolina, Columbia, SC, 29208
[4]Comunication Sciences and Disorders, Emerson College, Boston, MA 02116-4624

The distributed feature approach to semantic memory organization has been supported by data from patients with Alzheimer's disease (AD) (e.g., Gonnerman et al., 1997). This account makes specific predictions about the types of errors one would expect in AD as semantic memory deteriorates, with initially more contrast coordinate errors, followed by superordinates, and finally an increase in unrelated responses. We investigate these predictions using a picture naming task, with both natural kinds and artifacts.

Method

Participants

The young normal (YN) group included 25 USC undergraduates, the old normal (ON) group 24 healthy elderly, and the Alzheimer's (AD) group 15 individuals diagnosed with AD, matched with the ON group for age.

Materials and Procedure

Participants named 144 color pictures, with 12 items each from six natural kinds and six artifacts categories, controlled for familiarity, imageability, frequency, and typicality.

Results & Discussion

The YN group correctly named 86% of the pictures, ON 85%, and AD 62%, indicating a significant impairment in naming for the AD group, (t (15) = -4.15, p < .0009), but no significant difference between YN and ON controls.

To examine the types of errors AD patients made as their naming impairment progressed, errors were coded into three categories: 1) *contrast coordinate*, giving the name of another category member (e.g., calling a *zebra* 'horse'); 2) *superordinate*, giving the category label rather than the object name (e.g., 'bug' for *beetle*); and 3) *unrelated*, where the response was not from the same category (e.g., 'flute' for *cucumber*). No responses, 'I don't know', and machine errors were not included in the analysis.

To determine if the prevalence of a given error type was affected by the degree of damage, ratios of each error type over the total number of errors were calculated. Overall, there were initially significantly more contrast coordinate errors than superordinates (t(327)=-4.7, p <.00001), followed by unrelated responses (t(190)=-3.5, p <.001). This is consistent with the progression of errors in studies of patients with semantic dementia (Hodges et al., 1995).

We were most interested in the progression of errors within natural kind versus artifact categories (see Figure 1 below). The pattern of change varied by domain. As expected, there were more contrast coordinate errors in both natural kinds and artifacts early on, declining with increasing damage. Interestingly, while superordinate errors increased for natural kinds, they decreased for artifacts. The distributed feature approach provides a natural account of this pattern. As damage increases, the core features of natural kinds concepts are still available because they have more intercorrelations. The activation of these core features permits activation of the superordinate name, whereas the lack of similar correlations in artifact categories leads to a steady decrease in superordinate responses for artifacts. Finally, there is a greater increase in unrelated responses in artifacts compared to natural kinds in later damage stages.

Acknowledgments

This research was supported by NIA grant R01 AG-11774-04 and by NIH training grant 5T32MH20003-05.

References

Gonnerman, L.M., Andersen, E.S., Devlin, J.T., Kempler, D. & Seidenberg, M.S. (1997) Double dissociation of semantic categories in Alzheimer's disease. *Brain and Language*, 57, 254-279.

Hodges, J.R, Graham, N. & Patterson, K. (1995). Charting the progression in semantic dementia--implications for the organization of semantic memory. *Memory, 3*, 463-495.

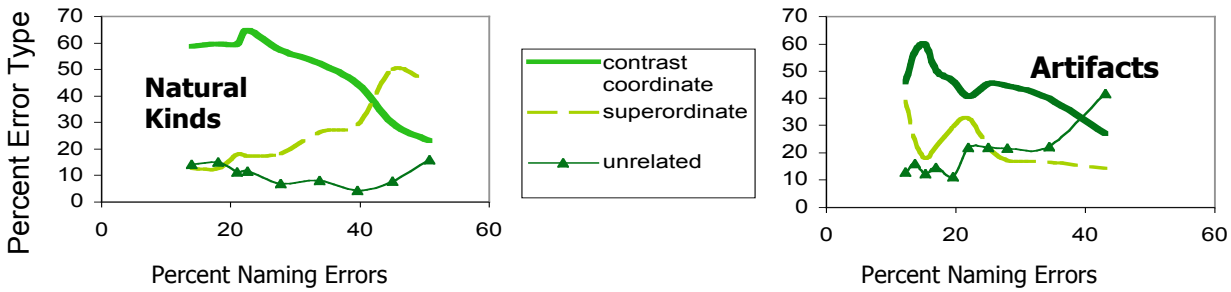


Figure 1. Percentage of error types as naming errors increase for natural kind (left) and artifact (right) concepts.

Probing the Paradox of the Active User: Asymmetrical Transfer May Produce Stable, Suboptimal Performance

Wayne D. Gray, V. Daniel Veksler, & Wai-Tat Fu

Cognitive Science Department
Rensselaer Polytechnic Institute
[grayw, vekslv] @rpi.edu

Psychology Department
Carnegie Mellon University
wfu@cmu.edu

The "paradox of the active user" (Carroll & Rosson, 1987) is the persistent use of inefficient procedures in interactive environments by experienced or even expert users when demonstrably more efficient procedures exist. In this study we examine the procedures that people adopt in response to minor changes in interface design and how these procedures adapt, or do not adapt, when the design changes.

Prior Knowledge versus Perceptual-motor Effort

For this work, subjects programmed one of two nearly identical VCR simulations. To program a setting, the subject must first click on the setting's radio button and then use the up or down arrow to reach the target value. As suggested by Figure 1, to set the start time the subjects must set start-hour, start-10min, and start-min. To set the end time, they set end-hour, end-10min, and end-min. In the interface used here the three radio buttons to set start time are in one row and the three buttons to set end time are in another row.

In this study we manipulated whether or not the interface had buttons above each column of radio buttons. For the button (BTN) condition (shown in the Figure 1), if a subject had just finished programming, for example, start-hour and now wished to program start-10min, they would first have to deselect the current column button and then select the next column button before they could click-on the "Start-10min" radio button. In the no button (noBTN) condition, subjects could select any radio button at any time.

How easy is it to manipulate the decomposition of the task-to-device rule hierarchy for a particular device? How persistent would the influence of practice in one task environment (either BTN or noBTN) be when subjects were transferred to the other task environment?

We hypothesized that due to the role of prior knowledge, subjects in the noBTN condition would adopt a *by-row* strategy in which they would program all of start time (hour, 10min, and min) before going off to program something else. In contrast, the BTN condition increases the perceptual-motor cost of this "natural" by-row strategy by requiring subjects to click column buttons on and off prior to selecting a radio button in another column. Consequently, the BTN interface would seem to encourage an unnatural *by-column* strategy (i.e. setting start- and end-hour, then start- and end-10min, and then start- and end-min).

Strategies Adopted on BTN and noBTN intefaces

Each subject programmed 8 different shows to the criterion of two successful trials per show. Half of the subjects (32) programmed the first four shows with BTN and half with noBTN. For the last four shows they switched interface conditions. The dependent variable in this report was the strategy used by subjects to program time; either by-row or by-column.

By trial 4, 30/32 noBTN(tr4) subjects used the by-row strategy. With all else equal, subjects preferred to program time as a unit. In contrast, 21/32 BTN(tr4) subjects used the by-column strategy. A very simple manipulation of the perceptual-motor cost resulted in dropping a strategy that was congruent with prior knowledge for a strategy that in some way ran contrary to prior knowledge.

Moreover, further analysis revealed that such minor interface differences affect effort and performance, as well. The average time spent memorizing the target values on the hard interface (BTN) was ~10.4% higher than on the easy interface (noBTN). This resulted in *~20% less programming errors on the hard interface than on the easy*.

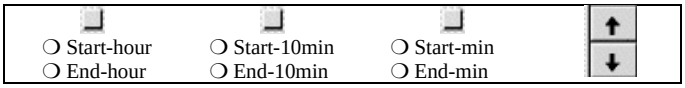


Figure 1. Partial simulated VCR 3.0 BTN interface.

Transfer from BTN to noBTN and vice versa

The results for trial 8 (see Table 1) show that when transferring from a hard interface (BTN) to an easy interface (noBTN), the methods acquired with the hard interface persist. However, in going from easy to hard, subjects quickly adapt to the hard interface.

From the subject's perspective, during the second phase of the study (shows 5-8) the perceptual–motor cost of the by-column strategy greatly decreased and the memory cost stayed about the same. Hence, a strategy that worked well under the conditions of the BTN task environment still worked and was easier to implement under the noBTN task environment.

	Trial 4	Trial 8
noBTN-BTN group	~94% used by-row	75% used by-column
BTN-noBTN group	~66% used by-column	~44% used by-row

Table 1. Strategy use on Trials 4 and 8.

Summary

People do not use one tool or one piece of software. Rather, we work in multiple task environments and our cognitive processes seem adapted to these environments, as opposed to particular subtasks in any given environment.

Acknowledgments

The research and writing was supported by a grant from the Office of Naval Research ONR #N000140310046.

References

Carroll, J. M., & Rosson, M. B. (1987). Paradox of the active user. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction.* Cambridge, MA: MIT Press.

Towards an Affective Cognitive Architecture

Markus Guhe, Wayne D. Gray, Michael J. Schoelles, Quiang Ji
(guhem, grayw, schoem, jiq@rpi.edu)
Rensselaer Polytechnic Institute

Affective state can influence users' cognitive processing capabilities and hence their productivity (Picard 1997). The first goal of our research is to develop methods to timely and efficiently recognize negative user affective states, model their influence on cognition and behavior, and provide the most appropriate intervention in a timely manner to return the user to his/her productive state. The second, more distal, goal is to develop an integrated architecture of affect and cognition. There are four challenges facing this initiative. (1) Users' affect develops over time, and its expressions vary significantly with individual and context. (2) Affective state observations from a given sensory source are ambiguous, uncertain, and incomplete. (3) The influence of cognition on affective state and vice versa is not well understood. (4) Interventions to improve user performance must be timely and effective.

Our approach contrasts with the state-of-the-art in augmented cognition as well as in affect-based augmentation. The former assumes normative performance and fails to adapt to the user's current affective state. The latter tends to have low-to-no cognitive fidelity, failing to understand the cognitive activities that lead to the observed user state. (But see Hudlicka 2003 for an overview of recent approaches.) Our framework addresses both sets of challenges.

The proposed framework has five major parts: data sensing, user affective modeling, user cognitive modeling, an integrated affective–cognitive model, and a probabilistic user assistance model. Data sensing entails various visual, physiological, and behavioral data.

The *Rensselaer Bayesian Affect Recognition System* (R-BARS) determines the user's most likely affective states using both current and stored sensory data. The model's *context component* represents information about relevant environmental factors such as time of day and type of work. The *affective state component* represents the affective states the system can infer. The affective state we are currently investigating is confusion. The *profile component* may include experience, skill level, etc. It enables us to adapt the model to individual differences. Finally, the model's *observation component* integrates current data with the longitudinal data record collected during the session.

The *Rensselaer Cognitive Architecture of Cognition* (RAAC) is based on ACT-R (Anderson & Lebiere, 1998). Our current focus is on "model tracing"; i.e., the step-by-step tracing of human performance in real-time. Although model tracing in real-time has been repeatedly demonstrated at the 10s level of analysis (Anderson, 2002), behavior at the 100ms level, such as point-of-gaze, is viewed as non-deterministic.

We use a dual-task to induce confusion in the user. The math task is a simple addition/subtraction of two-digit numbers. The user must decide whether the result presented on the screen is correct. The audio task is to determine whether a letter is lower or higher in the alphabet than the previously presented one. For example, for the sequence a–c–b the user must press the ⟨higher⟩ key first and then the ⟨lower⟩ key. The tasks are presented in eight 10min blocks that are subdivided into 36s intervals. For each task, one stimulus is presented every 2s, 4s, or 6s, so that there are 18, 9, or 6 stimuli per interval. By varying the rate of presentation between intervals for each task we get 9 different combinations, e.g. 6–18: 6 stimuli per interval in the math task and 18 stimuli in the audio task. Varying these combinations varies the user's level of confusion, which is confirmed by pilot data. Although performing the audio task at a rate of 2s is manageable, the math task proves very challenging – in particular in conjunction with the audio task. On trials with challenging (18–18) schedules the performance over a 10min block can drop below 20% for the math task (it is typically around 60%). The audio task is usually significantly better; even for challenging schedules the performance hardly drops below 80%.

The cognitive implications of the user's affective state are established by analyzing the deviation of user behavior from the optimal path determined by the model. We will interpret the difference between expected and observed behavior as the influence of affect on cognition and behavior.

In combining R-BARS with RAAC our proximal goal is to mimic the *effect of affect* by identifying low-level parameters of the cognitive architecture that, when varied, mimic the cognitive and behavioral consequences of affective state. Candidate parameters include noise in memory activation and noise in production choice, cf. Belavkin (2001).

We face a profound challenge. Even developing a reliable affective–cognitive model for the task at hand is demanding. Yet, even a partially validated integrated affective–cognitive model would be an important step forward for understanding of the relationship between cognition and affect.

Acknowledgments

Research supported by a grant from Office of Naval Research.

References

Anderson, J. R. 2002. Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science, 26*(1):85-112.

Anderson, J. R., & Lebiere, C. eds. 1998. *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Belavkin, R. V. 2001. *Modelling the inverted-U effect with ACT-R*. Proceedings of ICCM–4, Mahwah, NJ.

Hudlicka, E. 2003. To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies, 59*(1):1–32.

Picard, R. W. 1997. *Affective computing*. Cambridge: MIT Press.

Cognitive Science Needs Powerful Research Strategies

Bernadette Guimberteau (Bern0@mac.com)
Graduate School of Education, UC Berkeley
4533 Tolman Hall #1670, Berkeley, CA 94720 USA

It seems reasonable to believe that problem solving with the Tower of Hanoi (TOH) task has been studied thoroughly enough to be well understood. Yet, the discovery of a new class of affordance-based strategies was reported recently (Guimberteau, 2003), with the finding that problem solving strategies issued from that class are capable of explaining the famous TOH protocol from Anzai & Simon (1979).

The above discovery enriches the set of known strategies for the task, formalized several decades ago (Simon, 1975). Importantly, it also raises an issue, given the prolific nature of the research area concerned with problem solving with the TOH task: Why has not the new class been specified earlier?

Such a question may be dismissed on the grounds that certain aspects of scientific inquiry are not explainable. Another approach is to consider that question closely, as an opportunity to learn from the past. The present analysis takes a first step in the latter direction. It examines past modelizations of Anzai & Simon (1979)'s first problem solving episode (Episode 1), looking for clues to explain why those modelizations have not considered the affordance-driven explanation of the learner's problem solving behavior.

The strategy put forth to explain Episode 1 – Selective Search – constitutes a classic result in the cognitive science literature. The strategy simplifies search by not repeating moves (Anzai & Simon, 1979; Ruiz & Newell, 1989; VanLehn, 1991). It is made of three heuristics: "Don' t reverse a move just made", "Don' t move the same disk twice in a row," and "Don' t transfer the smallest disk and later return it to its previous peg." Those heuristics constrain move selection in such a way that search becomes unnecessary after the first move.

Good fits and convergence of results form the basis of the credibility of the Selective Search strategy. Past accounts simulate problem solving in Episode 1 using the Selective Search strategy and show that they can reproduce the transitions between the learner's strategies (Anzai & Simon, 1979), with a good fit to the learner's moves (Ruiz & Newell, 1989), and to her goal utterances as well (VanLehn, 1991). Specifically, Anzai & Simon (1979) build a production system that can make the same strategy transitions as the ones they hypothesize for their human learner. The Soar simulation from Ruiz & Newell (1989) produces moves that correspond to 77% of the subject's observed moves in her first problem-solving episode. The production system from VanLehn (1991) accounts for all but one of the subject's 130 observed moves and all but 3 of her 41 goal utterances.

A close examination of the Selective Search modelizations reveals three observations. Certain aspects of the episode (e. g., a bottom-disk focus) are not explained by the Selective Search strategy, requiring additional modeling mechanisms. In addition, two conflicts between the heuristics and data from the episode need addressing Finally, the fact that those simulations overlook certain aspects of the data raises the possibility that other mechanisms may be able to explain both the unexplained and the explained data. Indeed, a stronger explanation of the problem solving strategies used in Episode 1 is affordance-driven, and not based on the above Selective Search heuristics: The subject recognizes task-specific affordances during problem-solving, from which she devises affordance-driven strategies (e.g., moving the second smallest disk first because that disk affords moving less than the smallest one.) In other words, the Selective Search characterization rests on a foundation that does not address issues of exhaustive protocol coverage, conflict prevention, and parsimony.

The Selective Search strategy discussed here constitutes one of the classic results of cognitive science. Yet, the two major criteria underlying its credibility – good fits with the protocol data and convergence of past analyses of those data – are insufficient. Other classic results may suffer from similar limitations. The present finding calls for the definition of powerful strategies for cognitive science research.

References

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86*, 124-140.

Guimberteau, B. (2003). Developing problem-solving competence: A new model and a new class of strategies with the Tower of Hanoi task. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Ruiz, D., & Newell, A. (1989). Tower-noticing triggers strategy-change in the Tower of Hanoi: a Soar model. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society (pp. 522-529).* Hillsdale, NJ: Erlbaum.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology, 7*, 268-288.

VanLehn, K. (1991). Rule acquisition events in the discovery of problem solving strategies. *Cognitive Science, 15*, 1-47.

Probability Intervals and Sample Constraints

Patrik Hansson (patrik.hansson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Peter Juslin (peter.juslin@psyk.uu.se), Anders Winman (anders.winman@psyk.uu.se)
Department of Psychology, Uppsala University
SE-751 42, Uppsala, Sweden

The success of on agents realistic probability assessment for an unknown quantity are greatly enhanced if a pre-stated interval is evaluated, rather then produced by the same agent (Hansson, Winman and Juslin, 2004). In order to explain this *format dependence effect* we have developed what we call a *Naive Sampling Model* (here after NSM) (Juslin, Winman and Hansson, 2004).

The NSM assumes that a Subjective Probability Distribution for an unknown quantity is assessed by a retrieval of similar objects from memory which provide a sample distribution. This sample distribution is directly taken as an estimate of the corresponding population distribution. With interval production the sample dispersion is interpreted as an estimate of the population dispersion, with the fractiles in the distribution defining the upper and lower limit for the interval. Because of the fact that sample dispersion is a biased estimator of the population dispersion, failing to correct this bias (Kareev et al, 2002) leads to intervals that are too narrow, thereby producing overconfidence.

The current study tests the NSM with a special eye on sampling constrains. We manipulate how much knowledge (possibly sample size) that participants could use to make these inferences.

Experiment

The stimuli used in the current experiment were fictive income figures for 136 different companies. The companies were divided in to five different fictive regions (the regions were supposed to function as cues). Two conditions (13 participants in each) were used: one (4XTraining) where the participant trained on 4 x 136 trials, the other (2XTraining) where the participants trained on 2 x 136 trials. Feedback was given under the training phase. After going trough the training phase participants in both conditions completed a test phase consisting in making point estimates and producing intervals under three different confidence levels (50, 80 and 100%) regarding the income of the 136 companies.

Results and Discussion

Participants in the 4XTraining condition produced significantly more correct point estimates in the test phase than the participants in the 2XTraining condition ($t(24)=2.38$, $p=.03$). This indicated that they had received more knowledge (i. e. larger sample). Figure 1 (Left Panel): although the participants in the 2XTraining condition had learned less, they were not worse calibrated then those who participated in the 4XTraining condition. Both groups are overconfident in their interval productions. Monte Carlo simulation of the NSM on the same database used in the experiment showed that sample size (n) = 5 fitted the data best for both groups. One interpretation of these results is that the sample used to make these kinds of inferences is constrained by working memory limitations and that knowledge produced by the long-time summarizing of the complete sample of the observation experienced plays no part. One limitation with the model is that it does not predict the difference between the two conditions regarding the interval width (see Figure 1, Right Panel).

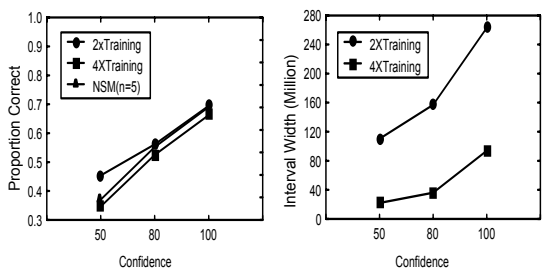


Figure 1: Left Panel: Mean proportion of correct values included in each confidence intervals in the two experimental conditions and the models performance with sample size (n) =5. Right Panel: Interval width (upper minus lower limit) for the produced interval by the participants in the two conditions

Acknowledgments

This research is supported by the Swedish Research Council.

References

Hansson, P., Winman, A., & Juslin, P. (2004). *Subjective Probability Interval: How to Cure Overconfidence by Interval Evaluation*. Manuscript submitted for publication.

Juslin, P., Winman, A., & Hansson, P. (2004). *The Naïve Intuitive Statistician: A Sampling Model of Format Dependence in Probability Judgment.* Manuscript in preparation.

Kareev, Y., Arnon, C., & Horwitz-Zeliger, R. (2002). On the Misperception of Variability. *Journal of Experimental Psychology: General*, *131*, 287-297.

Analysis of Attention Networks and Analogical Reasoning in Children of Poverty

Ruby C Harris (rcharr02@louisville.edu)
Tara Weatherholt (tnweat03@louisville.edu)
Barbara Burns (bburns@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
Louisville, KY 40292 USA

Catherine Clement (catherine.clement@eku.edu)
Department of Psychology, Eastern Kentucky University
Richmond, KY 40475 USA

Introduction

Researchers describe the development of analogical reasoning as a shift from similarity judgments based on simple perceptual feature comparisons to more complex reasoning based on common relational structures (Gentner, 1989). Given that this shift entails a selective focus on relational information, perhaps attentional development affects the development of analogy.

Recently attentional processes have been examined in terms of three networks of attention; orienting, alerting, and executive (Posner & Petersen, 1990). The executive network, an attentional control network required for the resolution of cognitive conflict, may be particularly important for analogical reasoning tasks. Recent research has shown that the same neurological pathways are activated in selective attention activities of the executive network and analogical reasoning (Duque & Posner, 2001; Luo, Perry, Peng, Jin, Xu, Ding, & Xu, 2003).

Children from low-income backgrounds have been shown to have impaired attentional and cognitive abilities (Norman & Breznitz, 1992). In the current study, individual differences in children's skills on the three attention networks are studied in order to understand the relationship between specific attentional processes and analogical reasoning in the context of poverty.

Methods

Participants were 78 children (Mean age = 56.88 mos, SD = 5.97) from low-income backgrounds. Children were assessed on computerized attention tasks designed to tap the three attention networks (Berger, Jones, Rothbart, & Posner, 2000) Analogical reasoning was assessed using the Matrices Subtest of the Kaufman Brief-Intelligence Test (Kaufman & Kaufman, 1990).

Results

Hierarchical regression was performed to predict analogical reasoning ability using median reaction time on the attention tasks. Controlling for cognitive ability, the overall model was significant, F (4, 72) = 4.091, $p < .005$. Performance on the executive attention task added a significant amount of variance (6.4%) to the model.

Discussion

The present study is unique in its examination of the relation between attention and analogical reasoning in the context of a high-risk environment. Future studies should examine how these findings relate to Halford's (1989) proposals about the impact of processing capacity on the development of analogy. Further, studies should examine how executive attention interacts with changes in domain knowledge to affect analogy task performance. The present results suggest that high functioning executive attention may be a protective factor in a high-risk environment.

Acknowledgments

Study funded by a grant from the Kentucky Science and Engineering Foundation, Inc.

References

Berger, A., Jones, L., Rothbart, M. K., & Posner, M. I. (2000). Computerized games to study the development of attention in childhood. *Behavior, Research Methods, Instruments, & Computers, 32* (2), 297-303.

Duque, D.F. & Posner, M. (2001). Brain imaging of attentional networks in normal and pathological states. *Journal of Clinical and Experimental Neuropsychology, 23*,1, 74-93.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.) *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.

Halford, G. (1989). Cognitive processing capacity and learning ability: An integration of two areas. *Learning and Individual Differences, 1*,1, 125-153.

Kaufman, A. S. & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test*. Circle Pine, MN: American Guidance Service.

Luo, Q., Perry, C., Peng, D., Jin, Z., Xu, D., Ding, G., & Xu, S. (2003). The neural substrate of analogical reasoning: an fMRI study. *Cognitive Brain Research, 17*, 527-534.

Norman, G. & Breznitz, Z. (1992). Difference in the ability to concentrate in first grade Israeli pupils of low and high socioeconomic status. *Journal of Genetic Psychology, 153,* 5-17.

Biological Limbic Systems: A Bottom-Up Model for Deliberative Action

Derek Harter (dharter@memphis.edu)
Department of Computer Science, University of Memphis
Memphis, TN 38152 USA

Evolution of Deliberative Actions

While bottom-up approaches to studying cognition have proved insightful in many ways, top-down approaches are still better at explaining deliberative cognitive processes. Deliberative actions are those that go beyond simple sensory-motor loops and seem to require some type of internal model, map or logical reasoning. Examples of deliberative actions include planning a route to navigate to a goal or performing a chain of logical inference to determine a likely course of action.

Bottom-up approaches such as Walter's tortoise (1951) and Braitenberg's vehicles (1984) are excellent models of how simple sensory-motor loops can combine to produce complex intentional behavior. Such behaviors are still mainly of the tropic type (e.g. phototropic, chemotropic), which rely on detecting and following some type of perceptual gradient in the environment. More recently, models such as Brook's (1990) subsumption architecture have shown us how collections of behavior patterns can combine in relatively flexible chains, in an emergent manner, to produce even more complex behaviors. Simple tropic behaviors are present in even the simplest of single celled organisms, while the more complex collection, chaining and combining of such sensory-motor behavior patterns appear with fish and insects.

Deliberative actions appear to require the development of more long-term memory mechanisms that allow for the storage of past experiences and for these experiences to be brought to bear on current situation. Evolutionarily, the development of the limbic system in simple vertebrates, such as amphibians, marks the first appearance of primitive hippocampal structures. The hippocampus plays the role of forming and remembering more long-term representations of experiences. It is known to participate in the formation of episodic memory, logical reasoning and cognitive maps (Dusek & Eichenbaum, 1997; Arbib, Érdi & Szentágothai, 1997). Building more deliberative systems in a bottom-up whole-system approach would therefore appear to potentially benefit from a more complete understanding of the biological limbic system.

K-IV: Basic Limbic System Model

The K-IV architecture is a model of what biologists believe may be the simplest neural architecture capable of basic intentional and deliberative actions, the limbic system (Kozma, Freeman & Érdi, 2003). The purpose of the K-IV is to model a complete autonomous organism, in a bottom-up manner, to understand better the neurodynamical mechanisms involved in intentional and deliberative behavior. The K-IV uses a neural population model (called K-sets) to describe the activity of large populations of neurons (as opposed to single unit or more abstract ANN models). It is a highly-recurrent multi-layer model of the important neurological structures of the basic limbic system.

We have been developing pieces of the K-IV for use as control mechanisms in autonomous vehicles for exploration and navigation problems for NASA. We have developed discrete simplifications of the K-set neural population models for use in such autonomous agent simulations (Harter & Kozma, submitted). In this work we will present some of our results on modeling and implementing pieces of the K-IV model, including how nonconvergent dynamics form perceptual categories (Harter & Kozma, in press) and how such dynamics may be used to learn and control behaviors in an autonomous agent (Harter & Kozma, 2004).

Acknowledgment

This work was supported by NASA Intelligent Systems Research Grant NCC-2-1244.

References

Arbib M.A., Érdi, P. and Szentágothai, J. (1997). *Neural Organization: Structure, Function Dynamics*. Cambridge, MA. MIT Press.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, *6*, 3-15.

Dusek, J.A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *National Academy of Sciences, 94*, 7109-7114.

Harter, D., & Kozma, R. (submitted). Chaotic neurodynamics for autonomous agents. *IEEE Transactions on Neural Networks*.

Harter, D., & Kozma, R. (in press). Navigation and cognitive map formation using aperiodic neurodynamics. *From animals to animats 8: The eighth international conference on the simulation of adaptive behavior (SAB'04)*. Los Angeles, CA. (in press).

Harter, D., & Kozma, R. (2004). Aperiodic dynamics for appetitive/aversive behavior in autonomous agents. *Proceedings of the 2004 IEEE international conference on robotics and automation (ICRA'04)* (p. 136). New Orleans, LA.

Kozma, R., Freeman, W. J. & Érdi, P. (2003). The KIV Model - Nonlinear Spatio-Temporal Dynamics of the Primordial Vertebrate Forebrain. *Neurocomputing, 52-54*, 819-826.

Walter, G. (1951). A machine that learns. *Scientific American, August* 60-63.

What Exactly Do Numbers Mean?

Yi Ting Huang (huang@wjh.harvard.edu)
Jesse Snedeker (snedeker@wjh.harvard.edu)
Elizabeth Spelke (spelke@wjh.harvard.edu)
Department of Psychology, Harvard University
33 Kirkland Street, Cambridge, MA 02138 USA

The correct semantics for number words has been a topic of much dispute in linguistics. This controversy bears directly on our understanding of the development of numerical concepts. The Neo-Gricean theory (Horn, 1989) posits that number words, like other scalar terms, possess a lower-bounded semantics and only receive exact interpretations pragmatically via scalar implicatures. For example, "two" would mean AT LEAST TWO and would only be interpreted as referring to exactly two entities because the speaker could use stronger terms such as "three" or "four" to refer to larger quantities. A second theory (Koenig, 1991) states that numbers have an exact semantics ("two" means EXACTLY TWO) that can generate both a set reading ("two" establishes numerosity of the set) and a distributed reading ("two" predicates the existence of two individuals of a given type). Situations that are compatible with the set readings of a number are also typically compatible with the distributed readings of all smaller numbers, leading to what appear to be lower-bounded interpretations of number words (e.g. if the number of fish in the bowl is four, then there are also three/two/one fish that are in the bowl). The salience of these distributed readings will depend heavily on the context in which the number word occurs. But critically the meaning of number words remains the same across contexts.

To test these theories we examined children's early interpretation of numbers words. Children acquire number words in a gradual and predictable sequence (Wynn, 1990) providing ample opportunity to test the initial semantics of each term. Previous research (Noveck, 2001) demonstrates that scalar implicatures appears relatively late in development. Therefore, if numbers are semantically lower bounded, we would expect to find evidence for this in children's interpretation prior to implicatures.

In Experiment 1, we presented 10 children (2;6 to 3;5) who have demonstrated knowledge of "two" but not "three" (i.e. "2-knowers") with a card displaying 1 fish and another with 3 fish and asked them to select the card with two fish. A similar procedure was repeated for 3-knowers (2;8 to 3;7) and 4-knowers (2;9 to 3;9) using their most recently acquired number. 2-knowers overwhelmingly chose the card with 3 fish, an interpretation that is consistent with lower-bounded semantics without implicatures. While these results support the Neo-Gricean account, two pieces of evidence lead us to refrain from that conclusion. First, according to an Exact Semantics account, 2-knowers in this task may assign "two" to mean EXACTLY TWO but simply select out a subset of two fish from a card with three fish using a distributive reading. Consistent with this idea, 7 out of 10 2-knowers pointed specifically to two fish on the three fish card. In addition, 4-knowers, who did not differ in age from 2-knowers, rejected both card choices, consistent with exact semantics.

In Experiment 2, we minimized the possibility of a distributive reading by pushing for the perception of stimuli as a bounded set. We also provided a way for children to demonstrate an exact interpretation without having to reject both choices. First, we taught 10 2-knowers (2;6 to 3;5) to find target animals that were located in uncovered or covered boxes. Then, in the test phase, we asked them to find the box with two fish when presented with uncovered boxes with one fish and three fish and a covered box (see figure 1). 2-knowers overwhelmingly selected the covered box, suggesting that they interpreted "two" as "exactly two" and inferred that this quantity must be in the covered box. A Neo-Gricean theory would have to provide an account for why children failed to select the visible option compatible with lower bounded semantics (3 fish) when they fail to show evidence of scalar implicatures for other terms until 7-9 years of age. The Exact Semantics account provides the most natural and parsimonious explanation of these results.

Figure 1: Experiment 2 Stimuli

Acknowledgments

This research was supported by a fellowship from Harvard University to the first author and NIH Grant #R37 HD23103 to the third author.

References

Horn, L. R. (1989). *A natural history of negation*. University of Chicago Press, Chicago.

Koenig, J. (1991). Scalar predicates and negation: punctual semantics and interval interpretation. *Chicago Linguistic Society 27, Part 2: Parasession on negation*, 140-55.

Noveck, I. (2001). When children are more logical than adults: experimental investigation of scalar implicature. *Cognition*, 78, 165-188.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, 36, 155-193.

Getting from Here to There: The Effects of Direction Type and Gender on Navigation Efficiency

Alycia M. Hund (amhund@ilstu.edu)
Department of Psychology, Illinois State University
Campus Box 4620, Normal, IL 61790-4620 USA

Introduction

Finding our way from place to place is essential to everyday functioning. Often, we rely on information from others to help us navigate. For example, people follow directions to get to unfamiliar destinations, such as airports and hospitals. One important goal is to determine the most effective way to give and follow directions.

What types of information are most effective? Previous research has focused on two common direction types: those using landmark descriptors (e.g., go toward the arena on Main St.) and those using cardinal descriptors (e.g., go east on Main St.). In general, landmarks are helpful naviagation tools. For instance, routes with landmarks are learned more quickly than routes without landmarks (e.g., Jansen-Osmann, 2002; McFadden, Elias, & Saucier, 2003; Saucier et al., 2002). Many studies have also examined the effects of gender on navigation (e.g., Lawton, 2001; Sholl, Acacio, Makar, & Leon, 2000). Findings have revealed gender differences in navigation tasks, with men often outperforming women.

The present experiment investigated whether landmarks or cardinal directions were more effective as navigation tools and whether there were gender differences in navigation efficiency using these cues. We predicted that people would navigate faster and more accurately when given cardinal directions than when given landmarks and that men might navigate more efficiently than women.

Method

Ninety-two undergraduate students (46 males, 46 females) participated for extra credit in psychology courses.

A fictitious model town (6 ft. 6 in. x 4 ft.) served as the experimental space. The town contained 17 landmarks marked by unique pictures and labels (e.g., hospital). The town also contained 30 streets marked by blue tape and street names (e.g., Memory Lane). Bound sets of note cards contained directions for navigation: one with landmark directions and another with cardinal directions. A toy car was used during navigation.

Participants were randomly assigned to either the landmark condition or the cardinal condition. Participants in the landmark condition received directions involving landmarks (e.g., go toward the arena on Main St.), whereas participants in the cardinal condition received directions involving cardinal descriptors (e.g., go east on Main St.). The routes were identical in both conditions; however, the descriptions differed based on condition. Routes started at a landmark, included four turns, and ended at a destination. The order of routes was counterbalanced across participants.

During the familiarization phase, the experimenter pointed out the four cardinal directions and the 17 landmarks. Then, participants were given 30 seconds to familiarize themselves with the town. On each trial, the experimenter placed the toy car at a starting location and said, "Go." Participants read a set of directions and moved the car so it followed the directions to the destination.

Navigation time was calculated by averaging the time for all 17 trials. The total number of errors was calculated by summing the errors for all 17 trials. Errors included reversing, making a wrong turn, ending at the wrong destination, and not finishing the route.

Results and Discussion

Our main objective was to examine how quickly and accurately men and women navigated based on cardinal and landmark directions. As predicted, participants were significantly faster and more accurate when following cardinal directions than when following landmark directions. In addition, men navigated significantly faster than did women. These findings generally support our predictions, providing valuable information about the processes by which men and women use landmarks and cardinal directions to navigate from here to there.

Acknowledgments

I would like to thank Jennifer Minarik, Laura Noonan, and Emily Slechta for their help with data collection and coding.

References

Jansen-Osmann, P. (2002). Using desktop virtual environments to investigate the role of landmarks. *Computers in Human Behavior, 18,* 427-436.

Lawton, C. A. (2001). Gender and regional differences in spatial referents used in direction giving. *Sex Roles, 44,* 321-337.

MacFadden. A., Elias, L., & Saucier, D. (2003). Males and females scan maps similarly, but give directions differently. *Brain and Cognition, 53,* 297-300.

Saucier, D. M., Green, S. M., Leason, J. MacFadden, A., Bell, S., Elias, L. J. (2002). Are sex differences in navigation caused by sexually dimorphic strategies or by differences in the ability to use the strategies? *Behavioral Neuroscience, 116,* 403-410.

Sholl, J. M., Acacio, J. C., Makar, R. O., & Leon, C. (2000). The relation of sex and sense of direction to spatial orientation in an unfamiliar environment. *Journal of Environmental Psychology, 20,* 17-28.

Conditions for the "Inverse Base-Rate Effect" in Categorization

Mark K. Johansen (johansenm@cardiff.ac.uk)
School of Psychology, Cardiff University
PO Box 901, Cardiff, CF10 3YG, UK

Nathalie Fouquet (n.fouquet@ucl.ac.uk)
David R. Shanks (d.shanks@ucl.ac.uk)
Department of Psychology, University College London
Gower Street, London WC1E 6BT, UK

Background

The inverse base-rate effect (Medin & Edelson, 1988) is a paradoxical result in human category learning. It occurs, see Figure 1 left, after participants have been trained over a series of trials with corrective feedback to categorize pairs of features into high-frequency (C) and low-frequency (R) categories, where each category has a perfectly predictive feature (PC or PR) and a shared, imperfectly predictive feature (I). The term "inverse base-rate effect" reflects the fact that when tested with the conflicting cues together (PC+PR, Figure 1 left), participants non-normatively tend to respond with R despite its low frequency relative to C even though both cues are otherwise equally predictive of their categories.

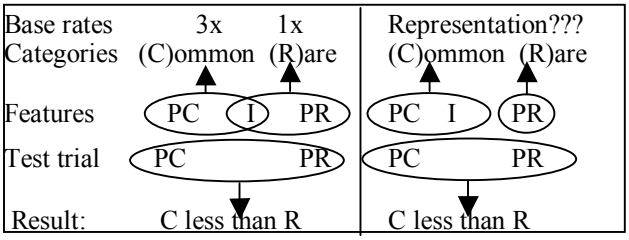


Figure 1: Left: Abstract category structure plus test trial (terminology from Kruschke, 1996). Right: Hypothesized asymmetric representation in relationship to the test trial.

Experiments

One of the most persistent theoretical explanations (e.g. Kruschke, 1996; Medin & Edelson, 1988) of the inverse base-rate effect is that the learned category representations are asymmetric, Figure 1 right, (for any of several reasons which we don't have space to describe) but that, based on similarity to this representation, the decision-making at test is normative. The purpose of our research was to evaluate whether asymmetric representation is a necessary and sufficient condition for the inverse base-rate effect in either a trial-by-trial category-learning task with corrective feedback or a purely decision-making task based on a single presentation of summary information: base-rate information together with feature-category relationships as specified in either the left or right hand sides of Figure 1, symmetric or asymmetric respectively. The four experimental conditions are in Table 1. Note that the results for the trial-by-trial learning of the symmetric structure are from Kruschke (1996).

Results

The results, see Table 1, for the pure decision-making task (N=33) on the summary information for the Symmetric structure indicate strong use of the explicitly presented base-rate information (C=0.94 > R=0.06) compared to the results from trial-by-trial learning of the Symmetric structure (C=0.35 < R=0.61). The results of trial-by-trial learning on the Asymmetric structure (N=16) show a significant inverse base-rate effect (C=0.27 < R=0.63) indicating that asymmetric representation in the context of the learning task is sufficient to produce an inverse-base rate effect. However, the results of the pure decision-making task on the Asymmetric summary information (N=33) show the absence of an inverse base-rate effect (C=0.58 > R=0.42). This indicates that asymmetric representation is not by itself sufficient to produce an inverse base-rate effect, possibly because the base-rate information is presented explicitly. Nevertheless, the fact that the results for the pure decision-making task on the Asymmetric structure are qualitatively closer to an inverse base-rate effect than the results for the Symmetric structure is consistent with Asymmetric representation being a necessary condition whose impact is overcome by the influence of the explicitly summarized frequency information.

In summary, asymmetric representation of the categories may be a necessary condition for the inverse base-rate effect, but it is not by itself a sufficient condition.

Table 1: (**C**)ommon and (**R**)are response proportions for perfectly conflicting cues (trials PC+PR) by task

	Task category structure			
Learning procedure	Symmetric		Asymmetric	
	C	**R**	**C**	**R**
Pure Decision Making	0.94	0.06	0.58	0.42
Trial-by-Trial Learning	0.35	0.61*	0.27	0.63

*Kruschke (1996) sum < 1 because of other possible responses.

References

Kruschke, J.K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 22:3-26.

Medin, D.L. & Edelson, S.M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, 117(1):68-85.

Dependency-Directed Reconsideration

Frances L. Johnson (flj@cse.buffalo.edu)
Department of Computer Science and Engineering; Center for Cognitive Science
University at Buffalo, The State University of New York, 201 Bell Hall, Buffalo, NY 14260-2000, USA

Stuart C. Shapiro (shapiro@cse.buffalo.edu)
Department of Computer Science and Engineering; Center for Cognitive Science
University at Buffalo, The State University of New York, 201 Bell Hall, Buffalo, NY 14260-2000, USA

Introduction and Background

If a knowledge representation and reasoning (KRR) system gains *new* information that, in hindsight, might have altered the outcome of an earlier belief change decision, the earlier decision should be re-examined. We call this operation *reconsideration* (Johnson & Shapiro, 2004), and the result is an optimal belief base regardless of the order of previous belief change operations. This is similar to how discussion in a jury room can help jurors to optimize their interpretation of the evidence in a trial, regardless of the order in which that evidence was presented.

To simplify our example, we assume a global decision function is used in the belief change operations, and it will favor retaining the most preferred beliefs as determined by a linear preference ordering ($\succeq$). Any base can be represented as a sequence of beliefs in order of decending preference: $B = p_1, p_2, \ldots, p_n$, where p_i is preferred over p_{i+1} ($p_i \succeq p_{i+1}$).

Reconsideration requires maintaining a set of all beliefs that have ever been in the belief base at any time (effectively, the union of all past and current bases), $B^{\cup}$. The base produced by reconsideration is defined as $B^{\cup}!$ where $!$ is a consolidation operation (which eliminates *any and all* inconsistencies) (Hansson, 1999).

A base, $B = p_1, p_2, \ldots, p_n$, is optimal if it has the most credible beliefs possible without raising an inconsistency: i.e. it is consistent and there is no $B' = q_1, q_2, \ldots, q_m$ s.t. $B' \subseteq B^{\cup}$, B' is consistent, and either $B \subset B'$ or $\exists q_i$ s.t $q_i \succeq p_i$ and $p_1, p_2, \ldots, p_{i-1} = q_1, q_2, \ldots, q_{i-1}$.

Dependency-Directed Reconsideration

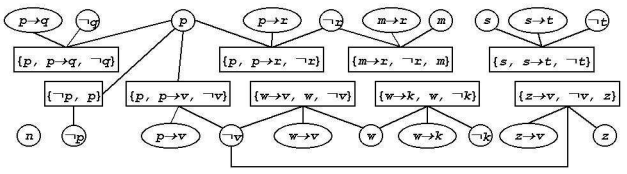


Figure 1: A graph showing the elements of $B^{\cup}$ (circles/ovals) of a KS connected to their minimally inconsistent sets (rectangles), where $B^{\cup} = \neg p, p, p \rightarrow q, p \rightarrow r, m \rightarrow r, s \rightarrow t, w \rightarrow v, w \rightarrow k, p \rightarrow v, z \rightarrow v, n, \neg q, \neg r, w, s, \neg v, m, z, \neg t, \neg k$.

Consider the base beliefs in Figure 1 *prior* to the addition of $\neg p$. The optimal base would be $B1 = \{p, p \rightarrow q, p \rightarrow r, m \rightarrow r, s \rightarrow t, w \rightarrow v, w \rightarrow k, p \rightarrow v, z \rightarrow v, n, w, s, m, z\}$, with $\neg q$, $\neg r$, $\neg v$, $\neg t$, and $\neg k$ removed. Adding $\neg p$ to $B1$ now forces the retraction of p. MOST SYSTEMS STOP HERE.

A literal implementation of reconsideration would examine *all* removed beliefs. Dependency-Directed Reconsideration (DDR), however, only reconsiders removed beliefs whose inconsistent sets have had *changes* in the belief status of their elements. It reconsiders these beliefs in decending order of preference, updating the base as it goes and maintaining a global priority queue of beliefs yet to be reconsidered. A removed belief can return as long as any inconsistency it raises is resolved through the removal of a *less preferred* belief.

As with a literal implementation of reconsideration, DDR first produces the following changes: (1) $\neg q$ returns to the base, and (2) $\neg r$ returns to the base with the simultaneous removal of m, because $\neg r \succ m$ (consistency maintenance). However, once DDR determines that $\neg v$ cannot return to the base (due to its being the culprit for the inconsistent set $\{w \rightarrow v, w, \neg v\}$), it would would prune off the examination of the inconsistent sets containing $\neg k$ and z. The inconsistent set containing s would also be ignored by DDR — it is not connected to p in any way. This latter case is representative of the possibly thousands of unrelated inconsistent sets for a typical belief base which *would* be checked during a literal $B^{\cup}!$ operation of reconsideration, but are ignored by DDR.

DDR is an anytime algorithm: if starting with a consistent base, a consistent base is always available, and the optimality of that base improves with increased execution time. Additionally, an interrupted DDR can be continued at a later time as long as the priority queue has been maintained. If run to completion, the base will be optimal (as with reconsideration) — thus, the KRR system can make the most reliable inferences, and belief change operation order will have no effect.

Acknowledgments

This work was supported in part by the US Army Communications and Electronics Command (CECOM), Ft. Monmouth, NJ through Contract #DAAB-07-01-D-G001 with Booze·Allen & Hamilton. The authors appreciate the insights and feedback of the SNePS Research Group.

References

Hansson, S. O. (1999). *A Textbook of Belief Dynamics*, vol 11 of *Applied Logic*. Kluwer, Dordrecht, The Netherlands.

Johnson, F. L. & Shapiro, S. C. (2004). Knowledge state reconsideration: Hindsight belief revision. To appear in *Proceedings of AAAI-2004*, Menlo Park, CA. AAAI Press.

Beyond common features: The role of roles in determining similarity

Matt Jones (mattj@psy.utexas.edu) and Bradley C. Love (love@psy.utexas.edu)
Department of Psychology, The University of Texas at Austin. 1 University Station A8000; Austin, TX 78712 USA

A common assumption underlying most category learning research has been that category information is represented in terms of intrinsic properties or features (cf., Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961). However, recently there has been a growing awareness that many concepts are determined not by features but by the relationships between category members and members of other categories (Gentner & Kurtz, in press; Markman & Stilwell, 2001). For example, while a *game* cannot be defined in terms of features (Wittgenstein, 1968), it has a simple definition as something that can be *played* (Markman & Stilwell, 2001).

One intriguing implication of this idea is that relational information may be a central component of object representations (in addition to feature information), with objects playing the same roles in predicates or events being perceived as similar. For example, the concepts *hammer* and *baseball bat* might be similar because they both regularly hit other objects. A further question is whether similarity is affected by roles per se, or whether involvement in the same relationship is all that matters. For example, are hammers and baseballs similar because they both participate in the relation *hit*(*x*, *y*)? This is the prediction made by models of word learning such as LSA (Landauer & Dumais, 1997) that derive meaning from co-occurrence statistics. Because these corpus approaches are insensitive to the role an object plays in an utterance, they predict that similarity will be thematically determined, in that objects that participate in common relations will be similar regardless of the correspondence between their roles.

Method

The present experiment tests the potential contributions to similarity of roles and relations independently. In the first phase of the experiment, each subject read 16 atomic sentences and rated them according to how realistic and interesting they were. Certain nouns varied between subjects in the relations they participated in and the roles they played within their relations. The second phase consisted of a series of forced-choice similarity comparisons among these nouns, in which subjects selected which of two base words was most similar to a target word.

Results

The effect of common role was assessed using similarity comparisons in which one base word played the same role as the target and the other base played the opposite role in the same relation. For example, among subjects who read "The polar bear chases the seal" and "The collie chases the cat," 65% later selected *cat* over *collie* when asked which was more similar to *seal*. Among subjects who instead read "The seal chases the fish," only 29% chose *cat*. An analysis combining eight contrasts of this type showed a significant effect of common role ($\chi^2[1] = 64.8$, $p < 10^{-15}$) with 73% of subjects choosing the base that matched the target's role.

Tests for the effect of common relation involved comparisons in which one base word matched the target in relation but not in role, and the other base was involved in a different relation. The analysis showed a significant effect of common relation ($\chi^2[1] = 7.43$, $p < .01$) with 61% of subjects selecting the base that had appeared in the same relation as the target.

Discussion

The present results demonstrate that similarity is affected by relational information in at least two ways. First, participation in the same relation increases the similarity between objects, even if they play different (or opposite) roles. Second, people are sensitive to structure within relations, such that playing the same role further increases similarity. This structure-sensitivity implies that word learning models like LSA need to be modified to discriminate among sentential contexts, according to the role (e.g., agent vs. patient) played by the word in question.

These results also support the claim that the processes by which humans learn similarity and categories extend beyond the purely feature-based approaches currently assumed. Incorporating relational information into experiments and models will lead to a more encompassing theory that may shed light on many current unsolved problems.

References

Gentner, D. & Kurtz, K. J. (in press). Learning and Using Relational Categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside of the lab: Festschrift in Honor of Douglas L. Medin*. Washington, DC: American Psychological Association.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence, 13*(4), 329-358.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39-57.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*.

Wittgenstein, L., (1968). *Philosophical investigations*, trans. G.E.M. Anscombe. New York: MacMillan.

Attentional shift within an object and between objects in 3D space

Tetsuko Kasai (tetsu@edu.hokudai.ac.jp)
Graduate School of Education, Hokkaido University, Sapporo 060-0811, Japan
Takatsune Kumada (t.kumada@aist.go.jp)
Institute for Human Science and Biomedical Engineering,
National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8566, Japan

Introduction

Object-based attention can be indexed by an advantage of attentional shift along the same object relative to that across different objects in the pre-cuing paradigm (Egly, Rafal, & Driver, 1994). There are two accounts of the effect: within-object benefit (WOB) and between-object cost (BOC). The former is explained by prior covert scanning of a cued object (Shomstein & Yantis, 2002); the latter is by a switching cost from the cued object to the other (Lamy & Egeth, 2002). So far, these accounts are indistinguishable because the object-based attention effect is defined by relative difference in RTs between the within-object and between-object conditions. This study examined the WOB and BOC separately presenting stimuli in 3D space and showed that these can operate in different space/object coordinates respectively.

Methods

Subjects: Twelve healthy volunteers participated in Exp.1 (6 females, 19-25 years) and Exp.2 (5 females, 19-28 years).
Stimuli: Fig.1 shows stimuli presented stereoscopically, using shutter goggles (frame rate: 60 Hz per eye) with a viewing distance of 57 cm. A square (16°×16°) was overridden by a bar (17.4°×3.6°) with horizontal or vertical orientation located in back of (segmented condition), the same as (flat condition), or front of (completed condition) the square. All stimuli had crossed binocular disparity, 27.4', 13.7', 41.0', and 45.6' (Exp.1, near space) or uncrossed disparity, -27.4', - 41.0', -13.7', and -9.1' (Exp.2, far space) relative to the CRT display for the square, bar in back, bar in front, and fixation, respectively. **Procedure:** An experimental block of each condition consisted of 640 target-present trials and 128 catch trials. After presentation of the bar and square for 1,000 ms, the cue (flashed at one corner of the square) was superimposed for 100 ms. After another 200 ms, a target (dot diminishment) was superimposed at one of the corners until the subject responded. The intertrial interval was 1000 ms with blank screen. The task was to detect a target as rapidly and accurately as possible by pressing a key. On target-present trials, the target appeared at the cued corner on 75 % (valid cue) and at an uncued corner on 25 % (invalid cue).

Results and Discussion

Mean hit and FA rates were 95.9 % and 5.4 % in Exp.1, and 97.9 % and 3.2 % in Exp.2. Summary of RT results is shown in Table 1. RTs for valid trials were faster than for invalid trials, confirming pre-cueing effects. Attention shift (indexed by (invalid – valid)) were faster for the within-region than between-region conditions in the segmented condition, replicating a typical object-based attention effect. Comparing with the flat and completed conditions, the object-based effect was due to WOB in Exp. 1, but to BOC in Exp. 2.

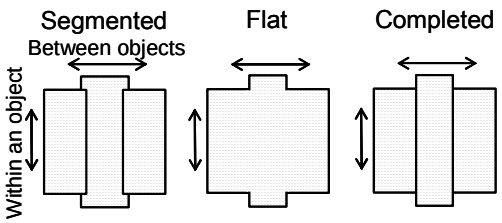


Fig.1 Schematic illustration of stimuli.

Table 1: Summary of mean RTs (ms).

	Valid	Invalid-Valid Within	Invalid-Valid Between	Object effect	WOB	BOC
Exp.1						
Segmented	300.8	3.4	12.9	9.5*	---	---
Flat	303.3	16.4	10.1	-6.3	13.0*	2.8
Completed	303.0	7.6	13.7	6.0	4.2	0.8
Exp.2						
Segmented	295.4	16.9	27.6	10.6*	---	---
Flat	293.4	17.5	20.0	2.5	0.6	7.6
Completed	297.6	17.0	19.7	2.7	0.1	7.9*

Note: WOB and BOC are shown for segmented display relative to flat and completed displays; * indicates significant effect ($p < 0.05$).

The present results showed separable mechanisms for WOB and BOC of attention. The benefit and cost may be associated with habits in different space regions (Previc, 1998): analyses of object shapes for action in near space associated with a WOB; search and orienting of objects in far space associated with a BOC. Different mechanisms to scan visual field can be driven according to stimulus context.

References

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between-object and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123, 161-177.

Lamy, D., & Egth, H. (2002). Object-based selection: The role of attentional shifts. *Perception & Psychophysics*, 64, 52-66.

Previc, F. H. (1998). The neuropsychology of 3-D space. *Psychological Bulletin*, 124, 123-164.

Shomstein, S., & Yantis, S. (2002). Object-based attention: Sensory modulation or priority setting? *Perception & Psychophysics*, 64, 41-51.

Effects of Interactivity and Spatial Ability on the Comprehension of Spatial Relations in a 3D Computer Visualization

Madeleine Keehner[1] (keehner@psych.ucsb.edu)
Daniel R Montello[2] (montello@geog.ucsb.edu)
Mary Hegarty[1] (hegarty@psych.ucsb.edu)
Cheryl Cohen[1] (cacohen@umail.ucsb.edu)
[1]Department of Psychology, UCSB, Building 551 Room 1332, Santa Barbara, CA 93106-9660 USA
[2]Department of Geography, UCSB, Ellison Hall Room 3611, Santa Barbara, CA 93106-4060 USA

Introduction and Method

This experiment was designed to investigate the roles of interactivity and spatial visualization ability in the comprehension of 3D computer visualizations.

Undergraduates were presented with a fictitious anatomy-like structure in the form of both printed 2D images and a 3D computer visualization that could be rotated in *x*, *y* and *z* dimensions. A superimposed vertical or horizontal line on the printed images indicated where they should imagine the structure had been sliced. The task was to draw the cross-section at that point. The drawings were assessed for spatial understanding using a standardized scoring scheme.

Sixty participants were randomly allocated to one of two conditions. The *active* group was allowed to rotate the computer visualization at will via keyboard controls during the drawing task. The *passive* group had no control over the movements. Using a yoked pairs design, the manipulations performed by the active participants were recorded and later played back to the passive participants, so that both members of each pair received the same visual information.

Spatial ability was measured via the Mental Rotation Test (Vandenberg & Kuse, 1978) and a modified version of Guay's Visualization of Views test (Eliot & Smith, 1983).

Results

There was no main effect of condition, indicating no significant difference between the active and passive control conditions. However, a main effect of spatial ability was found (median split; F=9.38, p<.005; Figure 1). Although the interaction between these two factors did not reach significance, pairwise comparisons revealed that high- and low-spatial participants differed significantly in the passive condition (t=2.80, p<.01), but not in the active condition (t=1.47, p>.1; Figure 1). In line with this finding, the correlation between spatial ability and performance was relatively attenuated under active control (r=.29, p>.1), compared to passive viewing (r=.51, p<.005; Figure 2).

Discussion

The data indicate that having active control of the computer visualization did not benefit overall performance. A more important predictor of success was individual differences in spatial ability. However, the contribution of this factor was stronger in the passive condition than in the active condition, i.e. when participants were allowed to manipulate the 3D model, the performance means of high and low spatial individuals were brought closer together. While low-spatial participants were helped by interactivity, this benefit did not extend to high-spatial individuals. We are currently undertaking a replication study with a more intuitive control mechanism, to establish whether these findings arose from the nature of the interface or from interactivity *per se*.

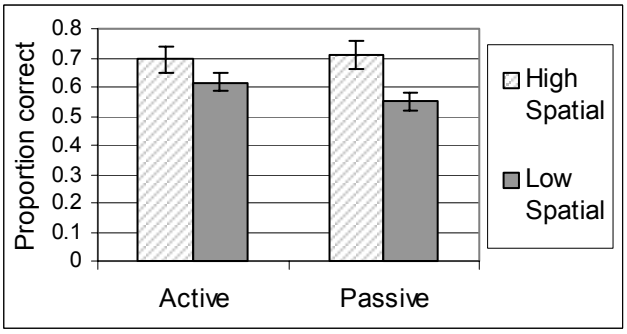


Figure 1: Performance on the cross-section drawing task by interactivity condition, as a function of spatial ability.

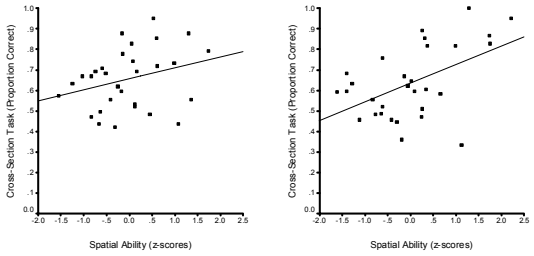


Figure 2: Correlation between spatial ability and performance, by interactivity condition.

Acknowledgments

This research was supported by NSF grant # EIA - 0313237.

References

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills, 47,* 599-604.

Eliot, J., & Smith, I M. (1983). *An international directory of spatial tests*. Windsor, Berkshire: NFER-Nelson.

Exploring the Bases of Causal Inferences

Say Young Kim (syk2@pitt.edu), Francisco J. Morales, & Erik D. Reichle (reichle@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara St., Pittsburgh, PA 15260, USA

Myers, Shinjo, and Duffy's (1987) paradoxical results showed that the degree of causal relatedness affected the likelihood of participants making causal inferences: Whereas memory for the sentences exhibited an inverted-U shape function, with participants showing the best performance in the case of moderate relatedness, sentence reading times increased linearly as the degree of causal relatedness decreased. Although recent efforts to explain these results implicate causal relatedness as being important (Reichle & Mason, in press), the factors that mediate this causal relatedness remain under-specified.

In this paper, we report the results of an experiment that examined one variable that also affects perceived causal relatedness and the propensity to make inferences—the topicality of discourse. Kim, Cho, and Han (2002) showed that topicality (a global variable that reflects overall text coherence) affects how well a topic maintains continuity in a text. In the present experiment, we manipulated the topicality of short passages by augmenting the Myers et al. sentence pairs so that they were preceded by four sentences that either did or did not maintain continuity (i.e., weak vs. strong topicality, respectively). Our main objectives were to determine if topicality affects the likelihood of making interferences, and to determine if topicality modulates the perceived causal relatedness of the actual sentences that were used by Myers et al.

Method

Participants. Thirty-eight University of Pittsburgh undergraduate students participated for extra class credit.

Stimuli Materials. Thirty of the sentence "pairs" from Myers et al. (1987) was adapted for our study. Each "pair" consisted of an outcome sentence and a sentence preceding it that could be highly, moderately, or distantly related to the outcome sentence. Each sentence "pair" was preceded by four sentences that maintained a strong or weak topicality. (Topically was assessed through ratings collected in an earlier normative study.) Six filler stories similar to the test items were also included.

Design. There were six different versions of each story, with each version being defined by the factorial combination of two within-subjects variables: relatedness (highly vs. moderately vs. distantly) and topicality (strong vs. weak). Stories were counter-balanced across participates using a Latin Square design so that each participant read five stories from each of the six conditions, plus six fillers.

Procedure. Participants read the stories in a self-paced manner as they were displayed sentence-by-sentence on a computer monitor. After reading each story, participants judged the degree of relatedness between the last two sentences (which were re-displayed in isolation) using keys corresponding to a 7-point Likert scale. After reading all of the stories, participants completed a recognition task in which they viewed 45 sentences (30 story-final sentences and 15 distractors) that were displayed one-at-a-time. Four dependent measures were recorded: (1) sentence-reading times; (2) relatedness judgments; (3) recognition accuracy; and (4) recognition latencies.

Results & Conclusions

Nine participants failed to comply with task instructions and were excluded from our analyses; data from the remaining 29 participants were analyzed using within-subjects ANOVAs. The degree of causal relatedness affected sentence reading times [$F(2,56) = 7.39$, $p < .01$], with reading times increasing linearly with decreasing relatedness. Although topicality did not reliably affect reading times ($F < 1$), it did lead to differences in the perceived relatedness, as indicated by the relatedness judgments [$F(1, 28) = 17.82$, $p < .001$]. Importantly, the Relatedness × Topicality interaction was not significant ($F < 1$), suggesting that these two variables differentially influence initial sentence processing and perceived relatedness. This finding suggests that local (inter-sentence relatedness) and global (topicality) variables independently affect the likelihood of a reader making an inference during reading. Finally, although recognition accuracy was at ceiling and hence did not differ condition ($F < 1$), the reliable Relatedness × Topicality interaction in the recognition latencies [$F(2, 56) = 3.49$, $p < .05$] suggests these variables may contribute similarly to text memory.

References

Kim, S. Y., Cho, S. W., & Han, K. H. (2002). The selective effect of cohesive devices on scientific text reading and comprehension in Korean. *The Thirteenth Annual Winter Conference On Discourse, Text & Cognition, Jackson Hole, WY, USA (Jan 26, 2002)*

Myers, J.L., Shinjo, M., & Duffy, S.A. (1987). Degree of causal relatedness and memory. *Jounal of Memory and Language, 26,* 453-465.

Reichle, E.D. & Mason, R.A. (2004). The neural signatures of causal inferences: A computational account. In F. Schmalhofer & C.A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes*. Manuscript in press.

Recency judgments and list context

Krystal A. Klein (krklein@indiana.edu)
Amy Criss (acriss@indiana.edu)
Richard Shiffrin (shiffrin@indiana.edu)
Department of Psychology, Indiana University
1101 E. 10th St., Bloomington, IN 47401 USA

Most memory experiments require participants to remember *what* events occurred, indirectly providing a measure of context availability. A more direct approach requires participants to remember *when* events occur. In Judgment of Recency (JOR) paradigms, participants study a list of stimuli and are asked to judge the recency of items from the list. In life, recency judgments can be made by associations to dates or autobiographical timelines. Although such cues are probably absent in list studies, participants can nonetheless make such judgments (Yntema & Trask, 1963), and the results can be used to make inferences about temporal context and its changes.

The current experiments utilize a study-test variant of the forced-choice judgment of comparative recency paradigm (Flexser & Bower, 1974). In each experiment, participants viewed lists of words on a computer monitor, and were subsequently tested in the following manner: two words from the list were presented, and participants indicated with a keystroke which word they had seen most recently.

A pilot study was completed in an attempt to obtain baseline data for JORs. Study lists were 90 items in length. Following the study phase, participants completed the forced-choice JOR for each of 20 pairs of words from the study list. Factors were lag (number of words studied between the two test items) and list type (fast or slow presentation time; each participant received one list of each type). Both factors were manipulated within-subjects. To our surprise, we found performance did not differ significantly from chance (50% accuracy) overall or in any of the individual experimental conditions.

Given that above chance performance had been found in earlier studies using a continuous study-test paradigm, we generated two hypotheses that might help explain this null result. First, the longest lag used in the study was 24 items, and context may change too slowly in a random word list without breaks for tests to allow above chance performance at short lags. Second, we had excluded the first ten and last ten study items from testing, in order to avoid any contamination by special strategies or effects due to primacy or recency. It could be that it is only during these parts of the list that context changes rapidly enough to allow for temporal discrimination.

Experiment 1 used longer lags (36) and compared performance between pairs in three conditions: those that contained one primacy item (primacy-middle), one recency item (middle-recency), or neither (middle-middle). Primacy and recency regions were set at length 12. The testing procedure was the same as in the pilot study. The longer lags did facilitate recency discrimination, illustrated by above-chance performance in the three conditions. However, the three conditions did not differ, even when the primacy and recency regions were limited to include only four items on each end of the list. These results suggest that while primacy and recency items receive a benefit in item encoding (as seen in recall), they do not receive better temporal encoding than other list items.

Because longer lags produced above chance performance, the hypothesis that context changes quite slowly during list presentation received some support. Nonetheless, the results seemed weaker than in earlier continuous study-test paradigms, leading us to ask what factors induce context change. In Experiment 2, participants studied a long list of items that was broken in half by the insertion of a 90-second task. There were four such tasks: 1) study of a list of faces; 2) a math task; 3), an old-new recognition test (on a subset of first-half items that would not be later tested for recency); 4) answering the following question (aimed to change internal context): "What would you do if you were invisible?" (Sahakyan & Kelley, 2002). For pairs containing one item from the first half (before the break) and one item from the second half (after the break), participants who received the recognition test performed best, followed by the 'invisible' answer condition. Performance in the face study and math problem conditions was not different from chance. These results are consistent with the idea that different tasks cause differential context change, and the pattern of results is consistent with certain puzzling results from standard memory paradigms (e.g., Shiffrin, 1970).

Acknowledgments

This research was supported in part by National Institute of Mental Health MERIT Grant 12717.

References

Flexser, A.J., & Bower, G.H. (1974). How frequency affects recency justments: A model for recency discrimination. *Journal of Experimental Psychology, 103,* 706-716.

Sahakyan, L. & Kelley, C.M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1064-1072.

Shiffrin, R.M. (1970). Forgetting: Trace erosion or retrieval failure? *Science, 168,* 1601-1603.

Yntema, D.B., & Trask, F.P. (1963). Recall as a search process. *Journal of Verbal Learning and Verbal Behavior, 2,* 65-74.

Visual and Verbal Interference in Recognition of Imitative and Mimetic Words

Yuki Kobayashi (yuki_kobayashi@nifty.com)
Department of Psychology, Kawamura Gakuen Woman's University
1133 Sageto, Abiko City, Chiba 2701138 JAPAN

Eriko Kawasaki (eriko.kawasaki@nifty.com)
Department of Psychology, Kawamura Gakuen Woman's University
1133 Sageto, Abiko City, Chiba 2701138 JAPAN

Osaka (2001) suggested that imitative words would be processed verbally and mimetic words would be processed visually. This study investigated whether visual or verbal second task would interfere with the processing of imitative or mimetic words. Our hypothesis was visual task would interfere with recognition of mimetic words, whereas verbal task interfere with recognition of imitative words.

Method

Experimental Design

The design used the reading span (high, low) as a between-participants variable, and the stimuli of memory task (figures, words) and the target word of sentence recognition task (imitative, mimetic) as within-participants variables. Dependent variables were reaction times and error rates for memory task and sentence recognition task.

Stimuli

Forty sentences including one imitative word and forty sentences including one mimetic word were used in sentence recognition task. In the half of the sentences, these sentences made sense and else did not make sense. In memory task, twenty combinations of three words or figures were used.

Participants

Participants were thirty-five female undergraduate students. All were Japanese native speakers and had normal or corrected vision.

Procedure

Reading Span Test. We measured each participant's working memory capacity by Japanese reading span test (Osaka, 2002).

Sentence Recognition Task and Memory Task. The fixation point was presented for 3000ms. After that, three figures or words were presented for 3000ms, so participants were required to memorize these stimuli. Participants answered whether the sentence presented after the figures or words could make sense as quickly and accurately as possible (sentence recognition). Three figures or words were presented again, participants answered whether these stimuli were presented previously by pressing allocated keys (memory task).

Results

Seventeen participants with a high reading span (more than six sets) and eighteen with a low reading span (less than five sets) were assigned to the high and low groups, respectively.

Recognition Task

The mean reaction times for correct responses were shown in Figure 1. Interaction between the target word (imitative, mimetic) and the stimuli of memory task (figures, words) were significant ($F [1, 33] = 7.85$, $p < .01$). Other main effects and interactions were not significant. About error rate, no main effect and interaction was significant.

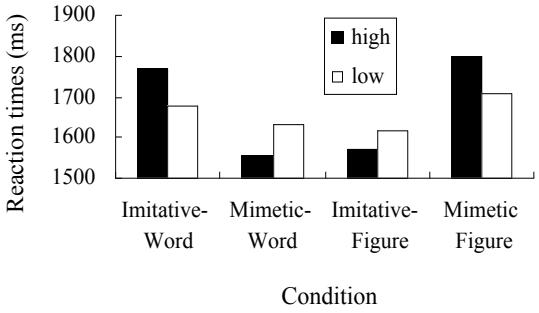


Figure 1: Reaction times for recognition of sentences (ms).

Memory Task

In word condition, reaction times for mimetic words were significantly longer than imitative words ($F [1, 33] = 4.75$; $p < .05$). In figure condition, error rate for imitative words were significantly higher than mimetic words ($F [1, 33] = 5.25$, $p < .05$).

Discussion

The results of sentence recognition task showed that verbal dual task interfered with the memory of imitative words and visual task interfered with the memory of mimetic words. These results support our hypothesis. The results of memory task suggested that there was a tradeoff between primary task and secondary task, however. The task switching between phonological loop and visuo-spatial sketchpad would affect these results. Working memory capacity did not related with performance of words' maintenance.

The relationship between Japanese spatial terms and visual factors in three- dimensional virtual space

Takatsugu KOJIMA (kojima@cpsy.mbox.media.kyoto-u.ac.jp)
Takashi KUSUMI (kusumi@mbox.kudpc.kyoto-u.ac.jp)
Graduate school of Education, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

Introduction

How spatial terms correspond to visual factors has been a topic of interest (e.g. Regier & Carlson, 2001). In particular, how a spatial term categorizes a space is an important problem (Hayward & Tarr, 1995). Most studies have considered a line as a prototype of a spatial term. In this view, some spatial terms have the same prototypical line. However, it is possible to choose an appropriate spatial term from similar spatial terms that are based on a line. Therefore, a spatial term also has a prototypical point on a prototypical line that distinguishes it from another spatial term, also categorized by a line (Kojima & Kusumi, 2002).

This study examined the prototypical points for three Japanese spatial terms categorized by a spatial line, which are recognized as differing from each other, by using a method of adjustment instead of rating tasks. We also examine the effects of visual factors on choosing a spatial term.

Method

In this experiment, we focused on three Japanese spatial terms, *mae* (front), *ushiro* (back), and *saki* (ahead), and three visual factors, the distance between objects, the height of the viewing point, and the position of the viewing point.

Forty-five Japanese graduate or undergraduate students participated in this experiment. They were divided into three groups of fifteen, and each group participated in each session.

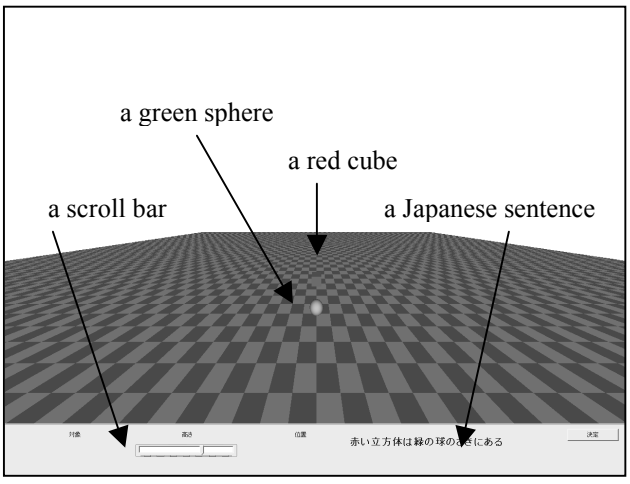


Fig.1. An example of a stimulus in the experiment

The experiment was run on a computer with a 17-inch monitor (Fig. 1) and consisted of one session for each of the three spatial terms. In any trial, only one of three scroll bars was shown and used to adjust a different factor: the distance between the green sphere and the red cube, or the height or the position of the viewing point. A Japanese sentence including a spatial term was presented in the lower part of the screen (*e.g.*, in English the sentence might be "A red cube is in front of a green sphere."). When the participant adjusted one factor, the coordinates of the other two factors remained fixed. Seven patterns were used in each session. The participants adjusted the pattern for each sentence.

Results and Discussion

The point of subjective equality (PSE) was computed from the data. The PSE values for each condition for the three spatial terms were analyzed using one-way ANOVA and Tukey's HSD. We found significant differences between all of the conditions: the position of the viewing point ($F(2,447)=54.06$, $p<.01$); the height of the viewing point in the near ($F(2,447)=92.28$, $p<.01$), middle ($F(2,447)=75.42$, $p<.01$), and far ($F(2,447)=48.27$, $p<.01$) distance conditions; and the difference in the distance between the green sphere and red cube in the near ($F(2,447)=58.22$, $p<.01$), middle ($F(2,447)=53.32$, $p<.01$), and far ($F(2,447)=68.75$, $p<.01$) distance conditions. Tukey's HSD indicated a significant difference between all but four of the conditions with respect to the spatial terms. The exceptions were the height in middle and far distance conditions, and the distance in middle and far distance conditions.

The result indicated that three visual factors affect the choice of spatial terms. It follows that humans can distinguish one spatial term from another, based on certain visual factors, even if the spatial terms are linked to a similar prototypical line in space. In addition, considering the values of the mean PSEs, it may be said that human beings choose an appropriate spatial term by differentiating visual differences accurately.

References

Hayward, W., & Tarr, M. (1995). Spatial language and spatial representation. *Cognition*, 55, 39-84.

Kojima,T., & Kusumi,T. (2002). The Structure of Linguistic Spatial Representation: A test for psychometric structure using Japanese spatial terms. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 1013.

Regier, T., & Carlson, L. A. (2001). Grounding Spatial Language in Perception: An Empirical and Computational Investigation. *Journal of Experimental Psychology: General,* 130, 273-298.

The effect of a character's emotional shift on narrative comprehension

Hidetsugu Komeda (komeda-h@xk9.so-net.ne.jp.ac.jp)
Takashi KUSUMI (kusumi@educ.kyoto-u.ac.jp)
Faculty of Education, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan

While reading a story, readers feel happy when good things occur and worry when characters are in danger (Zwaan, 1999). When readers understand the narrative, they construct situation models (Kintsch, 1998). Situation models are multidimensional representations consisting of five dimensions: time, space, causation, intentionality, and the protagonist (Zwaan & Radvansky, 1998). Zwaan, Langston, & Graesser (1995) developed the event-indexing model to explain how readers construct coherent multidimensional representations of situations. According to the model, events and the actions of characters are important in situation model construction. Readers can represent and update characters' emotions (de Vega, Le'on, & Diaz, 1996). The present experiments focus on the effect of a character's emotion when people read a story in which a change in the character's emotion is induced.

Method

Participants. Thirty Japanese speakers were recruited at Kyoto University.

Materials. The materials were 16 stories (4 themes × 4 emotional states: worry-relief, relief-worry, worry-worry, and relief-relief). Emotional-shift versions were worry-relief and relief-worry. No-shift versions were worry-worry and relief-relief. There were 24 sentences in each story. Presentation of the versions of the stories was counterbalanced with a 4 × 4 Latin square. Each participant read four stories.

Procedure. Participants were instructed to read the stories in order to appreciate the story and sympathize with the characters. Stories were presented one sentence at a time on a CRT. Reading was self-paced; readers pressed the space bar to proceed. Reading time of each sentence was collected. After finishing each story, readers rated their emotional response to each on five 7-point scales (sympathy, similarity between the character and the reader, experience, interest in the theme, and readability of the story).

Results and Discussion

We performed multiple regression analyses of reading times to assess that reading times could be predicted by the temporal breaks, causal breaks, and a character's emotional shift. Table 1 presents the b-weights from the multiple regression analyses. As Table1 indicates, temporal discontinuities caused sentence reading times to increase, suggesting that the temporal dimension is crucial. The result is consistent with Zwaan, Magliano, & Graesser's (1995) study. A character's emotional shift also caused sentence reading times to increase. Words and serial positions have consistently been robust predictors of reading times (e.g., Zwaan et al., 1995). These current results indicate that readers monitor temporal continuity and represent a character's emotion. The multiple regression analyses suggested that a character's emotional shift caused sentence reading times to increase; therefore, readers monitored a character's emotional shift during the on-line reading process. We conclude that, when readers monitor the dimensions of the protagonist in an event-indexing model, emotions similar to those of that character are invoked.

Table 1 B-Weights

Variable	B-Weights
Temporal breaks	134.0*
Causal breaks	8.6
Emotional shifts	209.1*
Words	49.5***
Serial positions	-32.2***
R^2	.30

p < .05, ** p < .01, *** p < .001

Table 2 The difference of score in characters' emotional shifts (standard deviation) range: 1-7

Theme	Examination		Moving		Party		Marriage	
	S	N	S	N	S	N	S	N
1	4.7 (1.3)	4.5 (1.7)	4.1 (2.2)	2.9 (1.8)	3.4 (2.0)	4.3 (2.2)	1.6 (.96)	1.2 (.58)
2	4.0 (1.5)	4.4 (1.8)	4.7 (1.8)	4.1 (2.1)	5.1 (1.3)	5.1 (2.2)	2.5 (1.1)	2.9 (1.4)
3	5.7 (1.1)	5.4 (1.4)	5.6 (1.1)	5.5 (1.3)	5.9 (.89)	6.5 (.76)	4.8 (1.4)	3.9 (1.7)
4	4.5 (1.6)	4.0 (1.7)	5.6 (1.2)	4.0 (1.5)	4.0 (1.3)	5.1 (1.8)	4.9 (1.6)	3.9 (1.4)
5	3.9 (1.5)	4.2 (1.6)	5.3 (1.4)	4.0 (1.9)	3.7 (1.3)	4.6 (1.4)	4.7 (1.6)	4.2 (1.4)

Note. S: shift, N: no-shift

1: sympathy, 2: similar thinking and action to the character 3: experience, 4: the interest in the theme, 5: the readability of the story.

Table 2 displays the score differences in characters' emotional shifts. As Table2 reveals, interest in the story theme was higher in the shifting version than in the no-shift version, for the moving and the marriage story (t (28) =3.3, p =.003, t (28) = 1.9, p = .074). Interest in the theme was higher in the no-shift version than in the shifting version, for the party story (t (28) = -1.9, p =.069). The readability of the story was higher in the shifting version than in the no-shift version for the moving story (t (28) =2.1, p =.043). On the other hand, the readability of the story was higher in the no-shift version than in the shifting version for the party story (t (28) =-1.8, p =.078). Because the reader's experience was higher in the no-shift version than in the shifting version (t (28) =-2.1, p=.049), the party theme exhibited a pattern different from that of the other two themes. The present findings suggest that a character's emotional shift influences the reader's on-line and off-line processes.

When Coordination is Worth a Thousand Words: the Role of Gesture in Grounding

Meredyth Krych Appelbaum (krychm@mail.montclair.edu)
Joan Schultheiss (schultheisj1@mail.montclair.edu)
Julie Banzon (banzonj1@mail.montclair.edu)
Department of Psychology, Montclair State University
1 Normal Avenue, Montclair, NJ 07043 USA

Introduction

In most conversations, people rely on a process of grounding, in which people establish the mutual belief that they have been understood (Clark, 1996). While the majority of research on grounding has focused on speech in conversation, we examine the grounding process when people may coordinate only through gesture, compared to speech + gesture, and to speech-only conditions.

Grounding is fundamentally about coordination between people. Clark & Krych (2004) demonstrated how speaking and listening are incremental processes and how many of those increments are determined jointly—whether through speech or gesture.

In this study, our aim is to achieve a better understanding of the role of gestures and speech in communication and what gesturing alone can further inform us of the grounding process.

Lozano and Tversky (2003) have studied gestures and speech+gestures when people are videotaped assembling a TV cart for an undefined audience who would later view the videotape. In our research, we are interested in how two participants interact and coordinate with one another under different conditions.

While one might predict that not being allowed to talk with one another would be a disadvantage, resulting in more time to achieve understanding, we predict that the gesture-only and speech + gesture conditions should be equivalent in timing, while the speech-only condition should take significantly longer.

Methods

Pairs of students worked together as one participant, the director, instructed the other participant, the builder, how to create duplicate models of Lego blocks. Their goal was for the builder to create identical models as efficiently as possible based on the director's instructions. Each pair had a practice trial to orient them to the task and then constructed nine other models. The models used were the same as in Clark & Krych (2004) and Krych & Clark (1997).

Thirty-nine subject pairs participated in one of three separate conditions. Each condition consisted of 13 subject pairs who were all undergraduate students. In one condition, the director could see the builder's workspace and they could converse normally using both speech and gesture as they wished (speech + gesture). In a second condition, the builder's workspace was not visible to the director, so participants could only communicate with speech (speech-only). A third group of subject pairs participated in a gesture-only condition in which the workspace was visible to the director, but the subjects could not use any words at all. They could communicate only by gesturing to one another and pointing to objects.

Results and Discussion

As predicted, there was a large difference in the average amount of time to complete each model, $F(2, 36) = 27.64$, $p < .001$). The speech-only condition took much longer-- 181 seconds compared to 94.5 seconds in the Speech + Gesture condition and 112 seconds in the Gesture-only condition. The latter two conditions were statistically equivalent to one another. This pattern held true even if the practice trial was included. Thus, participants who were restricted to gesturing were not at a disadvantage compared to participants who could speak and gesture. These results appear consistent with the findings of Lozano and Tversky's non-interactive study (2003) that language and gesture can supplement as well as complement each other.

In the future, we plan to focus on the process of how people ground information in the gesture-only condition. We suggest that studying the process of grounding when people may only gesture to one another will shed further information on the underlying processes involved in achieving understanding in face-to-face conversation.

Acknowledgments

We would like to acknowledge Montclair State University for funding some of this research. We also thank Joanna Musial and Cecilia Sullivan for their dedicated work.

References

Clark, H. (1996). *Using Language.* Cambridge: Cambridge University Press.

Clark, H. & Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language,* 50(1), 62-81.

Krych, M., & Clark, H. (1997). Coordinating Hands, Eyes, and Voice. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (p. 962). Hillsdale, NJ: Erlbaum.

Lozano, S. & Tversky, B. (2003). Demonstrations: Clues to Effective Animations. Interactive Graphic Communication, London.

Adaptation Effects on Word Recognition Times: Evidence for Perceptual Representations

Christopher A. Kurby (ckurby@niu.edu)
Katja Wiemer-Hastings (Katja@niu.edu)
Department of Psychology, Northern Illinois University
DeKalb, IL 60115 USA

Abstract

We present evidence for the involvement of perceptual feature detector cells in the initial stages of processing lexically activated concept knowledge. Participants were slower to respond to names for oriented objects after being adapted to gratings of matching orientation, relative to opposite orientation. The data are consistent with the view that knowledge is perceptually based at a fundamental level, and inconsistent with alternative views that perceptual representations are generated at a later stage based on amodal representations.

Perceptual Representation: Adaptation Effects

The view that knowledge representations have a perceptual basis has been supported in a variety of methodological approaches. However, many of these studies may provide limited evidence for the importance of perceptual processing. Perceptually based aspects of knowledge may be generated in a later phase based on an initial, perhaps perception-unrelated representation, or they may be the knowledge that gets activated initially. Influential theories in this domain (e.g., Barsalou, 1999) assume the latter, and need to be tested accordingly.

With the goal to test whether perceptual representations are the foundation of knowledge processing as opposed to a by-product, we examined participants' response times to names for objects that had standard, vertical or horizontal orientations after exposing them to grating patterns that were in the same or opposing orientation as the object. When processing such grating patterns, specialized orientation detector neurons in the visual cortex become activated (Hubel & Wiesel, 1959). After certain exposure times, these neurons adapt, leading to lower sensitivity. We predicted that if access to a perceptual representation is by necessity part of accessing object concepts, response times to object names should be slowed down after adaptation of neurons that process critical object primitives (Treisman, 1986). Specifically, adaptation to matching orientations should slow responses relative to adaptation to the opposite orientation. In contrast, response times should be equal (or faster via priming) for matching grating patterns if such cells are not involved in initial processing.

Method

Twelve graduate students who volunteered to participate in the experiment viewed grating patterns for time durations that have been shown to create and maintain adaptation of the feature detectors for vertical and horizontal orientations. Subsequent to viewing patterns, they performed a lexical decision task on names for vertical vs. horizontal objects.

A set of items with standard vertical or horizontal orientations was rated with respect to variables that may influence perceptual processing, such as their height to width ratios and view invariance. These ratings served as covariates in the analyses.

Results & Discussion

Adaptation effects were observed for items with strong view invariance and high ratios. A 2 (grating orientation: vertical vs. horizontal) x 2 (word orientation: vertical vs. horizontal) repeated measures ANOVA revealed an interaction of grating orientation and word orientation on response time, which was slower when grating and object orientation matched (Table 1).

Table 1: Response Times per Syllable (in msec.)

	Grating	
Object	Vertical	Horizontal
Vertical	463.40	326.30
Horizontal	294.58	435.74

Conclusions

Our data suggest that perceptual representations are an integral part of knowledge activated during the initial processing of words for oriented objects. These data are consistent with the assumption that conceptual processing activates brain regions that are involved in processing of the actual related percepts. This finding lends stronger support to theories of perceptually based representations than other, related studies because the adaptation effects on response times are inconsistent with the view that perceptual representations are activated at a later stage and are merely by-products of a potentially amodal representation.

References

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577-660.

Hubel, D.H., & Wiesel, T.N. (1959). Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology, 148,* 3574-591.

Treisman, A. (1986). Features and objects in visual processing. *Scientific American, 255,* 114-125.

What drives learning by classification?

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, Binghamton University (SUNY)
P.O Box 6000, Binghamton, NY,13905 USA

Elizabeth Gonzalez (bj90475@binghamton.edu)
Department of Psychology, Binghamton University (SUNY),
P.O Box 6000, Binghamton, NY,13905 USA

The classification learning paradigm has been the dominant technique across decades for the study of categorization (Murphy, 2002). The learning procedure consists of passes through a set of training items presented one at a time in random order. On each trial, an example is displayed with a forced-choice classification question. Responding elicits corrective feedback followed by an inter-stimulus interval.

The goal of this research is to look inside the classification trial in order to identify the locus of learning. The roles of feedback and intentionality have been investigated elsewhere (e.g., Love, 2002). The factors addressed here are: 1) unlimited access to the stimulus during responding; 2) availability of the stimulus during feedback; and 3) generation of a classification response.

In the Init+During condition, the classification trial is executed in standard fashion except the stimulus is removed during feedback. This allows us to evaluate the importance of coordinated evaluation of the stimulus and the correct label at the end of the learning trial. In the Init+Final condition, each stimulus is presented for 3s and then removed when the classification question appears. After the response, the stimulus re-appears along with corrective feedback to allow coordinated evaluation. Speeded classification has been a topic of past research (e.g., Nosofsky & Palmeri, 1997), but the present question is about limiting access to the perceptual stimulus without any requirement of fast responding. In the Init-Only condition, we test the combined effect of limited initial access plus absence of the stimulus during feedback.

Finally, in the No-Response condition, each trial consists of presentation of the stimulus and its correct category. The learner observes the association and presses a button to continue. This allows us to address the common intuition that generating a response and evaluating success plays a critical role in classification learning. Many models of category learning operate on the basis of error correction between an output and a feedback signal.

In order to compare these conditions, three category prototypes were designed using 4x4 grids of half gray and half white squares. The training set consisted of 16 examples of each category generated by distorting the prototype with exactly two squares of reversed color. Participants (n=199) were randomly assigned to one of five conditions. The study phase consisted of a maximum of 192 trials. After every twelve trials, performance was evaluated against a 90% criterion for stopping learning. The test phase consisted of standard classification of all items.

Ease of learning was measured by percentage of participants reaching criterion and performance on the test phase common to all conditions. Impaired performance in any experimental condition relative to the control group would highlight a critical aspect of classification learning.

Approximately half of all learners reached criterion (11/12 correct). In the test phase, participants were well above chance (33%), though quite far from ceiling.

Table 1: Learning performance across conditions

Condition	% Ss reach criterion	% correct at test
Standard	54	69
Initial Only	40	63
Init+During	47	69
Init+Final	65	72
NoResponse	--	72

To our considerable surprise, none of the experimental conditions differed reliably from the control group on either measure. The only significant difference was between the Initial-Only and Initial+Final groups. This appears to be attributable to a slight disadvantage in the Init-Only condition combined with a slight advantage in the Init+Final. We draw the preliminary conclusion that none of the elements considered, i.e., extended evaluation of the stimulus during responding, coordinated evaluation during feedback, nor response generation can be considered critical components of classification learning. Learners are able to adapt fairly seamlessly in each case. These data suggest that as long as the learning trial includes the item and its label, the rest is more or less bells and whistles.

Acknowledgments

We thank the members of the LaRC Laboratory.

References

Love, B. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829-835.

Murphy, G. (2002). *The Big Book of Concepts*. MIT Press: Cambridge, MA.

Nosoksy, R. & Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Representations in Simple Recurrent Networks Which are Always Compositional

David Landy (dlandy@indiana.edu)
Departments of Computer Science and Cognitive Science, Indiana University
107 S. Indiana Ave., Bloomington, IN 47405-7000

In classical cognitive models, representations of inputs are deliberately built into the operational structure by a model's designers. Network systems by contrast usually automatically construct responses following some generic learning scheme, and consequently lack overt representations altogether. Instead, the system's representations are read off the system according to a chosen analytical methodology. The performance of such models is therefore independent of how their representations are labeled.

Simple recurrent networks (SRNs) are among the most successful network models of cognition (Elman, 1990, 1995). These networks are often taken to represent inputs in the values of their hidden layer nodes, which can be analyzed using principal component analysis or hierarchical clustering. Under this interpretation, representations in networks are context-sensitive, static, and non-compositional. Significantly different properties result from taking as the representation of a sequence the function which that sequence causes the network to compute.

Consider a typical SRN with input weights W_{in}, output weights W_{out}, and recurrent connections in the hidden layer with weight matrix C, and call the vector of weights in the hidden layer H. Let S denote the closure of the set of legal inputs to the network under concatenation, so that S contains all legal sequences (and also an empty input, ε). Call the set of possible output vectors O, and call the function which maps input sequences to output vectors $i:S \rightarrow O$, so that $i(s)=o$ exactly when o is the output resulting from running sequence s through the network. Consider the following function:

$$r'(s,h) = \begin{cases} h & \text{if } s = \varepsilon \\ r'(u, Sigmoid(C \cdot h + W_{in} \cdot t) & \text{if } s = t.u \end{cases}$$

r' can be interpreted as the function which computes, for an initial value on the hidden layer, h, the value on the hidden layer which results after processing input sequence s. Define the family of functions which results from currying r' over s: $r_s(\vec{h}) = r'(s,\vec{h})$, $r_s \in R$. Then R is the representation scheme of the SRN. $i(s)$ can be easily reconstructed from r_s.

A straightforward homomorphism can be constructed between concatenation over S and function composition in R, making the representations of any SRN classically compositional, regardless of the prior training of the network (see Fodor & Lepore, 2002; Zadrozny, 1994). This analysis also reveals a limited form of inherent systematicity, in that the same representation function, and hence the same causal mechanism, is employed in processing a particular lexeme or sequence regardless of the context in which it appears (see Davies, 1991).

If the computation specified by r_s picks out a type of representation, then any particular application of that function can be taken to be a token. The computation performed is independent of its context, but the specific hidden layer value which results will not be. Therefore, tokens of computations can be picked out by specifying the input/output pair (where both input and output are hidden layer values) which that application involved. The method of hierarchical clustering which is so useful in analyzing hidden layer values can then be performed on this pair, and so this technique can be applied essentially unchanged. Additionally, the extra information stored in the source values allows the method to be applied to sequences as well as single inputs.

Since these representations are the system's disposition to respond to a particular lexeme, rather than the residue of state information which results from that response, these representations are active processes rather than static data structures. Since the important

Therefore, representations capture all of the knowledge which is involved in generating the internal state of the network.

Because the representation scheme given here appropriately encapsulates an SRN's knowledge and presents representations as dynamic processes rather than static structures, it is intuitively appealing as a model for how SRNs represent. Inasmuch as it is appealing, SRNs represent compound phrases compositionally and context-independently, which implies that these properties may not account for some of the interesting properties with which they have been credited (Fodor & Lepore, 2002).

References

Davies, P. 1991. Concepts, connectionism, and the language of thought. In William Ramsey, Stephen Stich, David Rumelhart, eds., *Philosophy and Connectionist Theory*. Lawrence Erlbaum.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14:179-211.

Elman, J. L. 1995. Language as a dynamical system. In Port, R. F. and Van Gelder, T., eds., *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, 195-223.

Fodor, J., and Lepore, E. 2002. Why meaning (probably) isn't conceptual role. In *The Compositionality Papers*. Oxford University Press.

Zadrozny, W. 1994. From compositional to systemantic semantics. *Linguistics and Philosophy* 17:329-342.

Categories among Relations

Levi B. Larkey (larkey@mail.utexas.edu)
Lisa R. Narvaez (grimmlr@mail.utexas.edu)
Arthur B. Markman (markman@psy.utexas.edu)
Department of Psychology, University of Texas at Austin
1 University Station A8000, Austin, TX 78712 USA

Introduction

Much research has been devoted to the way that categories are represented. Two of the most influential theories argue that concepts are represented by prototypes (e.g., Rosch & Mervis, 1975) or exemplars (e.g., Medin & Schaffer, 1978). While these views are typically analyzed in terms of their differences, they share the assumption that category coherence is a function of intrinsic features of category members.

In contrast, recent theories have proposed role-governed categories (Markman & Stilwell, 2001; Gentner & Kurtz, in press). Members of role-governed categories cohere because they fill similar roles within a relational structure. Role-governed categories differ from feature-based categories in that membership is determined according to external relations between categories rather than intrinsic features.

One reason for the prevalence of feature-based categories in the literature is that typical laboratory tasks are well matched with this type of category. The majority of categorization experiments employ artificial categories that are composed of entities isolated from any relational context. A critical problem with this approach is that it may not capture the essence of categorization outside of the laboratory; natural categories occur in relational contexts. We present an experiment that addresses whether artificial role-governed categories can be learned and manipulated in the lab.

Method

The experiment consisted of a learning phase and two transfer phases. Each trial, subjects were shown two objects that varied in size (big or small), color (blue or orange), shape (circle or square), and relational role (object that pushed or was pushed). After one of the objects was briefly highlighted, one of the objects moved across the screen and pushed the other object. In all phases, the task was to classify the highlighted object as type F or G. Subjects were given feedback in the learning phase, but not in the transfer phases. The learning phase lasted until subjects correctly classified 10 consecutive objects or 50 trials had been given. Each transfer phase was 10 trials. Taken together, the transfer phases were constructed to control for the presence of non-relational spatial and temporal cues.

There were two between-subjects conditions. In the first condition, categories were defined by a Shepard Type I rule (i.e., a unidimensional rule) involving a relational role and a redundant Type I rule involving a feature dimension (cf. Shepard, Hovland, & Jenkins, 1961). For example, an object could be classified as type F based on being a pusher or based on being blue. In the second condition, categories were defined by a Type I rule involving a relational role and a redundant Type II rule involving two feature dimensions. In the transfer phases of both conditions, the Type I rule involving the role was reversed such that subjects using the role would reverse their classifications while subjects using the features would not. In other words, the rule involving the role and the rule involving the features were deconfounded.

Results and discussion

In the Type I condition, 20 out of 39 subjects (51%) used the role, 11 (28%) used the feature, and 8 (21%) did not meet the learning criterion or could not be clearly classified. In the Type II condition, 28 out of 41 subjects (68%) used the role, 4 (10%) used the features, and 9 (22%) did not meet the learning criterion. In addition, subjects were more likely to use relational roles when the rule involving the features was more complex (51% for Type I versus 68% for Type II).

These data suggest that artificial role-governed categories can be learned and manipulated in the lab. The majority of subjects (51% in the first condition and 68% in the second condition) used relational roles to classify the objects. This manipulation sets the stage for a systematic set of laboratory studies to explore the acquisition of feature-based and role-governed categories.

References

Gentner, D., & Kurtz, K. (in press). Learning and using relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman & P. W. Wolff (Eds.), *Categorization inside and outside the lab*. Washington, DC: APA.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence, 13*(4), 329-358.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13, Whole No. 517).

Understanding and Modifying Procedural Versus Object-Oriented Programs: Where Does Domain Knowledge Help More?

Thomas LaToza (Student Member CSS) and Alex Kirlik (Member CSS)
Department of Psychology and the Beckman Institute
University of Illinois at Urbana-Champaign
<latoza, kirlik>@uiuc.edu

Software bugs are an ever-increasing societal problem as computers become more prevalent in our homes and workplaces, and programs grow in size and complexity. In response, many in the software engineering community have advocated a shift in the way programs are written: away from procedural (plan-like) code (e.g., Pascal, Fortran) and toward more declarative (map-like) Object-Oriented (OO) code (e.g., C++, Java). A central claim is that more declarative, map-like code allows developers to better deploy their knowledge of a problem domain (what the code is 'about') than does procedural code. The latter typically consists of a compact plan for solving a problem (such as a recipe or driving directions), rather than a more declarative description of domain entities and their relationships (OO). To test this claim, we conducted an experiment requiring experienced programmers presented with procedural and OO code *isomorphs* of the same algorithm to perform code modification tasks, and crossed this manipulation with whether or not participants were primed on the domain knowledge of the actual problem solved by the algorithm (scoring ten-pin bowling). In contrast to the claims of the OO community, our findings revealed that priming domain knowledge helped those modifying procedural, rather than OO, code: The procedural group whose knowledge of how bowling is scored was primed created significantly fewer bugs than the non-primed procedural group, while priming the OO group led to perhaps even slightly more bugs. This finding suggests that, when trying to modify or "tweak" problem solutions, knowledge of the problem domain is more important when trying to amend existing procedural solutions (e.g. a route given as a list of directions) than when trying to amend more descriptive, declarative solutions (e.g. a map to a destination). Implications for both software engineering and cognitive science will be presented.

Content with Causal Complexity

Daniel Hsi-wen Liu (hwliu@pu.edu.tw)
Division of Humanities, Providence University
200 Chung-Chi Rd, Shalu, Taichung County 433, Taiwan

Internal representation and causal relations have been generally taken as polemically contrasted in the cognitive architectures. Representation bears content; whereas, a causal line does not, which at best can be seen as its (the representation's) implementation. In discussing the nature of a certain type of low-level processes, such as those in the Watt Governor, that they are 'mere transition of forces' is taken as a reason to deny the existence of an immanent role of representation (Haselager *et al*. 2003). To think of it more sympathetically, causal complexity and content together can be seen as two separate strands to be reconciled (Wheeler and Clark 1999). The possibility of *content with causal complexity* has rarely been considered, as it is not easy to figure out a substantial sense of content bound intrinsically with the complexity of a causal line.

As an attempt, Bechtel (1999) argues that there is a legitimate sense of representation immanent in the control of the Watt Governor. The reason of its existence is grounded on the isomorphism between representation and the machine states. Such a reason, however, is challenged by Haselager *et al*. (2003) that it is risky to incur overwhelming representations. A problem facing a loose account of representation is "how a system can be shown not to be representational" (Haselager *et al*. 2003, p. 18). This paper will present a sense of content with causal complexity but avoid the aforementioned problem of overwhelming representations.

It is easy to understand that the content of intentional states *arises from* machinery with causal complexity, yet this is not the attempt of this paper. Alternatively, in this paper I will make clear the existence of a type of content that is immanent in the dynamic states of certain complex cognitive processes, with those states being possibly scattering across intentional states or spreading across a line of cognitive control. Such a special type of content can be named *dynamic content*.

Sub-symbolic Features with Complex Connections

Consider two examples of dynamic content. Firstly, units of the connectionist network represent certain sub-symbolic features, according to Smolensky (1988), which together represent intuitive cognitive content. Consider the role of sub-symbolic features in the representation of those units. Those features interact mutually under the connecting control of the connectionist network, with various weights in different between-units connections and certain algorithms controlling the activation of units. When the information transformation is in process, the network has not yet presented clear intentional (possibly conscious) content, but those sub-symbolic features really undergo transformation. That network, meanwhile, is by no means empty (though possibly unconscious) in its maintenance of cognitive features. The envisaged content, unlike the higher-level content manifested in the output layer, does not pertain to a cognitive state under a functional analysis. Rather, its nature causes it exist *in process*.

Motor Control

Secondly, motor control is the paradigmatic example of *dynamic content*. Dynamic content is conceived of as consisting of standing-ins of a system that serve as *guidance of that system's behavioral control*, and in turn as a way to supply the maintenance (with its causal power) of that system's performance in its environment with a certain degree of flexibility. Those standing-ins work systematically under a scheme which is non-isomorphic but can engage external conditions. Those standing-ins qualify the system as content-bearing because they enhance its capabilities of performance. The role of standing-ins in the constitution of content is to provide mediating entities for the systematic use in the course of behavioral generation.

A system's dynamic content *qua* content, as we can see, rests on the amenability of its behavioral guidance in the light of enhancing its performance. Because the amenability is a capacity of fine-tuning the causal connections of a system's complex behavioral control, dynamic content qualifies as content on grounds of its potentiality of amending complex causal connections.

Acknowledgments

This research is supported by National Science Council, Taiwan, under grant NSC 92-2411-H-126-003.

References

Bechtel, W. (1998). Representations and cognitive explanations: assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22*, 295-318.

Beer, R. (2000). Dynamical approaches to cognitive science. *Trends in cognitive science*, *4*, 91-99.

Haselager, P., De Groot, A., & van Rappard, H. (2003). Representationalism vs. anti-representationalism. *Philosophical Psychology*, *16*, 5-23.

Smolensky, P. (1988). On the proper treatment of connectionism, *Behavioral and Brain Sciences*, *11*, 1-74.

Wheeler, M. & Clark, A. (1999). Genic representation: reconciling content and causal complexity. *British Journal for the Philosophy of Science*, *50*, 103-135.

The Hippocampus: Where a Cognitive Model meets Cognitive Neuroscience

Bradley C. Love (love@psy.utexas.edu) and Todd M. Gureckis (gureckis@love.psy.utexas.edu)
Department of Psychology, The University of Texas at Austin; 1 University Station A8000; Austin, TX 78712 USA

Models and Cognitive Neuroscience

The goal of the present work is explore possible mappings between an existing model from cognitive psychology and functional brain regions. There are numerous possible mappings between these somewhat different levels of analysis. For the the Supervised and Unsupervised STratified Incremental Network (SUSTAIN; Love, Medin, & Gureckis, 2004; Sakamoto & Love, in press) model, the mapping is straightforward: aspects of the model appear to map onto functional structures in the brain.

SUSTAIN holds that humans represent category information in terms of natural bundles of information, referred to as clusters. For example, knowledge of mammals might be represented by several clusters (e.g., primates, four-legged mammals, whales, bats, etc.). SUSTAIN posits that learners form new clusters in response to surprising events, such as when a child is first told that a whale is a mammal and not a fish.

In this poster, we will focus on SUSTAIN's cluster formation process. Our hypothesis is that a healthy and intact hippocampus is necessary for forming new clusters to support cortical learning in the temporal lobe (cf., Gluck & Myers, 1993). Forming new clusters can be seen as constructing conjunctive codes. A wide variety of tasks rely on the formation of conjunctive codes such as episodic memory (a conjunction of item and context), sequence memory (item and position), list discrimination (item and list), and item relations (item and item). All of these tasks appear to rely heavily on the hippocampus (see Brown and Aggleton, 2001, for a review). Assuming reduced ability to form new clusters, SUSTAIN has been able to model developmental trends in infant learning (hippocampus not fully developed) and performance by amnesiacs with hippocampal lesions (Gureckis & Love, 2003).

Rules and Exceptions: An Aging Study

Our account of hippocampal function predicts that normal aging will disproportionately affect performance for exception items in rule-plus-exception classification studies. To master an exception, a cluster must be recruited to encode it, despite the fact that similar clusters or conjunctive codes likely already exist in memory. As we age, an accumulation of cortisol released in response to stressful events differentially leads to atrophy and reduceas the functioning of the hippocampus (Lupien et al., 1998). In the study design, three items from category A and three items from category B followed a simple rule (e.g., if large, then category A. if small, then category B.). The exception items ran counter to these rules. SUSTAIN predicts that older adults will form one cluster for category A and B, leading to increasing rule application and insensitivity to old vs. novel rule-following items with increasing age. In contrast, SUSTAIN predicts younger adults will recruit one cluster for each exception, storing them apart from rule-following items, which leads to predictions counter to those of the older population.

Human Results

Thirty-seven University of Texas undergraduates and thirty-seven healthy older adults (51-84 years-old, mean=67.9) recruited from the Austin VA outpatient clinic participated in the study. All of SUSTAIN'S predictions held. Only a subset of results are reported here. In the learning phase, item type (rule vs. exception) and population interacted such that the younger population exhibited a smaller difference in accuracy (.27 vs. .61) for exception and rule-following items than did the older population, $F(1, 72) = 35.39$, MSe = 1.09, $p < .001$. For the older population, performance on rule-following and exception items for the learning and test phase negatively correlated (r=-.38 and -.72, respectively), whereas these correlations were positive for the younger population (r=.52 and .49, respectively). In transfer, subjects from the older population made rule consistent responses to studied rule-following items and all novel items at about the same rate, .70 vs. .69, $t < 1$, whereas the younger population applied the rule more frequently (.88 vs. .77) to the studied examples, $t(36) = 4.99$, $p < .001$.

References

Brown, M. P., & Aggleton, J. P. (2001). Recognition memory: What are the roles of perirhinal cortex and hippocampus? *Nature Neuroscience*, 2, 51-61.

Gluck, M. A., & Myers, C. (1993). Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus*, 3, 491-516.

Gureckis, T.M., & Love, B.C. (2004). Common Mechanisms in Infant and Adult Category Learning. *Infancy*, 5, 173-198.

Love, B.C., Medin, D.L, & Gureckis, T.M (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111, 309-332.

Lupien, S.J., DeLeon, M, DeSanti S, Convit A, Tarshish, C., Nair, NPV, McEwen, B.S., Hauger, R.L., & Meaney, M. (1998). Longitudinal increase in cortisol during human aging predicts hippocampal atrophy and memory deficits. *Nature Neuroscience*, 1, 69-73.

Sakamoto, Y., & Love, B. C. (in press) Schematic Influences on Category Learning and Recognition Memory. *Journal of Experimental Psychology: General*.

The 2004 CogSci proceedings publication

Lozano, S. C., Martin, B. A., and Tversky, B. Perspective Matters for Action Learning.
Table of contents: 1590 Actual page: 1618

has been retracted.

All authors retract this article. Bridgette Martin Hard and Barbara Tversky believe that the research results cannot be relied upon; Sandra C. Lozano takes full responsibility for the need to retract this article.

Thinking With and Without Words: A New Model of Cognition, Language and Mind

Jack Lynch (jlynch2@comcast.net)
200 Lake View Avenue
Cambridge, MA 02138 USA

Viewing the mind as a sophisticated processor of analog signals rather than as a processor of symbolic information has led me to conclude that human intelligence emerged from nonhuman primate intelligence as a consequence, and only as a consequence, of the evolution of natural language. A nonhuman animal's primary cognition processes endow it with a sophisticated means of extracting useful features of its environment from the signals from its sensors, such as the eyes and ears, that enable, some species, flexible survival-enhancing and even innovative behavior.

The powerful primary cognitive system was greatly enhanced by the hominid and human species evolving natural syntactic speech—a two-tiered system of auditory communication. Meaningless sounds combine in restricted ways to produce words and words combine in restricted ways to form sentences. The secondary cognition system, also called the human reason system, not only allows the expression of internal representations of the primary cognitive system, it has created a new way of thinking that has resulted in a entirely new culture on the planet that has produced Shakespearian sonnets, jumbo jet aircraft, international corporations and a cure for infectious diseases.

This new model of cognition is presented in two books, *I am not a machine—Book I: Thinking without words* (Lynch, 2004) and *I am not a machine—Book II: Thinking with words* (Lynch, expected late in 2004) and is introduced on the website, NOTaMACHINE.org. Under the proposed model of cognition, there is no inner language of thought (mentalese), no formal computational rules and procedures underlying thought and no inner "engine of reason" that controls or guides our behavior or utterances.

This paper starts with a description of the primary cognition system upon which the new conceptualization of human intelligence rests. The first of four parts of the primary cognition model is a neural network pattern classification system (association). In addition, primates have a second-order pattern classification system (also called relational matching or tertiary cognition). This second-order process cannot be implemented by a feedforward neural network because the similarities to be noted are not in the patterns themselves but in the higher-order relationships in the patterns.

The second part of the model is a mental representation system that can be best understood by introducing a new term, *cogject* that labels how minds from hamsters to primates mentally represent physical objects, actions and events. A critical property of *cogjects* is that they are singular and affirmative—a dog cannot represent, *The cat is not on the mat*. Unlike the popular inner language of thought hypothesis, these internal representations are not more precise than natural language—they are much less precise than natural language. This primary representation system has evolved the capability to chunk perceptions and representations in appropriately sized "bins," so that a creature can gather statistical information of its world via neural network pattern classification processes.

The third part of the model includes the well-documented specialized core knowledge systems. Also known as a cognitive toolkit, this system includes know-how about objects, navigation skills and a number sense.

The fourth part of the primary cognition model is an action planning and evaluation system. An animal's next immediate motor control movements are planned by sequencing representations of action control signals, that is, *cogjects*, in a buffer. Evaluation of a planned action is by an additional pattern classification process that operates on the *cogject* contents of the action buffer. The result of the evaluation process is either a go-ahead or an abort response that is simply based on pattern classification. If an action is aborted, another action can be planned and perhaps also aborted, leaving a human observer to conjecture that the momentarily inactive animal is "thinking."

Human cognition is modeled by three systems, the primary cognition system just outlined, a language cognition system, and a secondary cognition or human reason system. Language cognition includes those cognitive processes that support natural language. I endorse the cognitive linguists' theory based on patterns and spaces rather than rules, procedures, symbols and formal systems.

The third human cognitive system called, "human reason," may be the least intuitive. Based on natural language, it establishes a new world of complexity by allowing us to create, name, describe, explain and communicate intricate concepts, procedures and theories. What can the human mind do with words? My answer is that we can tell stories, period. Human logic and truth are Greek myths that need comprehensive revision. Even the "truth" of science is a value we bestow on a story that has been corrected many times by many people who look for consistency with other stories (theoretical import) and consistency with their observations of the world (empirical import). Truth is based upon group consensus and is always subject to change and revision.

Human intelligence is built upon animal intelligence and natural language and not on a separate engine of reason.

References

Lynch, J. (2004). *I am not a machine—Book I: Thinking without words*. Imperial Beach, CA: Aventine.

Lynch, J. (expected late 2004). *I am not a machine—Book II: Thinking with words.* Imperial Beach, CA: Aventine.

A Model of Novel Compound Production

Dermot Lynott (dermot.lynott@ucd.ie)
&
Mark T. Keane (mark.keane@ucd.ie)
Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

PUNC: Producing and Understanding Novel Combinations

The PUNC model (Producing and Understanding Novel Combinations) is the first model to capture both production and comprehension aspects of conceptual combination within the same theoretical framework. The comprehension side of PUNC has been detailed elsewhere (see Lynott, Tagalakis & Keane, 2004), so here we give a brief overview of the production side of the model.

Lynott (2004) has proposed the Integrated Production and Comprehension (IPAC) Theory of conceptual combination. The IPAC theory seeks to describe the two sides of the conceptual combination coin, comprehension and production, within the same theoretical framework. Lynott draws together several factors that have been shown to influence both sides of conceptual combination in similar ways. Central to this view are the factors of Diagnosticity, Informativeness and Plausibility (inspired by earlier work by Costello & Keane, 2000). PUNC is a computational implementation of this theory, with the model taking as input short descriptions of novel entities (e.g., "a beetle that eats cacti" or "a prickly beetle") and, using the aforementioned factors, outputs candidate labels together with an overall acceptability score for each label. Below, we provide a brief description of the stages the model undergoes, from taking in an entity description to outputting candidate labels and assigning acceptability scores.

A description is input to PUNC (e.g., a beetle that eats cacti; a beetle that is prickly). Concepts are activated either by being explicitly mentioned in the description or through the description containing a feature that is diagnostic of another concept. So, the "is prickly" feature would also activate the concept *cactus*. Each of these concepts forms part of a set of candidate modifiers for the head concept (e.g., *beetle*). The individual concepts' features are then activated, prioritised by their diagnosticity. For example, for the concept *cactus* "is prickly" is more diagnostic than "can conserve water" and so has greater activation. These features are used to determine whether a modifier is informative with respect to the head concept as PUNC considers each candidate modifier in turn and whether it can form part of a valid, acceptable label for the entity being described. For example, *cactus beetle* is output as a valid label for the described entity since "is prickly" is a highly diagnostic feature of the concept *cactus* and this feature is also informative with respect to the head concept *beetle* (i.e., beetles are not by default prickly). Labels are considered informative if they incorporate some new information relative to the head concept. In this way, the informativeness of a label is a binary affair. A label such as "wood tree" meaning "a tree made from wood" would not be considered informative and so would be rejected as a possible label. As such, informativeness is a primary pragmatic constraint within the theory and model.

PUNC assigns overall acceptability scores to each of the candidate labels, based on the relative diagnosticity of the features used, the informativeness of the label and the plausibility of the relation that links the two concepts in the compound. For example, using a highly diagnostic feature of *cactus* to form a label contributes positively to the acceptability of a compound; if a less diagnostic feature had activated *cactus* the resultant score would be reduced.

Finally, the plausibility of the relation linking the head and modifier concepts contributes to the acceptability of the label. For the description "a beetle that eats cacti", the label *cactus beetle* would be considered highly plausible since there is a reciprocal "eats" relation between the concepts – beetles can eat things, and cacti, as vegetative matter can be eaten. On the other hand, *brick beetle* as a label for "a beetle that eats bricks" is considered less plausible as bricks are not usually considered edible.

Lynott (2004) has found that by using such pragmatic constraints in an integrated fashion PUNC not only reflects people's choice of label for novel entities, but its overall acceptability scores correlate highly with people's ratings of how good specific compounds are as labels for entity descriptions.

Acknowledgments

The work presented here has been funded through a grant from the Irish Research Council for Science, Engineering and Technology under the Embark Initiative.

References

Costello, F. J. & Keane, M. T. (2000). Efficient creativity: Constraints on conceptual combination. *Cognitive Science*, *24*, 299-349.

Lynott, D., Tagalakis, G, & Keane, M. T. (2004). Conceptual combination with PUNC. To appear in *Artificial Intelligence Review*.

Lynott, D. (2004). *Comprehension and Production in Conceptual Combination*. Doctoral dissertation, University College Dublin, Ireland.

Integrating Planning, Creativity and Exploration in Motivational Agents

Luís Macedo (lmacedo@isec.pt)
Department of Informatics and Systems Engineering, Engineering Institute, Polytechnic Institute of Coimbra, Quinta da Nora
3030-199 Coimbra, Portugal
Centre for Informatics and Systems of the University of Coimbra, Department of Informatics, Polo II
3030 Coimbra, Portugal

Abstract

Although, planning and exploration of unknown environments has been previously addressed in multi-agent environments, we believe that in addition to these activities, agents may also benefit from exhibiting creativity so that they are able to imagine or invent new things (objects) that may be helpful or simply pleasant for the agents that inhabit the environment. Psychological and neuroscience research (e.g.: Damásio, 1994) over the past decades suggests that motivations (emotions, drives and other motivations) play a critical role in these activities that involve decision-making, and action, by influencing a variety of cognitive processes (e.g., attention, perception, planning, etc.).

We have developed a multi-agent environment (Macedo & Cardoso, 2004) in which, in addition to inanimate agents (objects), there are two main kinds of animate agents interacting in a simple way: the creators, whose main function is to create things (objects, events), and the explorers whose goal is to explore the environment, analyzing, studding and evaluating it. In spite of this classification, there are agents that may exhibit the two activities, exploration and creation. In addition to these two activities, animate agents are able to generate plans. Planning plays a central role in the reasoning/decision-making by supporting the other two activities: exploration and creativity. Actually, in our approach, creativity and exploration involve planning: when exploring the environment an agent has to plan a sequence of actions required to visit an unknown region or entity; when creating, an agent has to plan the sequence of actions required to come up with an original and valuable object.

The architecture of an agent includes the following modules: memory (for entities, plans, and maps of the environment), goals/intentions, desires, motivations (emotions, drives and other motivations), and reasoning/decision-making. The planner is the core of the deliberative reasoning/decision-making module. The agent uses a planner that combines the technique of decision-theoretic planning with the methology of HTN planning in order to deal with uncertain, dynamic large-scale real-world domains. Unlike in regular HTN planning, the planner can generate plans in domains where there is no complete domain theory by using cases of previously successful plans instead of methods for task decomposition. It generates a variant of a HTN - a kind of AND/OR tree of probabilistic conditional tasks - that expresses all the possible ways to decompose an initial task network. The expected utility of alternative plans is computed beforehand at the time of building the HTN and it is based on the expected positive and negative feelings that the agent feels if the plan is executed. Plans that are expected to elicit more positive feelings (happiness, surprise, etc.) and less negative feelings (e.g.: hunger) are assigned a higher expected utility.

When performing exploration, the aim of an agent is twofold: (i) acquisition of maps of the environment – metric maps – to be stored in memory and where the cells occupied by the entities that populate that environment are represented; (ii) construction of models of those entities. Exploration may be performed by single or multiple agents. Each agent autonomously generates goals for visiting unknown entities or regions of the environment (goals of kind *visitEntity* or *visitLoc*) and builds a HTN plan for each one. Goals and plans that are expected to cause more positive feelings and less negative feelings are preferred. Thus, each agent performs directed exploration using an action selection method based on the maximization of the intensity of positive feelings and minimization of negative ones. Relevant motivations for directing exploration are for instance curiosity, surprise and hunger. The exploration strategy for multiple agents relies on considering a team leader that, based on the information provided to it by the members of the team as they perform their single exploration, builds a joint metric map, a joint episodic memory and a joint plan in order to be shared by all the members of the team.

When performing creativity, an agent generates goals for the creation of novel, original and valuable entities (goals of kind *createObj*) and builds a plan for each one. Like in exploration, goals and plans that are expected to cause more positive feelings and less negative feelings are preferred. Motivations such as surprise and curiosity that capture variables such as novelty or unexpectedness, respectively, are hence important for creativity.

Main References

Damásio, A. (1994). Descartes'error, Emotion Reason and the Human Brain. New York: Grosset/Putnam Books.

Macedo, L., & Cardoso, A. (2004). Exploration of Unknown Environments with Motivational Agents, Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems.: ACM.

Word Order Effects in Conceptual Combination

Phil Maguire (phil.maguire@ucd.ie)
Arthur Cater (arthur.cater@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland.

The combination of two existing words is a productive strategy used by speakers to convey new concepts and extend the limits of their vocabulary. In English compounds, the first word or modifier attaches further meaning to the second word or head, thus creating a reference to the intended concept. According to Gagné and Shoben's (1997) Competition Among Relations In Nominals (CARIN) theory, there is a fixed, relatively small taxonomy of standard relations that can be used to link the modifier and head noun concepts. According to this theory, the most available standard relation is the one most frequently used to interpret other compounds containing that same modifier. We investigated whether the alleged importance of the modifier in relation selection is due to the fact that it comes first or whether it can be attributed to the modifier's functional role. Accordingly, we conducted our study in French, a language in which the order of the nouns is the reverse of that in English.

Priming Experiment

We carried out two experiments using French noun-noun combinations which parallel a speeded sensibility study carried out by Gagné (2001). Gagné's study investigated the way in which recent exposure to a similar combination influences the processing of a subsequent combination. She found that when the prime and the target had the same head noun, there was no significant different in reaction times between cases where they shared the same relation and cases where they did not. However, when the modifier was the repeated constituent, primes that used the same relation were more effective than those that used a different relation. Gagné took this as evidence that the modifier is paramount in relation selection. We investigated whether the same effect would be apparent in a language in which the order of the constituent nouns was reversed.

No Evidence of Word Order Effects

As predicted by the CARIN theory, we found no influence of the prime's relation on reaction times when the prime and target shared the same head noun. However when the modifier was a shared constituent, reaction times were slower when the target was preceded by a combination with the same relation than when it was preceded by a combination with a different relation. Participants responded to targets following a same-relation prime 45ms quicker than they did to targets following a different-relation prime, $F_{subject}(2, 34) = 4.349$, $p < .05$; $F_{item}(2, 118) = 4.194$, $p < .05$. Hence *"ruisseau de montagne"* (mountain stream) was more effective than *"chaussures de montagne"* (mountain shoes) at priming *"glacier de montagne"* (mountain glacier) while *"sac de voyage"* (travel bag) and *"sac de cuir"* (leather bag) were equally effective at priming *"sac de sport"* (sports bag).

Table 1: Response Times for Same-Head Targets.

Prime			Target Response Time (ms)
Same Head	Same Modifier	Same Relation	
✓	×	✓	994
✓	×	×	999
×	×	NA	1153

Table 2: Response Times for Same-Modifier Targets.

Prime			Target Response Time (ms)
Same Head	Same Modifier	Same Relation	
✓	×	✓	998
✓	×	×	1043
×	×	NA	1062

Our results follow a similar pattern to those of Gagné (2001). They indicate that people's ability to select a relation that was used in a recently viewed combination is influenced by whether that combination shares the same modifier but not whether it shares the same head. Since these effects have been replicated in a language in which the order of the modifier and head are reversed, this suggests that modifiers and heads maintain the same role in the process of interpretation regardless of the order in which they are realized. Additionally, it appears as if relational information is predominantly associated with the modifier.

References

Gagné, C. L. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 236-254.

Gagné, C. L., & Shoben, E. J. (1997). The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 71-87.

Origins of Universality and Linguistic Diversity in Naming Human Gait

Barbara C. Malt (barbara.malt@lehigh.edu)
Department of Psychology, Lehigh University
17 Memorial Drive East; Bethlehem, PA 18015 USA

Mutsumi Imai (imai@sfc.keio.ac.jp)
Faculty of Environmental Information; Keio University at Shonan Fujisawa
5322 Endoh; Fujisawa, Kanagawa 252; JAPAN

Silvia Gennari (sgen@lcnl.wisc.edu)
Department of Psychology, University of Wisconsin
1202 West Johnson St., Madison, Wisconsin 53706 USA

Introduction

The search for universally shared elements of word meaning can shed light on how human languages are shaped as well as on shared aspects of human cognition: Do any such elements arise from a universal appreciation of structure that exists in the world, from universal properties of human information processing, or from universal human needs, interests, and concerns? Likewise, understanding how and why languages diverge in their word meanings can reveal how languages evolve word meanings and how cognitive processes shape this evolution.

We studied naming patterns across three languages in the domain of human locomotion (walking, running, skipping, etc.). This domain is universally experienced and highly structured, and there is an independent biomechanical description of that structure. Any universality is not likely determined at the level of the sensory apparatus (cf. the much-studied domain of color). These features allow us to ask whether a shared perception of structure in the world provides a constraint on the development of word meanings.

Portions of this domain vary in their centrality to human experience: Walking and running are universally the primary gaits, whereas hopping, skipping, etc. are more peripheral. And languages differ in how manner of movement is lexically encoded. Some typically encode manner in the verb ("She ran out of the room" [the English pattern]); others more typically encode path in the verb and manner only optionally in an adverbial phrase ("She exited the room {optional: running} [the pattern of e.g., Romance languages]. These features allow us to ask whether centrality to human experience and differences in verb lexicalization patterns across languages influence the degree of shared meaning.

We predicted that: (a) Strong universality would be found in the central portions of this domain; all languages tested would have manner verbs closely equivalent to "walk" and "run" in English. (b) Greater diversity would be found in more peripheral parts of the domain. (c) Manner verb languages would show greater linguistic differentiation of the more peripheral parts of the domain than other languages.

Experiment 1

Monolingual native speakers of English, Spanish, and Japanese named 24 video clips of a student locomoting on a treadmill that varied systematically in speed and slope from low to high. Speakers of all three languages showed strong within-language agreement on names for clips and switched from one label to a different one at exactly the same points in the stimulus continuum -- points that corresponded to biomechanical discontinuities in the movements produced.

Experiment 2

Monolingual native speakers of English, Spanish, and Japanese named 36 video clips of a student locomoting on a static walkway. The student performed a range of examples of various biomechanically distinct gaits (e.g., several different versions each of walking, marching, and jumping). The central biomechanical distinction of walking vs. running was largely observed by speakers of all the languages. However, there were points of between-language disagreement on exemplars even here. The between-language disagreement for gaits more peripheral to human experience was greater, and speakers of English (a manner verb language) showed greater linguistic differentiation of the more peripheral gaits than did speakers of Spanish (a path verb language) and Japanese (a path-and-ground language).

Discussion

These results indicate that structure in the stimulus array provides a constraint on the cross-linguistic construction of meaning, but not an absolute one. Even in a strongly structured domain, some diversity in the meaning of terms can arise. The formulation of meaning is less constrained by stimulus structure in parts of the domain that are more peripheral to human experience, and independently existing typological differences in verb lexicalization patterns are a force that can contribute to diversity in meaning.

Effect of Presentation Style on Children's and Adults' Use of Data Characteristics

Amy M. Masnick (psyamm@hofstra.edu)
Department of Psychology, Hofstra University
Hempstead, NY 11549 USA

Bradley J. Morris (morrisb@gvsu.edu)
Department of Psychology, Grand Valley State University
2117 AuSable Hall, One Campus Drive, Allendale, MI USA

In two studies, we examined children's reasoning in the interpretation phase of an experiment: data were presented as results of a completed experiment, and participants were asked to draw conclusions based on the information they had available. We set up situations with minimal theoretical background information, to make the variation in data characteristics particularly salient.

Two of the most important ideas about data involve expectations about data distribution and expectations about the effect of sample size, and we used these variables as the focal variables in our study. We asked participants to draw conclusions about whether there was a difference between two sets of data and to explain their reasoning (Masnick & Morris, 2002).

In the current study, we wanted to explore how the style of presentation might influence children's and adults' conclusions. Presenting all of the data to be considered at one time could be difficult to process. In addition, the pairwise presentation of data could facilitate comparisons of pairs of data points instead of comparisons of the entire column of data. We used the some of the same datasets as in Masnick and Morris (2002) but presented information in a different format.

Method

Twenty-two third grade students, 29 sixth-grade students, and 50 undergraduate students participated in this study. All participants were shown a cover story describing two engineers testing sports balls. The engineers programmed robots to throw or kick balls a certain number of times to test if they were different. Participants were then presented with three datasets, in one of two conditions.

In the pairwise condition, participants saw data presented in two columns. First, they were shown one pair of data points, then two, four, and then six pairs. After each presentation, participants were asked if there was a difference between the two variables, how sure they were of this difference, and whether they thought the engineers should test the balls again.

In the column condition, participants saw six data points in one column, and one in the other column. They were asked the same questions as in the pairwise condition, and then were presented with additional data points in the second data column (1, 2, 4, and 6 data points at a time). This condition was included to see if reasoning changes when pairwise comparisons are less salient.

Results and Discussion

The results of this study replicated the major finding of Masnick and Morris (2002): Across all grade levels, there was a significant effect of sample size and level of data variation that affected students' sureness that the two columns of data were different.

In this study, participants were asked if they thought the engineers should test the balls again. There were no differences between conditions in the frequency of replies: with fewer data points, most participants wanted more data. When there were six pairs in each column, about 40% of participants still thought more data should be collected.

Participants' justifications for their reasoning were most frequently based on data characteristics such as sample size and the magnitude of differences. There were large age differences, with older participants more likely to name more characteristics. Participants in the column condition were more likely to comment on an outlier affecting their judgment. These participants also said that the inequality of the number of data points in each column was the main reason for wanting the engineers to test at least one ball again.

The findings from this study suggest that students pay attention to data characteristics as early as third grade, and are able to use this information in drawing conclusions. Further, the presentation style of the data affects reasoning about only a small subset of features, suggesting that students are responding not just to the demands of the task, but are interpreting data however they are presented.

Acknowledgments

This research was supported in part by an NICHHD grant to David Klahr (HD25211).

Reference

Masnick, A. M., & Morris, B. J. (2002). Reasoning from data: The effect of sample size and variability on children's and adults' conclusions. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 643-648). Mahwah, NJ: Lawrence Erlbaum Associates.

How the Central System Works? It Uses Fast and Frugal Heuristics

Rui Mata (mata@mpib-berlin.mpg.de)
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

There are two serious contestants as to how the mind works: the Modularity (MT; Fodor, 1983) and the Massive Modularity theses (MMT; Tooby & Cosmides, 1992). Both visions have been targets of criticism Decision making research suggests that one criticism faced by the MT can be overcome by assuming the central system relies heavily on simple heuristics. In this paper fast and frugal heuristics are presented. It is argued that fast and frugal heuristics are the unencapsulated solutions to the central system's potential computational tractability problem, thus supporting MT. Moreover, it is discussed how these heuristics are task-specific but not domain specific, thus undermining MMT.

Modularity vs. Massive Modularity

The MT (Fodor, 1983) is the thesis that the mind is made up of a few modular systems plus a domain-general, unencapsulated central system that serves higher-order functions, like decision-making. The MMT (Tooby & Cosmides, 1992) is the idea that the mind is like a Swiss Army knife, a collection of specialized tools designed to solve adaptive problems. MMT contrasts with MT in that it claims that central capacities can also be divided into domain-specific mechanisms. The two approaches are at opposing sides of a debate and yet they share concerns like computational tractability and domain-specificity. However, for MMT domain-specificity is non-negotiable at all levels. For MT unencapsulation of the central system is the non-negotiable item.

Both visions face challenges. MT faces the obvious criticism that an unencapsulated central system is prone to computational intractability problems. MMT has been criticized for being based on the unwarranted premise that domain-specific mechanisms outperform domain general ones in principle, and for not accounting for the holistic nature of human thinking (Fodor, 2000).

Fast and Frugal Heuristics

The idea that individuals have limited resources, such as time, money, and cognitive capacity, has lead some to propose that people often rely on simple but accurate, fast and frugal heuristics (Gigerenzer, Todd, & The ABC Research Group, 1999).

Different fast and frugal heuristics have been so far identified and tested, including the Take-the-Best and the Recognition heuristics for pair-comparison tasks, and the Quickest heuristic for estimation tasks. These heuristics have been proven to be accurate (i.e., providing more often right than wrong decisions), but also faster (i.e., requiring less computations) and more frugal (i.e., requiring less information) than more standard decision models like multiple regression. Moreover, fast and frugal heuristics are more robust than these latter models when cross-validation is concerned. The reason simple heuristics are so successful is they exploit the structure of decision environments. Importantly, it has been shown that people use such simple heuristics (for a review see Gigerenzer et al. 1999).

Unencapsulated, Domain-general Heuristics

Crucial to the argument exposed here is that simple heuristics can be conceptualized as decision devices which are both information- and processing-frugal without being encapsulated. The fact that they are not encapsulated supports MT. The fact that they are not domain-specific contradicts MMT. Let us evaluate these claims by considering one prototypical heuristic, TTB.

In order to arrive at a decision about which of two objects scores higher on a criterion TTB does the following: (1) it retrieves the cue values of the best predictive cue for that criterion from memory; (2) assesses if one object has a higher value on that cue than the other; (3) if the cue discriminates it chooses the object with the highest value, if the cue does not discriminate, TTB looks up the second best cue, and so forth, until it makes a decision. If no discriminating cues are available TTB guesses.

TTB is not encapsulated in the sense that it has access to all beliefs *in principle* (e.g., beliefs about the value of a cue). However, for TTB there is a limited set of beliefs that it needs to use to reach a decision. In sum, TTB makes decisions in a computationally tractable way not by being encapsulated but by having a stopping rule (i.e., stop search and decide after finding a discriminating cue).

TTB is task-specific because it can only be applied to pair-comparison tasks. However, TTB is domain general because it can be applied to any domain (e.g., social, food choice). Importantly, it has been shown that people use TTB across domains (see Gigerenzer et al. 1999).

From this perspective, one might be lead to think that the central system is but a collection of simple heuristics. However, we still need MT's concept of a central system that chooses between heuristics and reasons about which cues are the best and should therefore be considered first (see also Payne, Bettman, & Johnson, 1993).

References

Fodor, J. (1983). *The modularity of mind*. Cambridge: Bradford Book.

Fodor, J. (2000). *The mind doesn't work that way*. Cambridge: Bradford Book.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambrige, UK: Cambridge University Press.

Tooby, J. & Cosmides, L. (1992). The psychological foundations of culture. In J.H. Barkow, L.Cosmides & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

The Levels of Processing influence the Mere Exposure Effect on Incidental Concept Formation

Ken MATSUDA (ken@p01.mbox.media.kyoto-u.ac.jp)
Takashi KUSUMI (kusumi@educ.kyoto-u.ac.jp)
Faculty of Education, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan

We investigated the effect of perceptual and semantic processing on concept and preference formation. Matsuda & Kusumi (2002, 2003) found three effects using a mere exposure (e.g., Zajonc, 1968) and concept formation paradigm (e.g., Barsalou et al., 1999). The first effect, concept formation by repeated exposure, is based on the event; the concept builds the prototype. The second, prototypical stimuli integrated dimension of each individual, are preferred if the value of that dimension is weighted. Finally, although formed concepts decrease cohesiveness as a function of interval, the prototype is retained.

In the present study, the learning condition was changed from intentional (Matsuda & Kusumi, 2002, 2003) to incidental learning. Category classification performance using incidental learning was found superior to performance based on intentional learning in Parkinson's disease patients (Reber & Squire, 1999). Because the typicality effect on learning by semantic processing is higher than that achieved with perceptual processing (Fujita & Shimizu, 1990), we also examined levels of processing (perceptual vs. semantic) as an independent variable.

Method

Design. A 3 (typicality of stimuli: high, medium and low) × 4 (exposure frequency: 0, 1, 3 and 5 times) × 2 (levels of processing: perceptual vs. semantic)× 2 (interval: 5 min vs. 2 weeks) design was used, with interval and levels of processing manipulated between participants, and typicality of stimuli and exposure frequency manipulated within participants.

Participants. Ninety-six Japanese university students participated in the experiment.

Material. Unfamiliar fish pictures based on Barsalou et al.(（1999) were used. The pictures were classified into types A and B. All stimuli consisted of 10 dimensions (D1-D10), and all stimuli shared D7-D10 dimensionality. Shared dimensions operated typically as independent variables. High-typical stimuli shared D3-D10, medium-typical stimuli shared D5-D10, and low-typical stimuli shared D7-D10. No-shared dimensions had a unique value. Within-distracters were unpresented prototypical stimuli that were integrated with the value of the same- exposure frequency condition. Between-distracters were non-presented stimuli integrated with A and B types.

Procedure. Participants studied unfamiliar fish pictures that consisted of 10 dimensions (0, 1, 3, 5 times incidentally), and classified the fish into two categories (A or B), based on perceptual (shape: round vs. slender) or semantic processing (nature: gentle vs. fierce). Each stimulus was successively displayed for 7 sec; response times were 2 sec, feedback times were 1 sec, and ISIs were 1 sec. After an interval (5 min or 2 weeks), the participants judged the typicality, familiarity, liking, prettiness, and nostalgia of each picture, using a nine-point scale, as well as indicating recognition of new and old items.

Results and Discussion

A. Typicality and Familiarity Judgment In the 5-min interval condition of the present stimuli, the main effects of stimuli typicality and exposure frequency were significant, but the main effects of levels of processing and interaction were not. In the 2-week condition, the effect of exposure frequency was not significant. The results suggested that interval effected a judgment criteria shift from episode to knowledge base. Based on analysis of the within-category distracters, the judged value was higher in the semantic processing condition than in the perceptual processing condition. The data suggest that semantic processing integrates each stimulus and, thus, promotes prototype formation. (Figure 1A).

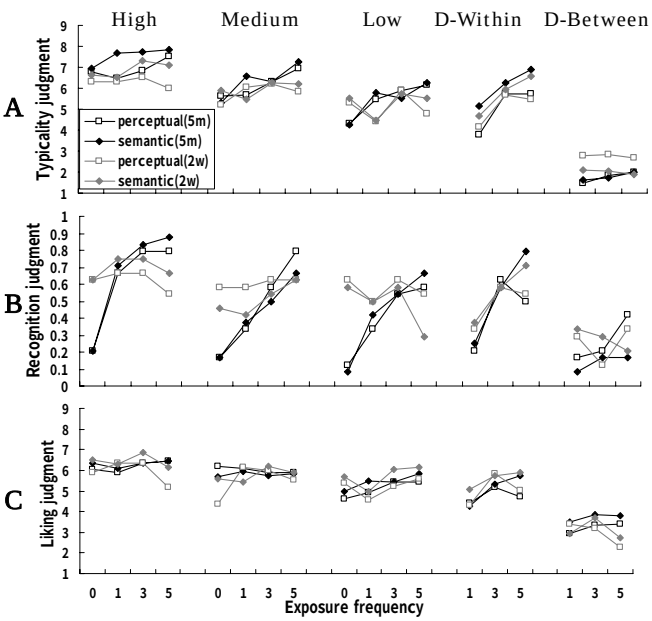


Figure 1 Typicality, Recognition, and Liking judgment scores

B. Nostalgia and Recognition Judgment The effect of exposure frequency on presented stimuli was significant in the 5-min condition but not in the 2-week condition. There was no effect of intention or levels of processing. Analyzing the data from the within-distracter condition revealed that recognition was higher in the semantic processing condition, which integrates high-frequency dimensions, than in the perceptual processing condition (Figure 1B).

C. Liking and Prettiness Judgment High-typical stimuli were preferred, and there was a significant effect of exposure frequency for the low-typical stimuli, as well as for intentional learning (Matsuda & Kusumi, 2002, 2003). There was no effect of levels of processing. The judged value of within-category distracters in the semantic processing condition was high, as compared to that in the perceptual processing condition (Figure 1C).

In conclusion, the effects of intention and levels of processing on judgments were weak. Furthermore, semantic processing in the study phase (compared with perceptual processing) enhanced typicality and affective judgments for the non-presented prototypical stimuli (within-category distracters).

References

Barsalou, L, W., Huttenlocher, J., & Lamberts, k. 1999 Basing Categorization on Individuals and Events, *Cognitive Psychology,* **36**, 203-272.

Matsuda, K. & Kusumi, T. 2002 The mere exposure effect in concept formation. Paper presented at 43rd Annual Meeting of the Psychonomic Society. Kansas City, KA, November.

Matsuda, K. & Kusumi, T. 2003 A long interval affects the mere exposure effect for the prototypes. Paper presented at 25th Annual Conference of the Cognitive Science Society. Boston, MA, August.

Event-based Priming

Ken McRae (mcrae@uwo.ca)
Department of Psychology, University of Western Ontario
Social Science Centre, London, Ontario, N6A 5C2 Canada

Mary Hare (hare@rowan.bgsu.edu)
Department of Psychology, Bowling Green State University
Bowling Green, OH 43403-0228 USA

Several recent articles have emphasized event representations and their role in language processing. For example, Vu et al. (2003) used subject nouns such as *astronomer* versus *director* to promote a situation model that led to disambiguating the meaning of a sentence-final word such as *star*. They found that their manipulation was sufficient to activate selectively the dominant or subordinate meaning of the ambiguous noun. Ferretti, McRae, and Hatherell (2001) used short stimulus onset asynchrony priming to provide evidence that verbs denoting events quickly activate knowledge of typical aspects of those events (verbs primed typical agents, patients, and instruments). McRae et al. (2004) found that nouns denoting typical aspects of events activate verbs, thus suggesting that event knowledge is computed quickly via means other than the name of the event (i.e., a verb). The goal of the present study was to extend this research by testing for priming between nouns that denote events or typical aspects of them.

We used generation norms to select six groups of items. For event nouns such as *baptism*, subjects were asked to "List the types of people and/or animals that are typically found at these events." The norming produced 18 prime-target pairs such as *baptism-priest*. A separate norming study asked subjects to "List the types of things that are typically found at these events." This produced 26 event-thing pairs such as *trip-luggage*. For location nouns such as *tavern*, subjects were asked to "List the people and/or animals that you commonly see in/at each of these locations." This produced 24 items such as *tavern-bartender*. This norming also was conducted for locations and things, producing 30 items such as *garage-car*. Similar norming was also conducted with instrument nouns such as *wrench*. These norming studies produced 24 instrument-living thing pairs such as *wrench-plumber* and 24 event-thing pairs such as *key-door*. Care was taken to exclude event and instrument nouns that are often used as verbs, and to exclude prime-target pairs that form common phrases.

We hypothesized that if people's memory representations are shaped by their experiences with events, then common aspects of events should activate one another. In our priming task, the prime was presented visually for 200 ms, followed by a blank screen for 50 ms, and then the target word was presented until the subject responded. For the people/animals experiments, subjects decided as quickly and accurately as possible whether or not the target referred to a living thing. For the "thing" experiments, subjects decided whether or not the target referred to a concrete object.

There was a 32 ms priming effect for event-people/animal pairs (related: M = 590 ms; unrelated: M = 622 ms; $F_1(1, 18)=5.30$, $p < .05$, $F_2(1, 16) = 7.74$, $p <. 05$), and a 32 ms priming effect for event-thing pairs (related: M = 738 ms; unrelated: M = 771 ms; $F_1(1, 18) = 7.74$, $p < .05$, $F_2(1, 24) = 4.71$, $p < .05$). Both priming effects for locations were also significant: 37 ms for location-people/animals (related: M = 728 ms; unrelated: M = 765 ms; $F_1(1, 20) = 4.39$, $p < .05$, $F_2(1, 22) = 5.29$, $p < .05$); and 29 ms for location-things (related: M = 646 ms; unrelated: M = 675 ms; $F_1(1, 18) = 8.60$, $p < .01$, $F_2(1, 28) = 5.16$, $p < .05$). Finally, there was a significant 58 ms priming effect for instrument-things (related: M = 735 ms; unrelated: M = 793 ms; $F_1(1, 16) = 9.59$, $p < .01$, $F_2(1, 28) = 10.72$, $p < .01$), but a nonsignificant -10 ms effect for instrument-people (related: M = 766 ms; unrelated: M = 756 ms; both F's < 1).

The present study provides additional evidence that semantic memory is organized so that knowledge regarding various aspects of common events can be computed quickly from multiple types of linguistic cues, thus providing valuable information for interpreting language on-line and generating expectancies during language comprehension. As Sanford and Garrod (1981) have stated, "we use a linguistic input to call up representations of situations or events from long-term memory as soon as we have enough information to do so" (p. 115). The present studies, and other recent ones related to them, suggest that nouns that denote typical aspects of common events are often "enough information".

Acknowledgments

This research was funded by NIH grant MH6051701A2 to both authors, and NSERC grant OGP0155704 to the first.

References

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, *44*, 516-547.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R. (2004). A basis for generating expectancies for verbs from nouns. Submitted to *Memory & Cognition*.

Sanford, A. J., & Garrod, S. C. (1981).*Understanding written language: Explorations of comprehension beyond the sentence.* Chichester, England: John Wiley and Sons.

Vu, H., Kellas, G., Petersen, E., & Metcalf, K. (2003). Situation-evoking stimuli, domain of reference, and the incremental interpretation of lexical ambiguity. *Memory & Cognition*, *31*, 1302-1315.

Two stages of visual feature binding: inside and outside the focus of attention

David Melcher (dmelcher@brookes.ac.uk)
Department of Psychology, Oxford Brookes University,
Oxford OX3 0BP, UK

Zoltán Vidnyánszky (vidnyanszky@ana.sote.hu)
Neurobiology Research Group, Hungarian Academy of Sciences
Semmelweis University, 1094 Budapest, Hungary

Despite of the fact that visual features are processed separately in specialized subsystems in the brain, our perceptual experience is of coherent objects. It has been suggested that visual attention acts as a "glue" to bind separate features–such as color, shape, size, motion, and location—into objects (Triesman & Gelade, 1980). This theory would suggest that features outside of visual attention remain unbound perceptually. We tested (1) whether binding occurs outside the focus of attention, and (2) whether feature binding was automatic or dependent on top-down attention to each feature individually. To determine whether two features were bound together, we probed whether paying attention to one feature (color) would also influence the processing of another, task-irrelevant feature (motion) of the same stimulus (cross-feature attention: CFA: Sohn et al., in press). These CFA effects were tested both within the focus of attention (focal attention) and outside the spatial location of attention (global attention).

Methods

The first test was to detect a luminance change in one of two colors of dots (red or green) clustered in one visual hemifield. After a beep, the subject was cued to pay attention to a cluster of dots in the opposite hemifield for a motion direction discrimination task. Unbeknownst to the subject, a brief (150 ms) sub-threshold motion prime was present in the unattended dots during the first (luminance) task (for details of the motion prime, see Melcher & Morrone, 2003). The use of a sub-threshold prime excluded the possibility that the motion signal was attended directly. The prime was presented either in same color of dots attended for the luminance task or in the other color.

In addition to this test of the influence of the prime outside the focus of attention, separate blocks were run in which the prime was in the same dots as those attended for the luminance task. In this case, only the dots in one hemifield were attended during the trial.

Results

We found that global CFA modulation outside the focus of attention spreads to spatiotemporally co-localized features, while inside the focus of attention CFA modulation spreads between all features belonging to the same surface or object. In other words, CFA effects outside the focus of attention were color-specific. An influence of the prime was found only when the prime dots were the same color as those attended in the opposite hemifield for the luminance test. When both tasks were in the same dots, however, the prime dots in the cluster always influenced the later motion test, suggesting that the entire cluster of dots was bound together as a surface.

Conclusions

These results suggest several implications for the role of attention in feature binding. First, these findings imply the existence of a binding mechanism at the local stages of visual processing that links spatiotemporally co-occurring features across the visual field, and that this mechanism is independent of attention. Secondly, these results support previous suggestions of another binding mechanism that acts at the level of coherent surfaces and links all features of the same surface (object-based attention: Duncan, 1984; O'Craven et al., 1999). Third, the CFA effects found here suggest that features are bound automatically, rather than depending on top-down attention to each feature separately. The use of the sub-threshold stimulus showed that attention influences visual processing even when the feature is below the level of conscious awareness.

References

Duncan, J. (1984). Selective attention and the organization of visual information. *J. Exp. Psychol. Gen.*, 113, 501-517.

Melcher, D. & Morrone, M.C. (2003) Spatiotopic temporal integration of visual motion across saccadic eye movements. *Nature Neuroscience.*, 6, 877-81.

O'Craven, K., Downing, P., & Kanwisher, N. (1999) fMRI evidence for objects as the units of attentional selection. *Nature*, 401, 584–587.

Sohn, W., Papathomas, T.V., Blaser, E. & Vidnyanszky, Z. (in press). Object-based cross-feature attentional modulation from color to motion. Vision Research.

Treisman, A.M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97-136.

The View Association Model of Embodiment Effects in Spatial Learning

Gareth E. Miles (GMiles@Glam.AC.UK)
School of Humanities, Law and Social Science, University of Glamorgan,
Pontypridd, S. Glamorgan, UK.

The View Association Model (VAM) is a novel account of how interacting with a spatial layout in different ways can lead to differing representation of the space. VAM provides an account of data demonstrating the learning of different spatial associations when the same layout is learnt through 3D Virtual navigation and 2D map navigation.

Embodied views of cognition suggest that representations of the environment can be understood by examining how they complement the perceptual-motor programs used to interact with the environment (Ballard et al., 1997). Inherent in this perspective is the suggestion that different environments will typically require different perceptual-motor programs that will in turn require different internal representations.

Miles & Howes (submitted) found that spatially close items became associated when participants learnt a space by navigating through it in a 3D Desktop Virtual Environment (DVE), but not when the space was a learnt by navigating a 2D map. The data presented by Miles & Howes suggest two questions. Firstly, what are the differences in the way the DVE and 2D map were learnt? Secondly, how do these differences lead to spatial association in the DVE condition, and its absence in the 2D Map condition? The View Association Model (VAM) attempts to explain these data and provide answers to both of these questions.

The View Association Model

VAM learns the locations of items in a space by learning the visual location of an item in a particular view of the space. A view is defined as a single static depiction of a space (or a portion of that space) from the viewer's perspective. In the plan view condition used by Miles & Howes participants only see a single view of the space. However in the DVE condition a large number of possible views of the environment are possible. A consequence of the views available in the DVE and 2D Map conditions is that different perceptual-motor programs are required for successful navigation.

Although, VAM moves around the DVE and Plan View in similar ways, the motor actions needed to facilitate movement are different. To move in the DVE VAM must point the direction of view toward the item it wishes to move to and press down the space bar. Subsequently VAM adjusts the direction of motion by moving the mouse. When VAM interacts with the 2D view a similar mode is adopted, but key presses are used to navigate. VAM presses a key to start moving in a desired direction and that key remains depressed until another key is preferred.

As VAM moves toward an item in both conditions a course correction algorithm is periodically engaged. VAM searches for the target item in the current view and then compares its current heading with the heading needed to get to the item. In the DVE it will then adjust the view so the target item is in the centre of the view (thus correcting the heading). When interacting with the 2D Map VAM will simply decide which key will move the red dot closest to the target item.

The algorithms used by VAM to move toward an item rely on bottom up processing to initiate course correction. The bottom up processing not only focuses attention on the target item but, by default, focuses attention periodically on other items that happen to appear in the current view. This occurs when VAM searches the current view for the target item, rejecting non-target items before it locates the target item. When attention is focused on these non-target items then there is a chance that VAM will elaborate the name of the item and an associative link will be formed between the item and the current goal item.

Crucially, only items that appear in view will be elaborated. Hence associative links will tend to form between items that appear in the same view. In Miles & Howes' 2D map condition all items are always in the same view. But in the DVE condition only subsets of items appear in any given view and typically items appearing in the same view will be proximal. Hence the interaction of the course correction algorithm and perceptual display lead VAM to predict the pattern of data observed by Miles & Howes.

References

Anderson, J. R. & Milson, R. (1989). Human Memory: An Adaptive Perspective. Psychological Review, 96, 703-719.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Diectic codes for the embodiment of cognition. Behaviour and Brain Sciences, 20, 723-42.

Miles, G. E., & Howes, A. (*submitted*). Spatial and Temporal Contributions to Priming Between Items Located in a Navigable Environment.

Attentional Modulation of Lexical Effects in an Interactive Model of Speech Perception

Daniel Mirman (dmirman@andrew.cmu.edu)
James L. McClelland (jlm@cnbc.cmu.edu)
Lori L. Holt (lholt@andrew.cmu.edu)
Center for the Neural Basis of Cognition & Department of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

A number of studies have demonstrated that the strength of lexical effects on phoneme processing can be modulated by attention (e.g., Cutler et al., 1987; Eimas, Hornstein, & Payton, 1990; Vitevitch, 2003). The TRACE model (McClelland & Elman, 1986) posits direct feedback from lexical processing to phonemic processing, thus accounting for lexical influences on phoneme identification. However, the TRACE model lacks a mechanism for modulation of this feedback through attention. Some researchers (Norris, McQueen, & Cutler, 2000) have argued that this is a weakness of the interactive view of speech perception and is one reason to prefer an autonomous model.

We consider biased competition (Desimone & Duncan, 1995) as a possible attention mechanism that fits within the interactive framework of TRACE. In the context of TRACE, when an input is presented, phonemes that are partially consistent with the input compete through lateral inhibition. This competition is biased by lexical feedback proportional to the magnitude of lexical activation. Activation of lexical items is based on excitatory input from the phoneme layer and lateral inhibitory interactions among lexical items. The magnitude and rate at which lexical items become active can be manipulated by a scaling factor on the lexical units' response to input. This, in turn, influences the strength of lexical influences on phoneme perception. That is, task or stimulus conditions that cause participants to direct attention away from lexical processing may operate by causing a dampening of lexical layer activity and thereby reducing lexical biasing of phoneme processing. To implement this mechanism in TRACE, an attentional scaling parameter (α) was added to the function specifying the change in activation for lexical units for each processing cycle. When α=1.0, this is the standard TRACE model as implemented by McClelland and Elman (1986), when α<1.0, the lexical activation is dampened and lexical effects should be reduced.

This mechanism was tested in two cases of lexical effects on phoneme identification. Ambiguous phonemes tend to be perceived as lexically consistent (Ganong, 1980), but the strength of this effect varies with task and stimulus differences (see Pitt & Samuel, 1993, for review and meta-analysis). The attention parameter captured this variability. When lexical attention is high, lexical items become more active more quickly, thus providing stronger and earlier feedback to the phoneme level and biasing perception of the ambiguous acoustic input. When lexical attention is very low, lexical items become active more slowly, thus providing less feedback to the phoneme level and causing a small and late-developing lexical bias.

A second lexical effect on phoneme recognition is that phonemes are recognized more quickly in words than nonwords. This word advantage has also been shown to be affected by task and stimulus factors (e.g., Cutler et al., 1987). Variation of the attention parameter also captures this variability: at high α values, TRACE is faster to recognize phonemes that are embedded in words; at lower α values, the word advantage disappears. This is because lexical items are less active, thus they provide less support to their constituent phonemes.

The addition of a scaling parameter that dampens overall lexical layer activation provides a simple mechanism that works within the interactive framework of the TRACE model to modulate the strength of lexical influences on phoneme processing.

References

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. Cognitive Psychology, 19(2), 141-177.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193-222.

Eimas, P. D., Hornstein, S. M., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. Journal of Memory & Language, 29(2), 160-180.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception & Performance, 6(1), 110-125.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18(1), 1-86.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. Behavioral & Brain Sciences, 23(3), 299-370.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. Journal of Experimental Psychology: Human Perception & Performance, 19(4), 699-725.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. Clinical Linguistics & Phonetics, 17(6), 487-499.

The Frame Problem in Text Analysis

Maki Miyake (mmiyake@dp.hum.titech.ac.jp)
Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

Hiroyuki Akama (akama@dp.hum.titech.ac.jp)
Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

Masanori Nakagawa (nakagawa@nm.hum.titech.ac.jp)
Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

General Formatting Instructions

The aim of this research is to evaluate the extent to which Thesaurus allow us to modify a researcher's knowledge frame. A well calculated Thesaurus has the power to overwrite existent knowledge frames, or habitual "heuristics" for the humanities, if word occurrence data are rigorously manipulated by the algorithms of statistical linguistics. However, even if all targets and means of data gathering and analyzing are readily available, there remains behind various interpretations of subjects a sort of Frame question of how we partition off the texts and documents to avoid arbitrary text segmentation. Frames are needed before we can gather and interpret the data for a word occurrence computation.

If the problem of text segmentation remains unresolved, any experiment in quantitative text analysis will be still far from being realized. We need a sort of "TextTiling" methodology enabling an objective segmentation based on objective criteria.

This holds true for the frame setting in parallel and variant texts as the synoptic Gospels. We have to point out that the basic frame for Biblical research was taken from second hand data called the Parallel Synoptic Table (in abbreviation, PST), which shows the order and arrangement of the "pericopes" belonging to the Synoptic Gospels. The frame traditionally prepared was built only by a "Form Criticism", which divided the texts into parts by the arbitrary unities coming from tradition or reduction.

The purpose of the PST framework has been the resolution of the problem of "who quoted whom" in writing respectively the first three Gospels. However we propose an alternative and more objective way of segmenting the parallel texts by using our web-based biblical software, named "Tele-Synopsis", which is designed to gather information of the word usage under various conditions and to help further statistical approach to the origin of the variant texts. A quantitative analysis (factor analysis) is applied to the lexical datasets obtained by changing framing conditions in order to verify some traditional hypotheses made to explain the mutual relationship of the synoptic Gospels. Our framing principle is that the entire texts can be classified into the following 7 categories which are A: common part of the three Gospels, B: part common to Matthew and Mark, C: part common to Mark and Luke, D: part common to Matthew and Lukas, E: part peculiar to Matthew, F: part peculiar to Mark, G: part peculiar to Luke. It is natural that the category to which each instance of word has to belong for constructing a biblical Thesaurus varies according to the way in which we partition off the parallel texts. But the results of the factor analysis applied to the multiple datasets showed high robustness in the sense that the types of loading matrix are more or less similar *except* for the dataset depending on the traditional PST.

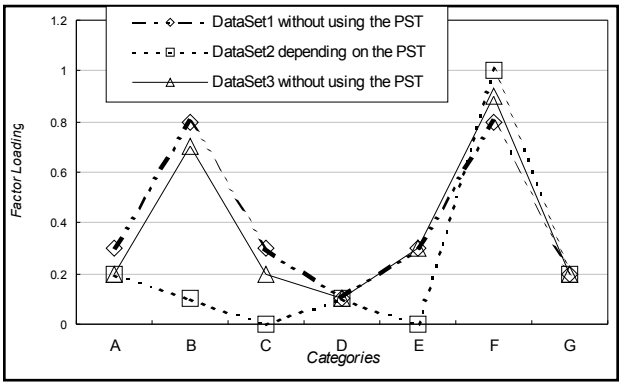


Figure 1: One of the factor patterns for the identical variables with different datasets.

Acknowledgments

This paper was made possible largely through grants from the 21st Century Center of Excellence Program "Framework for Systematization and Application of Large-scale Knowledge Resources". We would like to acknowledge here the generosity of this Center.

References

Minsky, M.L., (1975). A Framework for representing knowledge, *The Psychology of computer vision*, pp.211-277.

Hearst, Marti A.(1997). TextTiling: "Segmenting text into multi-paragraph subtopic passages", *Computational Linguistics* 23, pp.33-64.

Kurt Aland. (1989). *Synopsis of the Four Gospels*, German Bible Society Stuttgart.

Learning through verbalization (2): Understanding the concept of "schema"

Naomi Miyake, Hajime Shirouzu, & Yoshio Miyake ({nmiyake; shirouzu; ymiyake}@sccs.chukyo-u.ac.jp)
School of Computer and Cognitive Sciences, Chukyo University
101 Tokodate, Kaizu-Cho, Toyota, 470-0393 JAPAN

When one starts to learn a new topic, it is essential to understand the terminology, to the point where one can use them comfortably. For such learning, balance is required between concrete experiences and their abstract verbalization, but how to achieve the balance has not yet been studied systematically. In this report we compare three sets of learning activities to see the effects of the amount of concrete experiences and their verbalization on learning. While a short demo with high demand on abstraction does not yield significant verbalization, ample practices with reflection appear to solicit natural generalization.

Comparison of three classes

Three undergraduate classes were taught the concept of schema through structured activities around the "Day arithmetic" (Lindsay & Norman, 1977), where the students were to solve problems like

When Wednesday + Tuesday = Friday,
what is Tuesday + Friday?

The classes differed in the amount of practices of the problem, as well as in the types of verbalization required to summarize this experience. In Class 1, students solved three problems, while Class 2 solved 3 and then 20, and Class 3 solved 3, 72, 60, and 60 problems in chunks. This practice was followed by the question of what strategy they would choose to solve many Day arithmetic problems. After that, a transfer problem, "m+b=?" was posed. At the end of the unit, each class was asked to summarize their experiences. For the numbers of the students, see Table 2.

Amount of experiences and choice of strategies

Table 1 shows the students' choices of strategies to tackle many problems, either rote memorization of the answers, use of a table of answers, or of rules such as "to add a Monday, answer the next day of the addend." Rules are highly effective, but this fact was only graspable after a relatively many practices.

Table 1: Strategy choice

Class	No. of Trials	Strategy choice		
		Memory	Table	Rules
1	3	20.0%	70.0%	10.0%
2	23	16.2%	32.3%	51.5%
3	195	15.2%	15.2%	69.6%

Micro-generation of a schema

To the transfer problem of m+b, many students answer "o," paralleling this to the Day arithmetic. Some even extended its rule and solved this by just going down the alphabet two more letters from m, without counting. Both cases indicate that the students generate a schema-like understanding, applicable to a similar problem. Table 2 itemizes the ratio of types of this micro-generation. The success rate of the micro-generation of the schema is quite high, and sparing the practice time does not affect the generation pattern.

Table 2: Answer types of "m+b"

Class	Count-up	Transfer	No answer
1 (n=81)	50.6%	44.4%	4.9%
2 (n=71)	59.2%	38.0%	2.8%
3 (n=92)	63.0%	35.9%	0%

Abstraction at the end of the unit

At the end of this unit, the students were asked to summarize their experiences, in different instructions. The answers were categorized as "Concrete" when they only referred to specific examples and/or procedures; as "Moderate" when they referred to the strategies and effects; as. "High" when they included explicit comments on their commonality and/or generalizability. Class 1 students were asked to describe what kind of knowledge their rules were, which was too difficult to answer, particularly after a short demo. Class 2 students were encouraged to explain the Day arithmetic to their friends. Most students chose to stick to concrete procedures, ignoring the schemas. Contrastingly in Class 3, the students were asked to comment on the most important points of the unit. Possibly scaffolded by the ample amount of experiences as a base for reflection, this attempt was most successful among the three classes.

Table 3: Abstraction levels of summaries

	Ratio of answerers	Answer abstraction levels		
		High	Moderate	Concrete
1	23.4%	15.8%	42.1%	42.1%
2	91.5%	6.2%	1.7%	92.3%
3	100%	24.7%	25.9%	49.4%

The overall pattern indicates the importance of concrete experiences, as a basis for significant reflection. A short demo with highly abstracted explanation might appear to save time, but could impair the quality of learning.

Acknowledgments

This research is supported by CREST/JST to the 1st author.

References

Lindsey, N., & Norman, D. (1977). Human information processing. New York:Academic Press

Cognitive Style, Gender, Alignable Differences and Category Sorting

Marnie L. Moist (mmoist@francis.edu)
Lewis R. Ruddek, Jamie L. Bernazzoli, Stefanie N. Fedder, Nicole M. Lang,
Alyssa Stoehr, Linsey O'Donnell, Matt B. Baum, Scott F. Caldwell
Behavioral Sciences Department, St. Francis University
Scotus Hall Rm. 217, Loretto, PA 15940

Introduction

Witkin et al. (2002) notes that field independents process analytically, whereas field dependents process globally. Markman and Genter (1993) define alignable differences (AD) as arising from an underlying commonality (e.g.,'one has more legs' arises from 'both have legs'). It follows that:

1) Field independents may produce fewer AD, sort more categories, and create less variable-sized categories than field dependents.
2) Cognitive style may interact with artificial stimulus sets, which vary in shared attributes (characteristics true of multiple category members). Specifically, the "mixed" set, the only one of the three sets allowing selective attention to vary between either "common" (i.e., majority shared) or "idiosyncratic" (i.e., minority shared) attributes, may elicit the largest difference in AD production between field independents and dependents.
3) There should be gender differences in cognitive styles (Witkin et al., 2002), number of categories sorted, and/or category size variability (Pettigrew, 1958).

Methods

Participants

87 (23M/64F) college students (98% Caucasian; M age = 19) from a Catholic school, participated for extra credit.

Materials

The Group Embedded Figures Test (GEFT) by Witkin et al. identified cognitive style. Nine stimulus sets, each with 20 artificial animal line drawings, allowed category sorting. Four "common" sets each consisted of 8 common attributes (e.g., "tail") with varying values (e.g., 'peacock') shared by 16/20 animals. One "mixed" set consisted of 8 common attributes; 4 shared by 16/20 and 4 shared by 4/20. Four "idiosyncratic" sets each consisted of 8 common attributes shared by 4/20 pairs. All sets were counterbalanced to ensure the same attributes/values were used across all sets. Response sheets were used to record category sort answers and the first difference noticed for each of 20 animal pairs.

Design and Procedure

All individually tested participants were randomly assigned one ordered stack of 20 animals, which they sorted into as many categories as they wanted. Then they listed the first difference they noticed for the same 20 animal pairs, followed by a second, identical category sort task with the same 20 animals. Finally, all were timed and scored on the GEFT test as instructed by Witken et al.

Results and Discussion

GEFT inter-rater reliability was 99%; AD reliability was 95%. *Hypothesis 1*: A multiple regression analysis (see Table 1) predicting AD, model $F(4,71) = 4.30$, adj.$R^2 = .15$, $p = .004$, showed that alignable differences decreased as field independence increased and as animal pairs became more different from each other (i.e., shared fewer attributes). A simple correlation, $r(74) = -.31$, $p = .003$, showed that as field independence increased, the sorted category size variability at Time 2 decreased.

Table 1: Multiple regression on alignable differences (AD).

Four I.V.s	Stand. *B*	*SE*	*p*-value
GEFT scores	**-.31**	.01	**.009**
Categ. sorted Time 2	+.06	.01	.608
Stimulus Set	**-.26**	.04	**.021**
Categ. variab. Time 2	-.18	.03	.156

Note: 11 participants' data were removed here due to uncorrected vision.

Hypothesis 2: There was no cognitive style X stimulus set interaction, though a corrected confound and more equal numbers tested per condition may change this in the future. *Hypothesis 3*: An unequal variance independent t-test, $t(55) = 2.32$, $p = .024$, showed that females ($M = 6.20$, $SD = 2.3$) sorted more categories at Time 1 than males ($M = 5.17$, $SD = 1.61$). However, for the Time 2 category sort, a 2-way ANOVA, $F(5,81) = 2.46$, $p = .04$, showed a significant gender X stimulus set interaction, $F(2,81) = 4.88$, $p = .01$. Females ($M = 7.59$, $SD = .48$) sorted more categories for "common" stimuli than males ($M = 5.20$, $SD = .78$), but males ($M = 7.63$, $SD = .88$) sorted more categories for "idiosyncratic" stimuli than females ($M = 5.78$, $SD = .58$). No gender differences in cognitive styles or sorted category size variability occurred.

Acknowledgements

The authors thank St. Francis University for the Excellence in Education Award, 2003, and Dan Wetklow for support.

References

Markman, A.B & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

Pettigrew, T.F. (1958). The measurement and correlates of category width as a cognitive variable. *Journal of Personality, 26*, 532-544.

Witkin, H.A., Oltman, P.K, Raskin, E., & Karp, S.A. (2002). *Group Embedded Figures Test*. Mind Garden.

Very Brief Interruptions Result in Resumption Cost

Christopher A. Monk (cmonk@gmu.edu)
Deborah A. Boehm-Davis (dbdavis@gmu.edu)
Department of Psychology, George Mason University
4400 University Dr., Fairfax, VA 22030 USA

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
Naval Research Laboratory, NRL Code 5513
Washington, DC 20375 USA

Introduction

Recent research suggests that the disruptive effects of interruptions arise from decay of the activation associated with the primary goal while attending to the interrupting task (Altmann & Trafton, 2002; Monk, Boehm-Davis, & Trafton, in press; Trafton et al., 2003). This disruption is seen in the additional time required to resume a task after it has been interrupted; that is, in the reaction time (RT) from the onset of a display after an interruption until the first keypress is made (this RT is called the *resumption lag*). The purpose of this study was to test this interpretation by looking at the resumption costs associated with very brief interruptions, where the model predicts minimal goal decay.

Method

Twelve undergraduates from the George Mason University psychology subject pool participated for course credit. The experiment was a single factor within-subjects design with three interruption lengths (1/4 s, 1 s, and 5 s) and an uninterrupted condition. The dependent measure was the post-interruption *resumption lag*, which was the reaction time from the onset of the VCR display (after an interruption) to the first click on a VCR button.

The primary task was to program a simulated VCR, which consisted of four subtasks: entering the show's start-time, end-time, day of week, and channel number. The screen was blank during the interruptions (there was no task) and the participant was required to wait until the VCR was displayed again before resuming the programming task. The target information was posted next to the monitor on a 3x5-index card at all times.

The experimenter trained the participants through demonstration and practice of uninterrupted and interrupted trials. Each participant completed 20 experimental trials (five trials for each of four conditions). For each interruption trial, participants began with the VCR task and were interrupted every five seconds until the VCR program entry was completed.

Results and Discussion

The time between keypresses (*lag*) was measured every five seconds in the uninterrupted condition to provide a baseline comparison for the resumption lags in the interruption conditions. Figure 1 shows the mean resumption lags and confirms a significant main effect of interruption condition, $F(3, 33) = 48.88$, $p < .001$, $MSE = 7{,}190$. Paired comparisons showed that the 5-second interruption condition ($M = 1115$ ms, $SD = 129$) took significantly longer than the other three conditions, and the uninterrupted condition ($M = 706$ ms, $SD = 80$) was significantly shorter than each of the interrupted conditions. The resumption lags for the 1/4 s ($M = 974$ ms, $SD = 79$) and 1 s conditions ($M = 974$ ms, $SD = 107$) were not reliably different.

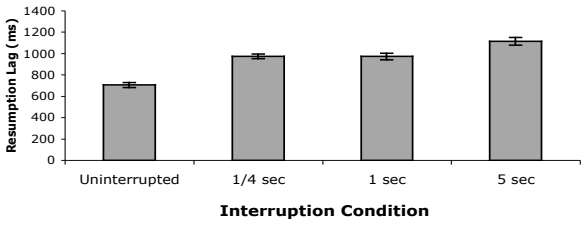


Figure 1. Mean resumption lags.

The difference between the interrupted and uninterrupted conditions confirms the prediction from the goal-activation model (Altmann & Trafton, 2002). Further, the presence of a resumption cost for both the 1/4 s and 1 s interruption conditions shows that goals decay quite rapidly. Even for the briefest interruptions, there is a penalty to be paid when resuming the primary task.

Acknowledgements

This research was supported in part by grant number 55-8122-01 from the Office of Naval Research to Greg Trafton.

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: an activation-based model. *Cognitive Science*, *26*, 39-83.

Monk, C. A., Boehm-Davis, D. A., & Trafton, J. G. (in press). Recovery from interruptions: implications for driver distraction research. *Human Factors*.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies, 58,* 583-603.

Evidence for Multiple Strategy Use Within a Single Logic Problem

Bradley J. Morris (morrisb@gvsu.edu)
Grand Valley State University
One Campus Drive, Allendale MI, 49506

Christian D. Schunn (schunn@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
3939 O'Hara St., Pittsburgh, PA 15260 USA

When solving a logic problem, do reasoners use a single processing step or do they use a series of steps? Further, are these steps derived from the same inferential mechanism or different inferential mechanisms? Traditional models of logical reasoning posit a single solution using one mechanism (e.g., rules, models). A modification of this model, the dual processing model, suggests that logical inferences are the result of a competition between two different mechanisms. Though there are two mechanisms, a single inferential step is executed based on a decision between two candidate solutions. A third possibility, the Logical Strategy Model (LSM), suggests that logical reasoning makes use of more than one inferential step making use of a variety of inferential mechanisms (i.e., strategies).

The LSM predicts that reasoners use strategies based on task demands such as believability. For example, given familiar content, a reasoner will likely use a knowledge-based heuristic. Reasoners may use multiple strategies at different points within a single problem. For example, reasoners might (1) begin with one strategy and shift to another strategy or (2) begin using one strategy and revise their approach using the same strategy. In either case, reasoners would be using a dynamic process in which they may begin with a strategy and change their approach based on changing problem factors and goal states (see Epstien, 1994). To examine this, we performed a verbal protocol study of a series of logical syllogisms.

Method

Subjects. Five University students were recruited from Introductory Psychology courses.
Materials. A series of 32 logical syllogisms were created varying the following dimensions: Abstract v. Concrete, Unfamiliar v. Familiar, Valid v Invalid X True v. False.
Procedure. Subjects were asked to "think-aloud" as they solved 32 syllogisms. As a warm-up, subjects were given a series of multiplication and word scramble problems to practice the verbal protocol.
Coding. Once completed, the session was transcribed. The resulting protocol was coded for strategy use. Strategy use was coded by matching elements of subject discourse to salient elements of proposed strategies. For example, a Token-Based strategy involves the creation (and search) of models derived from premises (e.g., "Some X are Y, so some X are not Y"). A Knowledge-Based strategy derives inferences from a match between problem elements and current knowledge (e.g., "It's not true that some fish have legs so this is false"). Finally, each problem was coded for cues indicating a change in current strategy (e.g., "that can't be right").

Results

All subjects used more than one strategy on a single problem, most (4/5) for each problem type. Subjects were most likely to use multiple strategies when validity and truth or falsity of the conclusion was in conflict (Valid & False, Invalid & True, see Table 1). In these cases, subjects were likely to re-examine their initial conclusion by using a new strategy than by using the same strategy in light of new information (i.e., a putative conclusion). Table 1 also reports whether the second strategy used was the same or a different than the initial strategy. The results indicate that reasoners commonly use multiple strategies in a single problem and that the type of strategies used can be predicted on the basis of task demands.

Table 1- Strategy use by problem type

Problem Type	Mean Number of Strategies	Most frequently used first strategy	Used Same	Used New
A	1.4	Token-based	**65%**	35%
C + U	1.4	Token-based	**70%**	30%
C + F + V + T	1.2	Knowledge-Based	**65%**	35%
C + F + I + T	1.9	Knowledge-Based	25%	**75%**
C + F + V + F	2.3	Knowledge-Based	40%	**60%**
C + F + I + F	1.4	Knowledge-Based	**80%**	20%

References

Epstein, S. L. (1994). For the Right Reasons: The FORR Architecture for Learning in a Skill Domain. *Cognitive Science, 18* (3): 479-511.

The Effect of Spatial Ability on Learning from Text and Graphics

Julie Bauer Morrison (morrison@bryant.edu)
Department of Applied Psychology, Bryant College
1150 Douglas Pike, Smithfield, RI 02917-1284 USA

Introduction

Previous research has shown that participants' comprehension of spatial information described in text improves when either graphics or animation accompanies the text (Morrison & Tverksy, 2001). This relationship was found to be true in all cases for low spatial ability participants; however, it was only true under the most difficult conditions for high spatial ability participants. When given enough time to do so, high spatial participants were able to mentally imagine what was described in the text, and therefore did not need externally provided graphics.

The high and low spatial ability distinctions described above were based on a median split of scores on the Vandenberg and Kuse (1978) Mental Rotation Test. The study was conducted with undergraduate students from Stanford University, raising concerns about the accuracy of the categorization and the generalizability of the results. The study presented here replicates the original with a more typical college population.

The Learning Study

Method

Fifty-nine Bryant College undergraduates participated in the study. Although specific GPA data was not available for the participants, the average high-school GPA of an entering student is 2.95 at Bryant and 3.90 at Stanford (Undergraduate Guide, 2003). After completing a test of spatial ability, the Vandenberg and Kuse (1978) Mental Rotation Test, participants began the learning phase of the experiment. They read through a learning interface with 7 novel rules four times, studying for as long as they wished. The interface included text, text plus static graphics, or text plus animated graphics. Following the learning phase of the experiment, participants completed three timed performance tests requiring applications of the rules. The scores on these tests have been combined into a composite Problem-Solving Score, which can range from 0-100.

Results

As with previous analyses, due to a lack of difference between the static and animated graphics conditions, these conditions were combined into a text plus graphics condition. The problem-solving data was analyzed with a one-way (text vs. text plus graphics) ANOVA with spatial ability as the covariate. Participants performed better in the graphics condition than in the text condition, $F(1,56)=21.1$, $p<.01$. High spatial ability participants performed better than their low spatial counterparts, $F(1,56)=16.0$, $p<.01$. Participants were separated into low and high spatial ability groups according to a median split of spatial ability scores. Figure 1 displays the Problem-Solving Score earned by the low and high spatial ability groups across the text and text plus graphics conditions.

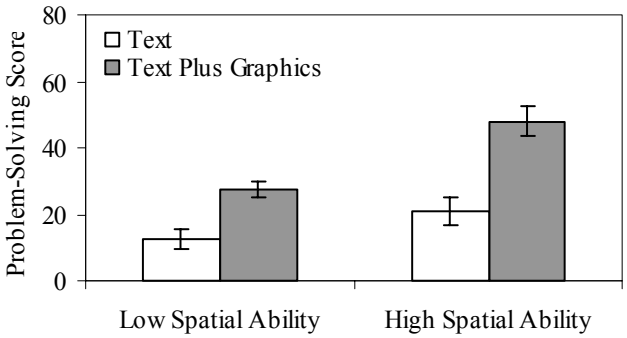


Figure 1: The relationship between spatial ability and interface type for Bryant students.

Discussion

Participants with high spatial ability and those who studied static or animated graphics were better able to learn the spatial rules described in the interface. These results replicated the previous research, in part. In the Stanford version of this study, the low spatial text participants performed more poorly than the three other groups, which did not differ, showing that graphics have benefits, but only for low spatial participants (Morrison & Tversky, 2001). However, the pattern of data seen above is identical to the performance of Stanford participants when limited to studying the interface a single time and with instructions to do so quickly. This suggests that when spatial ability decreases (MRT Score: Bryant M=7.53, SD=3.4; Stanford M=9.17, SD=4.2) and/or when the task becomes more difficult, the benefits of graphics become more pronounced. Although the participants in both studies were college students who may have higher aptitude than the general population, it is clear that there is a learning advantage when spatial material is presented with accompanying graphics.

References

Morrison, J.B. & Tversky, B. (2001). The (in)effectiveness of animation in instruction. *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems* (pp. 377-378). Seattle: ACM.

Undergraduate Guide to Four Year Colleges 2004 (34th ed.). (2003). Stamford, CT: Peterson's.

Vandenberg, S.G. & Kuse, A.R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47,* 599-604.

Baseline Explorations in the Spatial Construal of Time

Benjamin A. Motz (bmotz@cogsci.ucsd.edu)
Rafael E. Núñez (nunez@cogsci.ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Dr., La Jolla, CA 92093-0515

Introduction

Due in part to theoretical advances in cognitive linguistics (e.g., Lakoff & Johnson, 1980; Núñez, 1999) and recent studies in cognitive psychology (Boroditsky, 2000; Gentner, 2001), it has become well accepted that humans apply spatial principles to their conception of time. Statements such as, "Cogsci 2004 has arrived," reflect a mental model in which temporal abstractions are given meaning via spatial metaphors. Examining these metaphors, Boroditsky (2000) used spatial priming to influence subjects' interpretation of "forward" (as earlier or later) when disambiguating the meaning of *Move the meeting forward*. She reported the baseline interpretation of "forward" to be "about evenly split" (p. 9) between earlier (45.7%) and later (54.3%) when subjects received no priming.

As part of a larger research project, the present report focuses on these baseline interpretations. In order to standardize the responses and improve replicability, subjects were shown a graphic display with no reference to objects moving toward an observer, or an observer moving toward objects—the two situations widely believed to exclusively influence conceptual models when disambiguating the meaning of "forward" (Gentner, 2001).

Methods

66 undergraduate students at the University of California, San Diego participated in the study as part of their course requirements. Subjects were shown a static display of five stationary colored boxes. Two of the boxes contained balls; the others were empty. During this presentation, subjects were asked to respond to five questions, including: *What is the color of the box containing the black ball?* and *How many boxes are in the display?*

Immediately following these presentations, subjects were instructed to turn to subsequent pages in their questionnaires containing the following two target questions, the order of which was balanced across subjects.

Next Wednesday's meeting has been moved forward two days. On what day will the meeting now take place?

Tomorrow's 12:00 (noon) meeting has been moved forward two hours. At what time will the meeting now take place?

Results and Discussion

Subjects' responses to the target question are shown in Table 1. Chi-square analyses indicated that responses are significantly different from those expected by chance when the question on the scale of days is asked first ($p < 0.005$). This is not the case when preceded by the question on the scale of hours.

Table 1: Number of responses by order of target questions.

	Day then Hour	Hour then Day
Monday	8	17
Friday	24	16
10:00am	7	16
2:00pm	25	18

It is widely believed that time is construed specifically in terms of observers and motion (Gentner, 2001). Our methodology, which bore no explicit reference to such concepts, elicited responses significantly different from those expected by chance. Furthermore, this difference is sensitive to the time scale of the target question. These findings suggest that the specific spatial metaphors for time are more complex than previously assumed.

The data in the upper left cells of Table 1, which are analogous to Boroditsky's (2000) baseline data, can hardly be interpreted as "about evenly split." This suggests that there may be many pragmatic constraints to be considered and that there are more complex variations of conceptual mappings from space to time. Without making claims as to exact baseline responses to ambiguous questions about time, the present findings suggest a reconsideration of what is necessary and sufficient to elicit priming effects in the spatial construal of time.

Acknowledgements

The authors are grateful to Ursina Teuscher for her helpful comments and involvement.

References

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75*, 1-28.

Gentner, D. (2001). Spatial metaphors in temporal reasoning. In M. Gattis (Ed.) *Spatial Schemas and Abstract Thought* (Chapter 8, pp. 203-222). Cambridge: MIT Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Núñez, R. (1999). Could the future taste purple? *Journal of Consiousness Studies, 6*, 41-60.

Low Frequency Waves on EEG Recordings during Stimulation of Sound

Hiroyuki Murakami (murakami@otsuma.ac.jp)
Department of Social Information Studies
Otsuma Women's University
2-7-1 Karakida, Tamashi, Tokyo 206-0035, Japan

Introduction

It has been reported that delta band waves were observed during sleep (Gevins & Cutillo, 1986; Melnechuk, 1988), during altered states of consciousness (Jovanov, 1997) and during hypnagogium (Faber et al., 2003). In this article, the delta band waves are shown in spectral power of EEG recordings during sensory stimulation of sound. Four patterns of spectral power are shown in stimuli of Song and Talk by female and male participants.

While the participants were stimulated by sound, the EEG data were recorded into ESA-16 (Musha, 2000) from 10 electrodes according to the international 10-20 system. We present the content and the strength of the delta (1/4 ~ 5 Hz), theta (5 ~ 8 Hz), alpha (8 ~ 13 Hz) and beta band (13 ~ 20 Hz) waves from the spectral power of EEG recordings.

Experiment

Eight university students (4 females and 4 males) aged between 21 and 23 years participated to the experiment. The materials were song and talk.

Song Boy Soprano: Green grasslands in England (60"), by Anthony Way. Rock: Why I'm me (60"), by RIZE. Cheer Song: Aida! Decide a goal! (60"). Chorus: Barbie Girl (60"), by AQUA. Hip Hop: Return of the Ripper (1'29"), by LL Cool J.

Talk Man's DJ: The voice of the man in the 40's who talks slowly and softly in a low voice can be heard (60"). Woman's DJ: The voice of the woman in the 20's who talks clearly in a cheerful voice can be heard (24"). Woman's Voice: The voice of the woman in the 20's who talks fast in a high-pitched voice can be heard (30"). Conversation: The voice of two native speaker men who talk fast in a low voice in English can be heard (60").

Results

We obtained the following four patterns of spectral power.

Pattern 1 (Song, female participants): The delta band waves of the content 21.0% with small strength were observed on the frontal region, the theta band waves of the content 13.6% with rather large strength on the lateral and back regions, the alpha band waves of the content 57.6% with large strength on the frontal, lateral and back regions, and the beta band waves of the content 7.8% with small strength on the lateral and back regions.

Pattern 2 (Song, male participants): The delta band waves of the content 44.6% with rather large strength were observed on the frontal region, the theta band waves of the content 3.3% with very small strength on the frontal region, the alpha band waves of the content 47.6% with large strength on the frontal, lateral and back regions, and the beta band waves of the content 4.5% with very small strength on the lateral and back regions.

Pattern 3 (Talk, female participants): The delta band waves of the content 16.9% with very small strength were observed on the lateral and back regions, the theta band waves of the content 9.7% with small strength on the lateral and back regions, the alpha band waves of the content 66.7% with large strength on the frontal, lateral and back regions, and the beta band waves of the content 6.7% with very small strength on the lateral and back regions.

Pattern 4 (Talk, male participants): The delta band waves of the content 41.8% with small strength were observed on the frontal region, the theta band waves of the content 6.7% with very small strength on the frontal region, the alpha band waves of the content 37.5% with large strength on the lateral region and with small strength on the frontal and back regions, and the beta band waves of the content 14.0% with very small strength on the lateral and back regions.

Acknowledgments

We thank Sonomi Arai, Megumi Degawa, Yuko Murakami, Makiko Nakajima, Aya Sawada, Asami Watanabe and Mariko Yasutomi for their diligent help with data collection

References

Faber, J. et al. (2003). Electrical Brain Wave Analysis during Hypnagogium. *Neural Netword world Vol. 13.*

Gevins, A.S., & Cutillo, B.A. (1986). Clinical Applications of Computer Analysis of EEG and Other Neurophsiological Signals. In F.H. Lopezs da Silva, W. Soorm van Leeuwen & A. Remond (Eds.), *Handbook of Electroencephalography and Clinical Neurophysiology, Vol.2.* Amsterdam: Elsevier Science Publishers.

Jovanov, E. (1977). On The Methodology of EEG Analysis During Altered States Of Consciousness. *Proceedings of the First Annual ECPD International Workshop on Scientific Bases of Consciousness.* Rakic, L et al.: European Centre for Peace and Development (ECPD) of the United Nations University for Peace.

Melnechuk, T. (1988). Dynamics of Sensory and Cognitive Processing by the Brain. In E, Basar (Ed.), *Springer Series in Brain Dynamics, Vol.1.* Berlin: Springer.

Musha, T., Kimura, S., Kaneko, K., Nishida, K. & Sekine, K. (2000). Emotion Spectrum Analysis Method (ESAM) for Monitoring the Effects of Art Therapy Applied on Demented Patients. *Cyber Psychology & Behavior, 3,* No.3, 441-446.

The effect of repeated presentation and aptness of figurative comparisons on preference for metaphor forms

Keiko Nakamoto (kenakamoto@nifty.com)

Takashi Kusumi (kusumi@mbox.kudpc.kyoto-u.ac.jp)

Department of Cognitive Psychology in Education, Graduate School of Education, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Introduction

Figurative comparisons consisting of two nouns (a target and a base) can be expressed in two grammatical forms, i.e., in simile form (X is like Y), or in metaphor form (X is Y).

Recently, many studies have shown that there are substantial differences between metaphors and similes. In particular, much evidence has been found for people's grammatical preference for figurative comparisons (e.g., Chiappe & Kennedy, 1999). Among hypotheses proposed for explaining this grammatical form preference in relation to figurative comparisons, we focused mainly on the career of metaphor hypothesis, proposed by Bowdle & Gentner (1999). The career of metaphor hypothesis suggests that the repeated use of a particular base term, as intending a certain metaphorical sense, will result in lexicalization of the metaphoric sense as a secondary meaning to that of the base term, and that this conventionalization process causes the metaphor form preference. In contrast, Chiappe & Kennedy (1999, 2001) have claimed that the metaphor form might be preferred when a comparison is highly apt, because the metaphor form implies that the target will inherit almost all the features of the base term. In other words, the metaphor form implies the category assertion.

In this study, we conducted an experiment to test the career of metaphor hypothesis, based on Bowdle & Gentner's (1999) "in vitro conventionalization". In addition, we observed the interaction between the aptness of comparisons and the repeated presentation of the base terms.

Method

Design Aptness of comparisons (High/Moderate) X The number of repetition of base terms in the study phase (0 /5 times). Both were within subject variables.

Participants Thirty-six undergraduates participated in the experiment. All were native Japanese speakers.

Materials and Procedures The experiment consisted of two phases; the study and the test phase.

For the test phase, we prepared 16 comparisons as the test items in the test phase. Half of the test items were rated as highly apt, and the other half as moderately apt in a preliminary study (M=3.52 and 2.39 on a 5-point scale, respectively). We defined the aptness according to Chiappe & Kennedy (1999). For each comparison, two grammatical forms, a metaphor and a simile, were prepared.

For the study phase, we prepared five target terms for each base of the comparison in the test phase. For example, for the test "An encyclopedia is (like) a goldmine", the new target terms such as {library, book…} were selected. These terms were combined with the base (goldmine) and made up the comparison in simile form. Filler statements were prepared for both phases: 40 comparisons for the study phase, and two category-pairs (e.g., elephant-animal) and two literally-similar-pairs (e.g., lemon-orange) for the test phase.

In the study phase, the participants were presented with the study items in random order and required to write down their interpretation of the comparison in a few words; they were also required to rate the comprehensibility on a 5-point scale. After a five-minute delay, they were asked to rate which grammatical form (metaphor or simile) was more natural or reasonable for each target – base pair, on a 7-point scale.

Results and Discussion

The mean grammatical preference rating for the comparisons in the test phase are shown in Table 1, transformed so that higher numbers indicate a preference for the metaphor form over the simile form. Table 1 shows that previous repetition of the base term increased the participants' preference for the metaphor form. Moreover, a tendency emerged, in that the effect of the repetition differs by the aptness of the comparisons. A 2 X 2 repeated measures analysis of variance on the subject means showed that the main effect of repetition was significant ($F(1,35)=11.52$, $p<.01$). The interaction between repetition and aptness was marginally significant ($F(1, 35)=3.03$, $p<.10$).

Table 1. Mean metaphor form preference ratings (and standard deviations) as a function of repetition

Aptness of comparison	Repetition in Study Phase	
	None	5 times
High	3.13 (.86)	3.77 (1.41)
Moderate	3.17 (.82)	3.35 (1.03)

In summary, these results support the career of metaphor hypothesis. Furthermore, they suggest that the aptness of a comparison promotes the conventionalization of the base term. The implication of this interaction is that the newly created metaphoric meaning is more likely to be lexicalized when it is highly apt, that is, when it has more metaphoric implications.

References

Bowdle, B. F., & Gentner, D. (1999). Metaphor comprehension: From comparison to categorization. In M. Hahn & S. C. Stoness (Eds.), *Proc .of the 21st Annual Conference of the Cognitive Science Society* (pp.90-95). Mahwah, NJ: Lawrence Erlbaum Associates.

Chiappe, D., & Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin and Review*, **6**, 668-676.

Eye-tracking and Simulating the Temporal Dynamics of Categorization

Marissa Nederhouser (mneder@usc.edu)
Department of Psychology, University of Southern California
316 Hedco Neuroscience Building, MC 2520
Los Angeles, CA 90089 USA

Michael Spivey (spivey@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Temporal Dynamics of Categorization

Recent research in categorization has seen a growing emphasis on the temporal dynamics of classification responses (e.g., Lamberts, 1998, 2000; Nosofsky & Palmeri, 1997). These dynamic models generally predict that the degree of fit between an exemplar and the possible categories to which it might belong is a gradually increasing function over hundreds of milliseconds for the correct category and a gradually decreasing function for the incorrect category (or categories).

The development of experimental techniques that can provide evidence for these simultaneously partially-active category representations during the early moments of the categorization process has faced some methodological obstacles, such as imprecision in response deadlines, or limited reaction-time ranges, extensive repetition of stimuli, and potential strategies resulting from speeded classification instructions. The present work recorded eye movements as a semi-continuous, real-time measure of partially activated categories during a normal-speed categorization task (cf. McMurray, Tanenhaus, Aslin, & Spivey, 2003).

Eye Movements During Categorization

Participants were presented with a pair of category bins and given several toy animals sequentially. Eye movements were recorded while they placed the toy animal in one or the other category. We were thus able to calculate an indirect estimate of the moment-by-moment partial activation of the categories being adjudicated among. Figure 1 shows example data (averaged over 17 subjects) for two of the eight critical toy animals used in this experiment.

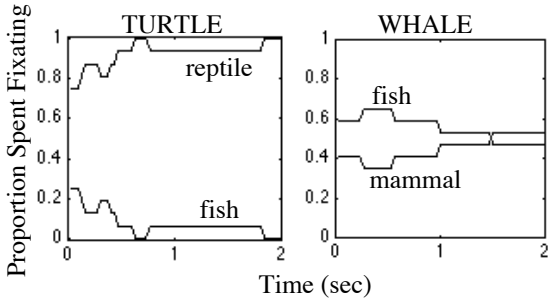


Figure 1: Proportion of time spent fixating two different classification bins while categorizing a toy animal.

Localist Attractor Network Simulations

A simple version of the normalized recurrence competition algorithm (Spivey & Tanenhaus, 1998) was constructed with five feature banks (limb type, environment, blood temperature, oxygen source, birth method) and four taxonomic classes (mammal, reptile, bird, fish). The resulting activation curves over time approximated the eye movement data (compare example items from Figures 1 and 2). Thus, experimental data and network simulations coincide with the general predictions of current temporally dynamic models of categorization (Lamberts, 1998, 2000; Nosofsky & Palmeri, 1997).

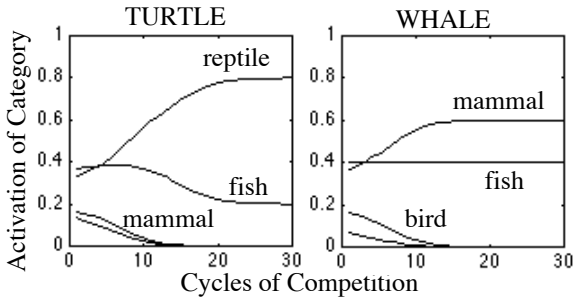


Figure 2: Activation of taxonomic classes as normalized recurrence settles into a stable state.

References

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 695-711.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychol. Review, 107*, 227-260.

McMurray, B., Tanenhaus, M., Aslin, R., & Spivey, M. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research*, *32*, 77-97.

Nosofsky, R. M. & Palmeri, T. J.. (1997). An examplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.

Spivey, M., & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling effects of context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1521-1543.

Finding useful questions in a natural environment

Jonathan D. Nelson (jnelson@cogsci.ucsd.edu)
Cognitive Science Dept., University of California at San Diego
9500 Gilman Dr., Dept. 0515, La Jolla, CA 92093-0515 USA

Identifying useful questions (tests, experiments, or queries) is important for a host of situations, including scientific reasoning, word learning, and vision. If a probabilistic belief model is used to describe an inquirer's knowledge, then each question's usefulness may be calculated using an explicit sampling norm (utility). Prominent sampling norms in psychological literature include Bayesian diagnosticity and log diagnosticity (Good, 1950), information gain (mutual information or Kullback-Liebler distance: Oaksford & Chater, 1994, 1996), probability gain (minimal error: Baron, 1985), and impact (absolute difference: Klayman & Ha, 1987). Strong claims about both the psychological reality and normative basis of particular norms have been made, in papers that calculate only a single sampling norm. Yet a literature review produced no treatment of when the sampling norms disagree with each other, and whether there are theoretical or empirical reasons to prefer a particular norm.

Skov & Sherman (1986) provided an early probabilistic study of information gathering. Participants were told (for instance) that on the planet Vuma 50% of creatures are gloms and 50% are fizos; that 28% of gloms and 32% of fizos wear a hula hoop; and that 10% of gloms and 50% of fizos smoke maple leaves. Given the goal of finding out whether a novel Vumian was a glom or fizo by asking either whether they wear a hula hoop or whether they smoke maple leaves, most participants asked about maple leaves. Skov & Sherman took this as evidence that people are sensitive to diagnosticity. Unfortunately, this result does not show what sampling norm is closest to people's intuitions, because diagnosticity, log diagnosticity, information gain, Kullback-Liebler distance, probability gain, and impact make the same prediction. Other studies have also made claims about particular sampling norms' psychological reality or normative preeminence, without considering other sampling norms.

One frequent task in daily life is to visually ascertain a person's gender. A simplified version of this task (which negates low-resolution information available from outside the center of gaze) is to learn a person's gender by viewing one feature at a time. This task is formally equivalent to the Vuma task. We collected statistics of the gender and features of interest of about 500 passerby, 51% of whom were male, in one natural environment (Table 1). Goals were to determine (1) whether different sampling norms make contradictory claims about what features are most useful, and (2) what sampling norms would best serve in this task.

Results showed that asking about hair length maximizes information gain, Kullback-Liebler distance, probability of correctly identifying the gender, and impact (absolute change in beliefs). Skirt and beard, however, have infinite diagnosticity and log diagnosticity. This is because in the rare event that a person is wearing a skirt or dress, or has a beard or other facial hair, their gender is known with certainty. Using diagnosticity or log diagnosticity to select questions would be inefficient in this environment.

Future work will examine what sampling norms' predictions best match human questions, and whether human questions are sensitive to symmetries and other class-conditional feature dependencies of natural objects.

Table 1: Features' distribution and usefulness.

	Skirt/ dress		Glasses (s=sun)			Beard		Earring		Short hair	
	n	y	n	s	y	y	n	y	n	y	n
% males	100	0	67	6	27	16	84	2	98	93	7
females	98	2	83	3	14	0	100	47	53	7	93
diag.	infinite		1.412			infinite		7.056		13.296	
$\log_{10}$ d.	infinite		0.093			infinite		0.532		1.123	
info.	0.010		0.025			0.084		0.235		0.634	
prob.	0.010		0.065			0.062		0.220		0.420	
impact	0.010		0.080			0.080		0.225		0.430	

Acknowledgment

J. D. N. was funded by NIMH grant 5T32MH020002-05 to the Salk Institute for Neural Computation, UCSD.

References

Baron, J. (1985). *Rationality and Intelligence*. Cambridge: Cambridge University Press.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. New York: Charles Griffin

Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information. *Psychological Review, 94*, 211-228

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608-631.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103(2)*, 381-391.

Skov, R. B. & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Psychology, 22*, 93-121.

Fluency in Categorization

Daniel M. Oppenheimer (Oppenheimer@psych.stanford.edu)
Stanford University, Department of Psychology
Building 420 – Jordan Hall, Stanford, Ca 94305 USA

The meta-cognitive experience of ease of processing, also known as fluency, is a central element of our reasoning repertoire and influences a wide array of judgments. Fluency has been shown to have an effect on a wide array of judgments such as intelligence (Oppenheimer, under review), frequency (Tversky & Kahneman, 1973), familiarity (Monin, 2003), risk for disease (Rothman & Schwarz, 1998) and confidence (Norwick & Epley, 2002).

As the widespread impact of fluency becomes more recognized, researchers have begun accumulating evidence that fluency plays a part in classification. For example Whittlesea and Leboe (2000) have proposed a heuristic model of categorization learning in which fluency plays a central role. In an elegant set of experiments, Whittlesea and Leboe (2000) constructed a set of words that varied in fluency and demonstrated that the fluency of an item had a tremendous impact on categorization judgments.

One potential shortcoming of this set of studies is that Whittlesea and Leboe (2002) restricted their stimuli to artificial words. While this was essential to ensure rigor and avoid confounds, it leaves open the question of what would happen if participants had access to more information than fluency and perceptual similarity of features. When people reason about categories about which they already know a great deal, will fluency still play a role, or is it only used in novel situations when there are few other cues available? This question is the impetus for the current study.

Method, Results, and Discussion

71 Stanford University undergraduates participated as part of a course requirement. Participants rated how good a category member a given exemplar was on a nine-point scale. Four categories (bird, mammal, vehicle, and unusual foods) with 15 exemplars each were used. Exemplars were selected so as to vary in both typicality and commonness.

A standard font manipulation was used to operationalize fluency (Norwick & Epley, 2002). A third of the questionnaires were printed in standard 12 point, Times New Roman font. A reduced fluency condition was created by printing a third of the questionnaires in 10 point, Mistral font. An example of the fonts can be seen in Figure 1.

Times New Roman Font
Mistral Font

Figure 1: Examples of the different fonts used.

Results are summarized in Figure 2. For all categories, participants in the fluent condition rated the exemplars as better category members than participants in the nonfluent conditions. ($t(14) = 4.7$ to 1.5, $p = .000$ to .07).

These results suggest that even in a domain about which individuals know a great deal and likely have pre-experimental notions about what features are relevant to category membership, fluency still plays a significant role.

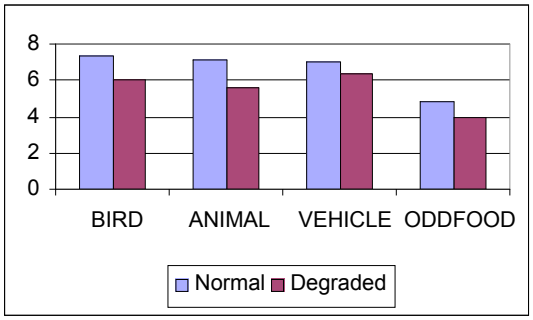


Figure 2: Results of Study 1.

Acknowledgments

This material is based upon work supported under a National Science Foundation Research Fellowship. Thanks to Max Abelev, Norbert Schwarz, Doug Medin, Michael Ramscar, and David Collister for helpful feedback.

References

Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology, 85,* 1035-1048.

Norwick, R.J., & Epley, N. (2003). *Experiential determinants of confidence*. Poster presented at the Society for Personality and Social Psychology, Los Angeles, CA.

Oppenheimer, D.M. (under review). Consequences of Erudite Vernacular Utilized Irrespective of Necessity: Problems with using long words needlessly. *Journal Experimental Social Psychology.*

Rothman, A. J., & Schwarz, N. (1998). Constructing perceptions of vulnerability: Personal relevance and the use of experiential information in health judgments. *Personality and Social Psychology Bulletin, 24*, 1053-1064.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5(2),* 207-232.

Whittlesea, B.W.A. & Leboe, J.P. (2002). The Heuristic Basis of Remembering and Classification: Fluency, Generation, and Resemblance. *Journal of Experimental Psychology: General, 129*, 84-106.

The Memory Consequences of Study after Successful Recall

Philip I. Pavlik (ppavlik@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

John R. Anderson (ja+@cmu.edu)
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Introduction

Our lab has been investigating data and models of spaced memory practice with the long-term goal of applying these models to optimizing the learning of material like vocabulary items. These continuous paired-associate experiments have utilized a recall-or-study trial procedure on both of 2 sessions. On first sessions (Session 1), items are randomized into conditions where they will receive a number practices at various spacing intervals. The memorial consequences of these conditions (distributed continuously across Session 1) are assessed during second sessions (Session 2) in which all of the items are retested several times to determine the effects of the practice by spacing conditions.

The procedure in these experiments was to introduce each paired-associate with an initial 5-second study presentation of the cue-response pair. Subsequent trials were then presented as tests of this knowledge. Because we wanted each trial to count as a single practice in the model, we provided a restudy presentation only when participants responded incorrectly. If the response was correct, we assumed that the correct response constituted a practice of the item. We felt that this recall-or-study procedure resulted in roughly equal practice for each trial.

However, a review of our work suggested that our assumption might not be so uncontroversial. Because of this we designed an experiment where we compared our procedure with a more typical test-and-study procedure where a study opportunity was always presented after a test.

Experiment

The basic procedures for the experiment are described above. The retention interval was 2 days. We looked at our results in terms of both session 1 and session 2 performance. On session 1, we compared recall performance for the two procedures for test trials 2 and 3 (where the effect should be strongest since it had not yet approached ceiling). The first test was excluded because the difference between conditions occurs depending on the success of this test. Means for test 2 and 3 performance were .684 and .639 for the test-and-study and recall-or-study procedures, respectively. This was significant $t = 2.372$, $p < .05$. However, a follow-up conditional analysis suggested that some portion of this effect was merely noise.

Not surprisingly, very little of this benefit persisted into Session 2 in which performance averages were .9 and .883 respectively, and the difference was not significant. Furthermore, session 2 first test results, which were farther from ceiling (*M*s = .760 and .746 respectively for test-and-study and recall-or-study conditions) also showed no significant difference.

Discussion

Subsequent to the experiment an ACT-R (Adaptive Character of Thought – Rational) (Anderson and Lebiere, 1998) model was created using modifications designed to capture the spacing effect described in Pavlik and Anderson (2003). This model captures the small differences in performance by proposing that study trials immediately following successful recall have little effect on long-term memory because the effect of these studies decays more quickly.

The data and model have implications for teaching material such as vocabulary items because they showed that in the typical paired-associate procedure the study trial after a correct recall is redundant and thus inefficient. Further, the data suggest that it is not crucial for models to consider the study after successful recall because its effect is so small. Finally, the model was shown to agree with arguments and data from Kimball and Metcalfe (2003) which proposed a theory of why delayed judgments of learning (JOLs) are more effective than immediate JOLs. The model agrees that this effect, which occurs only when there is no study after the JOL, is not due to enhanced metamemory.

Acknowledgments

Preparation of this poster was supported by grant N00014-96-01-1491 from the Office of Naval Research.

References

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah. NJ. US: Lawrence Erlbaum Associates. Publishers.

Kimball, D. R. & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*, 918-929.

Pavlik Jr., P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In F. Detje, D. Dorner & H. Schaub (Eds.), *Proceedings of the Fifth International Conference of Cognitive Modeling (pp. 177-182).* Germany: Universitats-Verlag Bamberg.

Comparing Acceptability in Magnitude Estimation Tests to an Unsupervised Model of Language Acquisition

Bo Pedersen and Shimon Edelman Zach Solan, D. Horn, E. Ruppin
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
se37@cornell.edu

Faculty of Exact Sciences
Tel Aviv University
Tel Aviv, Israel 69978
{*zsolan,horn,ruppin*}*@post.tau.ac.il*

Traditionally language models have been evaluated by testing their ability to mark sentences as grammatical or ungrammatical. But with the emergence of probabilistic, connectionist models etc. on the computational side and magnitude estimation tests etc., on the linguistic side, it might make sense to go all the way and evaluate the models graded predictions.

We present a language acquisition algorithm that can learn structural regularities from raw data without any prior knowledge about the data. When trained on corpora the extracted language structures can be tested with new sentences to which a graded score is assigned.

Three experiments were conducted. The algorithm was trained on text from the English CHILDES database [MacWhinney and Snow. 1985. The child language exchange system] and then tested on linguistic acceptability data collected by Keller [Keller, Frank. 2000. Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. PhD Thesis, University of Edinburgh] and the algorithm was partially successful on these.

A linguistic acceptability experiment was performed on a large set of well controlled data from an ESL multiple choice (English as Second Language) test and a modest but highly significant correlation with the algorithm score was found.

Finally a linguistic acceptability experiment was performed on sentences generated randomly from a small CFG. 25% of the sentences had 2 neighbor words permuted and another 25% of them had 2 random words from anywhere in the sentence permuted. Both groups got, as expected, significantly lower acceptability score but furthermore the latter had a significantly lower score and a higher variance suggesting that global permutations are more violating but also sometimes by chance get acceptable. The algorithm gives a more clear cut division of the permuted and non-permuted sentences (when trained on similar sentences) but it remains to be investigated whether it can distinguish the two different permutations.

These experiments show that our scoring function is still somewhat unstable and only performs well when variations are small or the data is highly structured as in the CFG experiment. But it also shows that the algorithm is productive under even slightly absurd circumstances like when we train it on CHILDES and test it on the more complex sentences from the ESL data. Furthermore, if we administer the ESL sentences as a multiple choice test the algorithm performs as "intermediate" according to the norms for that test.

What's in a Name?
The effect of sound symbolism on perception of facial attractiveness

Amy Perfors (perfors@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
77 Massachusetts Ave, Rm NE20-388, Cambridge, MA 02139

Introduction

The Saussurean assumption that there is nothing inherent in the relation between a sound pattern and a concept is taken for granted in most of cognitive science. Though the notion that sound-meaning pairings are arbitrary is rarely challenged, there is some evidence indicating that this conjecture may not be wholly true. Sapir (1929) first suggested that cross-linguistically, front and back vowels are robustly associated with specific connotations: front vowels like [i] and [ɪ] are perceived as "smaller" than back vowels like [u]. Other researchers have further explored this idea, documenting that the same association occurs in many languages and cultures (e.g. Ultan 1978; Jakobson 1937). A non-arbitrary sound-meaning relation has also been suggested of some consonants: for instance, Kelly, Leben, and Cohen (2003) suggest that obstruents like [g], [b], and [k] are perceived to be 'hard' and masculine, while sonorants like [l], [n], and [r] are 'soft' and feminine.

Most of these findings, though intriguing, rely on asking subjects what connotations they associate with certain sounds. To date, there is little research that rigorously uses implicit and unconscious measures to study whether sound symbolism is a psychologically real and robust phenomenon. This work does so.

Method and Results

24 photos of men and women paired with names of varying phonology and gender connotation were rated for attractiveness on a 10-point numerical scale on the website *hotornot.com*. Each photo consisted of a frontal shot taken in a naturalistic background; names were saliently located in the upper corner of each photograph. Each photograph was posted multiple times (though never simultaneously) with names that differed systematically in gender connotation, vowel type, and consonant type. As a control, each name was ranked on a 7-point Likert scale by 14 English-speaking subjects based on how much they "liked" it, in general as well as when considered specifically for a male or a female.

Results indicated that phonology played a significant role in perception of facial attractiveness. (see Figure 1). For men, pictures matched with names with front vowels were consistently perceived as more attractive than pictures with back vowels; for women the relation went in the opposite direction ($p<0.01$). Consonants played a smaller but still significant role, but only for women ($p=0.01$).

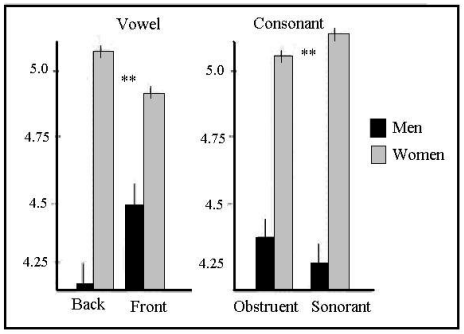


Figure 1: Effect of consonant / vowel type on attractiveness.

Interestingly, names with back vowels were liked less than names with front vowels, but only when the name was considered a guy's name – when the exact same name was considered for a girl, there was no effect of vowel type. (Males: 'back' mean 3.10, 'front' mean 3.59, $F=6.52$, $t=3.627$, $p<0.001$; Females: 'back' mean 2.86, 'front' mean 2.93, $F=2.748$, $t=-0.502$, $p=0.604$). This suggests that although some of the effect of sound symbolism on facial attractiveness may be mediated by how much a name is liked, it cannot be the full story.

Conclusion

This research argues against the Saussurean notion that word-referent associations are completely arbitrary pairings. It suggests that at least under some circumstances, there is a systematic and significant link between some sounds in a language and the semantic associations belonging to words with those sounds.

Acknowledgements

Thanks to Lera Boroditsky, Daniel Casasanto, Jesse Carton, and Lauren Schmidt for helpful comments and suggestions.

References

Jakobson, R. (1937) *Lectures on sound and meaning.* MIT Press: Cambridge, MA

Kelly, B., Leben, W., & Cohen, R. (2003) The meanings of consonants. *Lexicon Branding, Inc.*

Sapir, E. (1929) A study in experimental symbolism. *Journal of Experimental Psychology,* 12:225-239.

Ultan, R. (1978) Size symbolism. In Greenberg, J. (ed) *Universals of Human Language*, vol. 2. Stanford, CA: Stanford University Press.

Task-Set Specific Preparation Prohibits the Expression of Repetition Benefits in Task Switching

Edita Poljac (e.poljac@nici.kun.nl)
Ab de Haan (dehaan@nici.kun.nl)
Gerard P. van Galen (vangalen@nici.kun.nl)
Nijmegen Institute for Cognition and Information, University of Nijmegen,
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

Task Switching: Top-Down or Bottom-Up?

The research on task switching has gained a lot of attention in cognitive psychology in the last decade. The phenomenon observed is the so-called switch cost, which is the decline in performance after a task switch, with the base-line performance being measured on task repetition trials.

Generally speaking, the theoretical interpretations for switch costs can be divided in two groups: top-down and bottom-up interpretations. One of the most prominent top-down interpretations is the reconfiguration theory proposed by Rogers and Monsell (1995), and one of the most recent bottom-up interpretations is the activation theory proposed by Altmann (2004). The reconfiguration approach assumes a functional switching process, with switch costs as an index of this process, while the activation approach assumes a more distributed, general activation of a task representation in memory, with switch cost as a side effect.

The aim of this study was to test the validity of predictions the reconfiguration and the activation theory make about task switching.

Methods and Results

In 2 experiments, the preparation interval duration and the preparation interval type (self-paced vs. externally paced) were manipulated. These manipulations occurred within subjects in Experiment 1 (900 and 200 ms) and between subjects in Experiment 2 (self-paced, 900, 600, 300 and 200 ms). Color and form matching tasks were presented repeatedly in switch and no-switch blocks of 8 trials each. A written cue specified the nature of the upcoming task. The cue appeared at the beginning of a task block and disappeared as soon as a preparation interval was over. No switching between the two tasks occurred within the blocks.

The results showed switch costs, restart costs, and generic performance improvement for longer preparation intervals. A task-switch specific preparation effect (reduction of switch costs with longer preparation intervals) was only observed in Experiment 1.

Conclusions

The data of this study can just partially be explained by the two approaches. On the one hand, our results showed that task-switch specific preparation effect is design dependent. This contradicts the assumption of reconfiguration theory for this effect being robust. On the other hand, the self-paced condition showed switch costs but no restart costs. This observation is at odds with the activation theory, which assumes that the basic processes involved in switch and repeat trials are qualitatively the same.

Therefore, we propose an alternative model of task switching (see Figure 1). The model focuses on processes taking place around the preparation interval, which starts with a cue and lasts until the first imperative stimulus. Generic preparation is the main part of this model, which activates the system if no task switch is required and inhibits this generic activation if a task switch is required. The generic activation compensates for costs accompanied with rule reactivation if the preparation period is sufficiently long. The generic inhibition reduces the chance of making errors but cannot compensate for rule activation costs. Therefore, irrespective of the preparation interval duration, the costs of rule activation become apparent if a task switches.

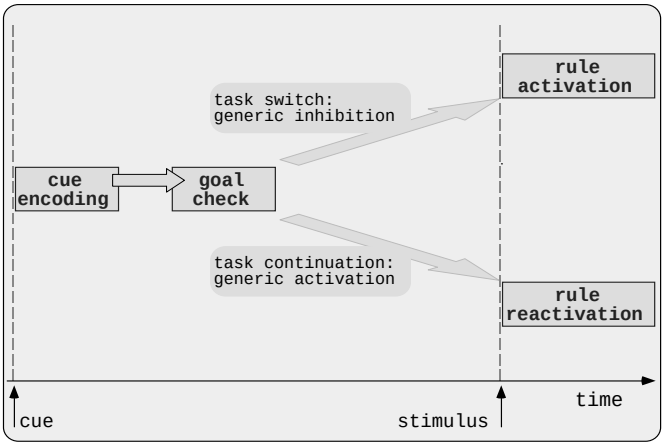


Figure 1: An alternative model of task switching, with generic preparation (activation for task continuation or inhibition for task switch) as its main part.

References

Altmann, E.M. (2004). The preparation effect in task switching: Carryover of SOA. *Memory and Cognition, 32,* 153-163.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207-231.

Localization of cognitive processes using Stroke patients and fMRI

V. Prabhakaran, S.P. Raman, M.R. Grunwald, A. Mahadevia, J.K. Werner, L. E. Philipose, N. Hussain, H.H.Alphs, P. Sun[+], H. Lu[+], B. Biswal*, B.Rypma*, P.C.M. van Zijl[+], A.E. Hillis,

Johns Hopkins University School of Medicine, Rutgers University*, Kennedy Krieger Institute[+]

Localization of cognitive processes to brain regions have mainly utilized the location of infarcted brain regions in stroke patients or fMRI in normal subjects. The BOLD effect in fMRI studies may be difficult to interpret in stroke patients who have areas of hypoperfusion (with resultant reduction in hemodynamic response) due to arterial stenosis. This study was undertaken to examine the influence of hypoperfused regions, in addition to the area of infarct itself, on cognitive processes and fMRI in stroke patients.

Methods

Subjects with subcortical strokes in the left MCA or right-MCA territories, along with normal controls, were imaged while performing a verbal fluency task. The experiments were performed on a 1.5 T whole-body scanner (Philips Medical System, Best, The Netherlands). The study population included six normal participants (3M, 3F, ages 24-57) and six stroke patients (3M, 3F, ages 28-58) with MCA distribution subcortical infarcts. Patients were given a verbal fluency task of 1 min. in duration, compared to rest of 30 secs, organized in an alternating block design, while being scanned with a whole brain fMRI/Stroke MRI-Protocol that included perfusion weighted imaging (PWI) that reveals areas of hypoperfusion as well as structural scans (FLAIR, DWI, T2 sequences)

Results

While normal subjects displayed a left-lateralized fronto-temporal and bilateral cingulo-striatal-thalamic-cerebellar network, the activation pattern of stroke patients was determined both by the hypoperfused regions and/or infarcted areas of the brain. Specifically, the left frontal-temporal network showed diminution of activity in our left MCA patients that had cortical hypoperfusion in the corresponding regions, although their infarcted areas were subcortical.

Table 1:

	Left Middle Cerebral Artery Stroke				**Right Middle Cerebral Artery Stroke**	
Patient	MS	SB	MZ	HS	TG	JH
Gender	Male	Female	Female	Female	Male	Male
Age,Race	58, W	34,AA	50, W	52, W	50, AA	28,W
Infarct **Perfusion Defect Occurence FMRI test**	Minimal left posterior temporo-parietal infarct Posterior temporo-parietal hypoperfusion 04/18/02 11/07/02	Minimal left frontal infarct Frontoparietal hypoperfusion 02/99 10/08/02	Left caudate and Centrum semiovale Fronto-temporo-parietal Hypoperfusion 08/18/01 12/12/03	Left basal ganglia and Centrum semiovale Frontal hypoperfusion 07/00 01/07/03	Right parietal watershed Fronto-temporo-parietal hypoperfusion 1990 02/05/04	Right anterior temporal lobe and basal ganglia Parietal Hypoperfusion 07/19/02 09/30/02
Signs and Symptoms	Impaired word retrieval Impaired sentence comprehension Right upper extremity tingling	Alexia, agraphia Minimal word retrieval difficulty Right-sided weakness	Impaired word retrieval Right upper extremity weakness	Impaired word retrieval Right upper extremity numbness Slurring of speech	**Impaired word retrieval** Right arm numbness Slurring of speech	Left-sided facial droop Slurring of speech Right temporal headache

Conclusions

The observation of a diminished BOLD signal in hypoperfused regions of cortex could either reflect reduced activation in these areas due to tissue dysfunction or reflect normal activation accompanied by increased oxygen extraction without a normal hemodynamic response. The results raise the possibility that localization studies should take into account brain regions that are hypoperfused, as well as infarcted brain regions, in trying to map cognitive processes to brain regions.

References:

Schlosser R, Hutchinson M, Joseffer S, Rusinek H, Saarimaki A, Stevenson J, Dewey SL, and Brodie JD (1998). Functional magnetic resonance imaging of human brain activity in a verbal fluency task. J Neurol Neurosurg Psychiatry, 64(4): 492-8.

Gaillard WD, Sachs BC, Whitnah JR, Ahmad Z, Balsamo LM, Petrella JR, Braniecki SH, McKinney CM, Hunter K, Xu B, Grandin CB (2003). Developmental aspects of language processing: fMRI of verbal fluency in children and adults. Hum Brain Mapp, 18(3):176-85.

Gaillard WD, Hertz-Pannier L, Mott SH, Barnett AS, LeBihan D, Theodore WH (2000). Functional anatomy of cognitive development: fMRI of verbal fluency in children and adults. Neurology, 54(1): 180-5.

Understanding of Principles of Arithmetic with Positive and Negative Numbers

Richard W. Prather
University of Wisconsin
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Martha W. Alibali
University of Wisconsin
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Many models of problem solving include intuitive knowledge components such as principles. Principles are general rules that capture regularities within a domain. Principles reflect conceptual understanding of the underlying structure of a problem domain. For example, in arithmetic, when adding two positive numbers (A + B = X) the answer (X) will always be greater than both operands (A and B) (Dixon, Deets & Bangert, 2001). Problem solvers have been shown to use principles in a variety of problem domains, including counting and arithmetic.

Dixon et al. (2001) investigated participants' understanding of principles that apply to arithmetic operations involving positive numbers. In their study, participants viewed sets of sample problems that had been solved by hypothetical students, and rated the level of understanding that each hypothetical student appeared to have. The analysis compared participants' ratings of problem sets that violated principles and sets that did not violate principles.

The present study built on Dixon et al.'s prior work to investigate participants' understanding of arithmetic operations involving negative numbers. Problem sets were created to test participants' understanding of principles that apply to addition and subtraction with a positive and a negative number, as well as addition and subtraction with positive numbers. The specific principles tested were: (1) *Relationship to Operands*, which specifies the magnitude of the sum or difference relative to the operands, (2*) Direction of Effect*, which specifies how the magnitude of the sum or difference changes as the magnitude of one of the operands is changed, and (3) *Sign*, which specifies the sign of the sum or difference as a function of the relationship between the magnitudes of the operands. As in Dixon et al.'s study, participants rated problem sets that violated principles and sets that did not violate principles. Participants used a scale ranging from 1 (very bad) to 7 (pretty good) to rate the sets. In each case, the relevant analysis compares participants' ratings of violation and nonviolation sets.

As seen in Table 1, participants represented the Direction of Effect principle for operations involving positive numbers and for operations involving negative numbers. Participants represented the Relationship to Operands principle only for addition with positive numbers.

Table 1:
Mean Ratings Provided for Problem Sets
with and without Principle Violations
for Each Principle, Operation, and Number Type

Principle	Operation	No. Type	M Non	M Vio	T
RO	Addition	Positive	3.56	2.89	3.50**
RO	Addition	Negative	3.31	3.15	0.72
RO	Subtraction	Positive	3.11	2.92	0.82
RO	Subtraction	Negative	3.60	3.58	0.10
DE	Addition	Positive	3.89	2.34	6.26**
DE	Addition	Negative	3.06	2.63	2.46*
DE	Subtraction	Positive	3.69	2.53	4.66**
DE	Subtraction	Negative	3.68	2.79	4.75**
Sign	Addition	Negative	2.79	2.81	0.07
Sign	Subtraction	Positive	2.79	2.56	1.17

RO = Relationship to Operands, DE = Direction of Effect, Vio = Violation, Non = Non-violation
* $p < .05$, ** $p < .01$.

The work of Dixon et al (2001) laid a solid foundation for investigating the principles governing arithmetic operations. Our findings replicate some of Dixon et al.'s results, and expand this line of inquiry to negative numbers. Our findings suggest that adults' representations of operations with negative numbers are not as well-established as their representations of operations with positive numbers.

Acknowledgments

This research was funded by an APA Minority Fellowship to Richard Prather. We thank Colleen Moore and Arthur Glenberg for helpful feedback.

Reference

Dixon J.A., Deets, J.K., & Bangert, A. (2001). The representation of the arithmetic operations include functional relationships. *Memory & Cognition, 29*, 462-477.

Part-Whole Statistics Training: Effects on Learning and Cognitive Load

Jodi L. Price (gtg231d@prism.gatech.edu) Richard Catrambone (rc7@prism.gatech.edu)
Georgia Institute of Technology, School of Psychology
654 Cherry Street, Atlanta, GA 30332-0170 USA

Part-Whole Presentations and Cognitive Load

The acquisition of a problem solving procedure is a challenging task often made more difficult by examples or presentation methods that heavily tax working memory and result in the learner being unable to identify and learn the key elements of the example. In general, cognitive load refers to the amount of mental effort required to complete a task within a given time frame (Xie & Salvendy, 2000). Cognitive load theory is based on the observation that working memory capacity is limited. Because of these limitations, cognitive load theory suggests that the methods used to present information should be designed to reduce the demands on working memory (Tindall-Ford, Chandler, & Sweller, 1997) to allow for better processing of examples and ultimately more learning. One technique shown to reduce cognitive load and improve learning is a part-whole (PW) presentation method (Mayer and Chandler, 2001). Mayer and Chandler suggest that initially studying a part (piece by piece) rather than a whole presentation allows the learner to progressively build a coherent mental model of the material without experiencing cognitive overload.

Overview of Experiment

To directly test how a PW presentation would affect cognitive load ratings and skill acquisition in the statistics domain (learning to calculate t-tests and ANOVAs) 84 undergraduate students at the Georgia Institute of Technology studied and completed statistical calculation training and testing materials. The training materials were paper based and consisted of 7 different portions, which contained a brief introduction to statistical calculations and worked examples of how to calculate a t-test and a 2-group ANOVA. Those in the PW condition initially received each of the 7 portions of the training materials one at a time (part) and then were given all 7 portions at the same time (whole). This order was reversed for those in the WP condition.

The test booklet contained three test problems: 2 near transfer problems that were isomorphs to those studied during training and a third far transfer problem that required participants to conduct an ANOVA with three groups.

Participants were asked to rate their cognitive load using the NASA-TLX (NASA Human Performance Research Group, 1987) three times: at the conclusion of the first presentation method (either P or W), at the end of the second presentation method but before testing began, and after they completed the test problems.

Results and Discussion

Contrary to expectations, those who studied the training materials in a PW order performed significantly worse on the test than those who received a WP presentation order, $F(1, 83) = 1.21$, $p = .07$; 4.34, $p < .05$; 4.12, $p < .05$, for the t-test, 2- and 3-group ANOVA problems, respectively. The mean NASA-TLX cognitive load ratings were also found to vary as a function of presentation order with participants in the PW condition rating the part as more difficult than the whole and those in the WP condition reporting the whole more difficult than the part. This yielded a significant main effect of ratings and a significant interaction between ratings and presentation order, $p < .01$ for both. Together these data suggest the PW benefit was not obtained in this experiment but it remains unclear whether this was due to the domain or our implementation of the PW method. Perhaps a paper-based implementation in the domain of statistics is too different from Mayer and Chandler's (2001) multimedia science lesson to obtain the PW benefit. Further research is necessary to tease apart these issues.

Table 1: Variables as a Function of Presentation Order

Presentation Order	Part First			Whole First		
	M	*SD*	N	*M*	*SD*	N
Test Performance (out of 6 possible)						
- T-test	5.38	1.21	41	5.65	1.02	43
- 2-group ANOVA	4.34	1.50	41	4.92	.99	43
- 3-group ANOVA	2.94	1.94	41	3.75	1.74	43
NASA-TLX Cognitive Load Ratings (100=greater load)						
- Part Portion	71.68	11.05	40	64.51	15.35	43
- Whole Portion	63.15	15.09	40	72.20	13.27	43
- Test	62.27	10.65	40	58.84	15.23	43

References

Mayer, R. E. & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of EducationalPsychology, 93* (2), 390-397.

NASA Human Performance Research Group (1987). *Task Load Index (NASA-TLX) v1.0 computerized version.* NASA Ames Research Centre.

Tindall-Ford, S., Chandler, P. & Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied, 3* (4), 257-287.

Xie, B. & Salvendy, G. (2000). Review and reappraisal of modeling and predicting mental workload in single- and multi-task environments. *Work and Stress, 14* (1), 74-99.

Sian Proctor
Patrick Bartshe

Topographic Map Learning Strategies

An important skill of geology is being able to visualize the landscape using contoured topographic maps. This study investigates how students develop topographic map learning strategies, and apply these strategies toward three-dimensional maps. Participants were geology students from an urban university in the Southwest. A topographic map memory test was developed by the authors using Authorware 6.5. One component of the test required participants to study a two-dimensional map, and then select the corresponding three-dimensional map representation from four possible choices. Another component of the test asked participants to describe their strategy for learning the two-dimensional map.

The results indicate differences between participant topographic map learning strategies. For example, participants who used directional terms (for example, North, South, or center) to describe their map learning strategy were more successful in selecting the corresponding three-dimensional map representation then participants who used geological terms (river, mesa, or hill). Gender differences of map learning strategy were also suggested. In conclusion, a better understanding of how students approach the learning of a topographic map is gained, and implications for further research are defined.

Analogy-making as Predication Using Relational Information and LSA Vectors

José Quesada, Walter Kintsch, Praful Mangalath ([quesadaj, wkintsch, praful]@psych.colorado.edu)
Institute of Cognitive Science, University of Colorado, Boulder
Boulder, CO 80309-0344 USA

Current models of analogy comprehension use hand - coded representations. As Hummel and Holyoak (2003) put it, "the problem pf hand-coded representations is among the most serious problems facing computational modeling as a scientific enterprise: All models are sensitive to their representations, so the choice of representation is among the most powerful wild cards at the modeler's disposal" (p. 247). French (2002) reviews different computational models of analogy-making, and points out one of the most fundamental problems of the field: case representations are authored (hand-coded) to make the model work. Between the challenges and future directions he presents "the systematic exploration of experimenter-independent representation-building and learning mechanisms" (p. 204).

In this poster, we propose LSA as a method to generate the much-wanted non-hand-coded representations. However, LSA has severe limitations to represent structure. Turney and Littman (2003) pointed out that the similarity of *semantic relations* between words is not directly reducible to the semantic similarity of individual words. This is also the leitmotiv of some analogy models like Gentner's (1983; 1989).Thus, LSA alone would fail to explain analogy, where relations (structure) between words are fundamental. We use a predication (Kintsch, 2001) to represent structure comparisons in the LSA semantic space. Predication is able to select the features (neighbors) of one component of the analogy (the source) that are relevant to the other (the target).

Table 1(a): a sample SAT question.

Ostrich : bird	
(a)	Lion : Cat
(b)	Goose : Flock
(c)	Ewe : Sheep
(d)	Cub : Bear
(e)	Primate: Monkey

Table 1(b): predication using analogy domains

	Number	Percent	T&L (2003)
Correct	157	41.20%	47.10%
Incorrect	150	40.10%	51.60%
Skipped	67	17.20%	1.30%
Total	374	100%	100.00%
Precision	157/307	0.51%	47.70%
Recall	157/374	0.42%	47.10%
F		0.46%	47.10%

We calculated the predication vectors for all the targets and alternatives of 374 items from the Scholastic aptitude test (SAT). This dataset of analogies was collected by Turney and Littman (2003). An example of a SAT item can be seen in Table 1(a). To calculate the correct alternative, we computed the cosine between the target vector and each alternative, and selected the alternative with the highest cosine. However, this method had poor results: using LSA this way leaves out most of the relational information. For example, relations such as is-a, part-of, causal-agent-of, etc. are all substituted by a very basic semantic distance measure when we compute the cosine between the target and the alternatives. To include this relational information in the comparison, we constructed a set of ten possible relations between the components in the 374 SAT analogies (table 2). Then we computed the cosine between the list of words that define the analogy domain and each analogy predication vector in the dataset. That is, for each analogy we created a vector of ten features, where each feature indicates how similar the analogy is to each of the analogy domains. For example, Ostrich::bird would load primarily in the taxonomy and Hyponymy domain components, but also in endonymy, synonymy, and degree. Then, we correlated these loading vectors for the target and each alternative, and selected the alternative that best correlated with the target to solve the SAT question.

Table 2: Ten analogy domains and their characteristic words

Hyponymy	X is a type of Y (for example - Maple:Tree)
[Subordinate of, superordinate to, rank, class, category, family, genus, variety, type of, kind of, hyponym]	
Degree	X means Y at a certain degree (Pour:Drip)
[level, stage, point, magnitude, extent, greater, lesser, intensity, severity, extreme, degree]	
Meronymy	The parts of X include the Ys (Body:Arm)
[part, whole, component, made up of, portion, contains, constituent, segment, piece of, composite, meronym]	
Taxonomy	X is an item in the category Y (Milk:Beverage)[classification, containing, structure, relationship, hierarchy, system, framework, taxonym]
Synonymy	is the same as Y (Work:Labor)
[equivalent, equal, likeness, match, interchangeable, alike, same as, similar, close to, like, synonym]	
Antonymy	is the opposite of Y (Find:Hide)
[opposite, unlike, different, antithesis, opposed, contradiction, contrast, reverse, anti, not the same as, antonym]	
Characteristic	X is a characteristic of Y (Dishonesty:Liar)
[indicative, representative of, typical of, feature, attribute, trait, property, mannerism, facet, quality, characteristic]	
Plurality	X is many Ys (Throng:People)
[mass, bulk, several, many, lots of, numerous, crowd, group, more, number, plural]	
Endonymy	X entails Y (Coop:Poultry)
[entails, require, evoke, involve, suggest, imply, presuppose, mean]	
Use	X is used to Y (Scissors:Cut)
[do with, manipulate, operate, function, purpose, role, action, utilize, employ, use]	

The results are displayed in Table 1(b). The performance of our model is very close to the state of the art in automatic analogy making when considering correct answers (42% vs. 47%, Turney & Littman, 2003), and precision, recall and F measures. Furthermore, our model is psychologically plausible.

Differentiating the Contextual Interference Effect from the Spacing Effect

Lindsey E. Richland (lengle@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Jason R. Finley (jfinley@ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Robert A. Bjork (rabjork@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095-1563

Interleaving, as opposed to blocking presentations of stimuli sets, can impair learning during training yet enhance retention after a delay or on transfer tasks (Battig, 1972; Shea & Morgan, 1979). Since the initial studies, these effects have been shown in diverse cognitive and motor tasks. These studies have in common that two or three stimuli sets were developed such that materials within each set were distinct yet shared features with the other set(s) (e.g. two ball toss patterns). The similarity was designed to create competition for the learner, such that the learner had to both learn the sequences and distinguish them. Battig (1972) described this competition as the contextual interference effect (CI).

While the CI effect has been widely documented, a natural confound has been integral to the studies. Interleaving materials also introduces spacing between the presentations of each set of learning materials.

The current experiment addresses the relationship between the CI effect and the spacing effect. The spacing effect is one of the most robust cognitive scientific findings (see Dempster, 1990); however, its relationship to contextual interference is less well understood. In this study, foreign language vocabulary words were used to test the prediction that CI is distinct from spacing, and that the CI and spacing effects are additive.

Eighty undergraduates were taught translations of eight English words into both Swahili and Estonian, in a task designed to maximize CI (materials from Pashler, UCSD). Subjects completed six anticipation trials, (prompted generation with feedback) and after a brief delay completed a transfer test. The test required subjects to discriminate between the two languages and present both translations for an English word. During training, the languages were either interleaved (I) or blocked (B) between subjects. The spacing between repetitions of a word was kept constant (CS) across the blocked and half of the interleaved stimuli (7-10 items between repetitions) to isolate the effect of interleaving. Spacing was doubled for the other half of words in the interleaving condition (DS) (15-18 intervening items).

Learning curves and final test accuracies are shown in Figure 1. An interaction between performance at trial six and performance on the transfer test revealed that while accuracy during training was highest for B-CS items and lowest for I-DS items, the opposite was true on the test. Test performance was highest for I-DS items and lowest for B-CS items.

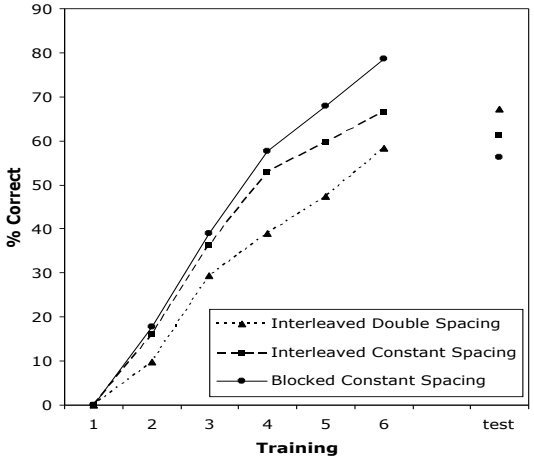


Figure 1. Interleaving and spacing effects during training and on a transfer test.

The data support the prediction that interleaving and spacing are distinct phenomena, and both impair performance during learning yet enhance retention and transfer. Further, the data suggest the effects are additive.

References

Battig, W.F. (1972). Intra-task interference as a source of facilitation in transfer and retention. In R.F. Thompson & J.F. Voss (Eds.), *Topics in learning and performance* (pp. 131±159). New York, NY: Academic Press.

Dempster, F. N. (1990). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627-634.

Shea, J.B., & Morgan, R.L. (1979) Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 179-187.

Acknowledgements

This research was supported by a grant from the Institute of Education Sciences, CASL Grants to Robert Bjork, Award # R305H020113.

Prosodic Feature of Focus in Korean Speech

Jeong Ryu (zjazan@yonsei.ac.kr)
Cognitive Science Program, Yonsei University
134 Shinchon-dong, Seodaemun-gu, Seoul, Korea 120-749

Jae Won Lee (jwlee@yonsei.ac.kr)
Cognitive Science Program, Yonsei University
134 Shinchon-dong, Seodaemun-gu, Seoul, Korea 120-749

Jiwon Chun (occipital@catholic.ac.kr)
Department of Psychology, The Catholic University
43-1 Yokkok 2 dong, Wonmi-gu, Puchon-Shi, Kyonggi-do, Korea, 420-743

Abstract

A focus being defined as the part in a sentence where new information is given, it is assumed that foci in verbal communications show distinctive prosodic features as well as semantic ones. Sentences were given in pairs and each pair contained a question inducing a certain focus and an answer to it. Suprasegmental features were investigated in priority to detailed physical features of separate sounds.

Generally, a narrow focus didn't show any special correlation with stress. An accentual phrase before a focus showed 1onger duration in the ratio of 1:1.5, and 75% of accentual phrases were actualized as intonational ones. It is suggested that a focus in Korean sentences becomes distinct not by being embodied with stress, but by remarking an accentual phrase before it as pause in an intonational phrase, which is quite different from the cases in Indo-European language.

Methods and Analysis

Materials

1. 'What's up?'
2. 'Who give the TV to her mother?'
3. 'What is given by Sumi to her mother?'
4. 'To Whom Sumi give the TV?'

→ '**Sumi** give **the TV** to **her mother**.' (in Korean)

Participants

20 persons, male and female university students who were born in and grew up at Seoul, Korea.

Analysis

- Pitch tracks were made with the Praat program.
- A K-ToBI transcription was made by the author of the redorded sentences. Analyses of break indices were confirmed by other listeners.

Results

In this paper, there is no correlation between focus and accent. If an AP become focus, the AP and a AP in front of the AP is extended. And, in length, a length of the last syllable in the AP is very extended than a total length of the AP. Especially, at the front of focus, it is revealed IP(ex. complex tone etc.)

Therefore, it could be assumed that in stead of accent, focus appeared through breaking speech in Korean.

References

Cho, Yong-Hyung (1998). A Prosodic Labeling System of Intonation Patterns and Prosodic Structures in Korean, *Journal of Speech Science 4(1), 113-133*

Choi, J. W., Lee, M. H. (1999). chojum('focus'in Korean), in Kang, B. M. et. Al. *Formal Semantics Theory and Korean Language Skill, 113-133, Seoul: Hanshinmunhwasa,*

Hansson, Petra (2000). Contraints on Prosodic Phrasing in Spontaneous Speech, Working Papers 48. http://www.ling.lu.se/disseminations/pdf/48/Hansson.pdf

Gundel, J. K. (1994). On Defferent Kinds of Focus, in P. Bosch et. Al., eds., Proceedings of a conference in Celebration of the 10th Anniversary of the Journal of Semantics, 457-466.

Hoskins, Steve "A Phonetic Study of Focus in Intransitive Verb Sentences," 2001. http://www.asel.udel.edu/speech/reports/icslp96b/a421.pdf

Krifka, Manfred (2001). Prosodic Manifestations of Focus, Humbolt Universitywintersmester, 2001. http://amor.az.huberlin.de/~h2816i3x/TopikFokus4.pdf

Jun, Sun-Ah (2000). K-ToBI(Korean ToBI) Labeling Conventions(Version 3.1) http://www.humnet.ucla.edu/humnet/linguistics/people/jun/ktobi/K-tobi.html

Selkirk, Elisabeth (2001). Contrastive *FOCUS* vs. presentational *focus*: Prosodic Evidence from Right Node Raising in English http://www.lpl.univ-aix.fr/sp2002/pdf/selkirk.pdf

Selkirk, Elisabeth (1995). Sentence Prosody: Intonation, Stress and Phrasing" J. Goldsmith ed., *Handbook of Phonological Theory, 550-569, Oxford, UK: Blackwell*

Type/Token Information in Category Learning and Recognition

Yasuaki Sakamoto (yasu@psy.utexas.edu)
Bradley C. Love (love@psy.utexas.edu)
Department of Psychology, The University of Texas at Austin
Austin, TX 78712 USA

Items that violate a salient category regularity are remembered better than items that follow the regularity (Palmeri & Nosofsky, 1995). A memory advantage for violating items is also found in the schema research (e.g., Rojahn & Pettigrew, 1992). Furthermore, work in the schema and category learning research suggests that the memory for inconsistent items is stronger when the violated regularity is more salient (e.g., Rojahn & Pettigrew, 1992; Sakamoto & Love, in press).

In Sakamoto and Love (in press), the salience of a regularity was manipulated by varying the number of items that conformed to it. Category A contained eight items that followed the regularity, whereas category B contained only four. The classification learning procedure encouraged subjects to entertain the rules "If value 1 on the first dimension, then category A" and "If value 2 on the first dimension, then category B." Each category contained an exception item that violated the rule (i.e., the regularity). The category B exception violated the category A rule, whereas the category A exception violated the category B rule. After learning, these exceptions were remembered better than the rule-following items, replicating Palmeri and Nosofsky (1995). Furthermore, following findings from the schema research, memory for the category B exception, which violated the more frequent category A rule, was enhanced (cf., Rojahn & Pettigrew, 1992). While SUSTAIN (Love, Medin, & Gureckis, 2004), a clustering model, correctly predicted these findings, current exemplar and hypothesis-testing models could not.

Type vs. Token

The category A rule-following items were more numerous in two ways. There were not only more rule-following tokens (i.e., instances of the rule) but also more rule-following types (i.e., distinct stimuli) in category A (cf., Barsalou, Huttenlocher, & Lamberts, 1998). Thus, the strength of the category A's regularity was attributable to both more tokens and more types. These two notions of "more" have perfectly co-occurred in the schema literature. The goal of the current research is to test the contributions of types and tokens independently of each other.

When repeating rule-following items from the category containing fewer types equated tokens, the exception that violated a regularity consisting of more rule-following types was remembered better (.86 vs. .65) than the exception that violated a regularity consisting of fewer rule-following types, $t(51) = 3.27$, $p < .01$. Preliminary results from experiments examining the effect of tokens independently of types are mixed across manipulations.

Discussion

The current results demonstrate that when tokens are held constant, items that violate a regularity consisting of many item types are remembered better than items that violate a regularity consisting of only a few item types. Future research will resolve the effect of tokens on recognition of violating items when types are equated. Stronger manipulations are currently being examined that avoid contrastive categories often used in category learning research. Work along these lines will illuminate future schema and category learning research and will advance our understanding of how humans represent rules, exceptions, and type/token information.

References

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, *36*, 203–272.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Human Category Learning. *Psychological Review 111*, 309–332.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 548–568.

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology*, *31*, 81–109.

Sakamoto, Y., & Love, B. C. (in press). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*.

The Leverage of a Self Concept in Incremental Learning

Alexei V. Samsonovich (asamsono@gmu.edu)
Krasnow Institute for Advanced Study, George Mason University
4400 University Dr. MS 2A1, Fairfax, VA 22030 USA

Kenneth A. De Jong (kdejong@gmu.edu)
Department of Computer Science and Krasnow Institute for Advanced Study
George Mason University, 4400 University Dr., Fairfax, VA 22030 USA

Introduction

This work is an attempt to bring together three topics that belong to three different levels of science: (1) symbolic unsupervised learning, (2) the self of a cognitive system, and (3) a universal criterion for conscious experience. We study cases when a cognitive system develops new abilities by reinterpreting its own episodic memories, using the self-concept and the schema of a motivated voluntary action. Within our framework (Samsonovich & DeJong, 2003), the subject-self per se is not represented as a virtual entity in the cognitive system. Instead, the set of axioms that constitute a self-concept (Aleksander & Dunmall, 2003; Samsonovich & Nadel, in press) are implemented via dynamical rules and constraints. These principles are demonstrated in a model paradigm and have implications for the philosophy of mind.

Approach

The proposed approach is based on the general framework of schemas, charts and mental states described previously (Samsonovich & DeJong, 2003). The term "schema" was introduced by Kant (1781/1929). Here schemas are units of semantic knowledge, primitives of action, reasoning, sensation, etc. A schema has a header (specifying rules and conditions of binding and expected effects of execution) and a body (specifying how, if at all, the schema is executed).

Paradigm: Leveraging Self-Learning with the Self

In this paradigm, a set of specially designed virtual worlds is used as a "training facility" to help the virtual robot to develop useful and powerful schemas. Innate schemas may include elementary moves and senses, as well as relevant reasoning primitives. The robot "wakes up" in a first-level world and starts by repeating the following procedure:

1. Select an action schema and mutate its header to produce an idea of an action that is not straightforward.
2. Take the new header as a challenge and solve it in each of several encountered situations (execute the solutions).
3. Reinterpret own behavior: find an apparent common motivation in the performed intermediate steps in all cases.
4. Based on the above, write the body of the new schema and add the schema to semantic memory.

As the robot learns essentials at the first level, it is taken to the next level, and so on. At each new stage, previously developed schemas are used for solving new challenges.

Demonstration by Example

The scheme outlined above will be demonstrated in the poster by computer simulations based on a push-push puzzle setup. A minimal set of innate schemas includes a one-step move and some useful cognitive primitives, e.g., the notion of Euclidean distance. At the first stage the robot learns to move in an open space. Then it learns to navigate a maze, to push blocks, to avoid irreversible moves, etc. After that, when given a goal, it is capable of solving simple puzzle configurations and learns to deal with more complex ones.

Philosophical Implications

The above analysis has interesting implications for the philosophy of mind. Some philosophers believe that the phenomenon of conscious experience will always remain a mystery, while others maintain that this mystery is illusory. How can one decide, when and whether this phenomenon should occur? Chalmers (1994) answers with the Principle of Organizational Invariance. His answer sounds like this: *there is an abstract mathematical model M of a functional organization of a cognitive system, such that, whenever M can be mapped onto a given physical object, that object is conscious.* Therefore, a criterion for consciousness can be given in terms of *M*. We define it as follows: *M* must instantiate the self as a noumenon (a notion introduced by Kant, 1781/1929), which we understand as an imaginary thing that seems to determine system's dynamics, and yet it cannot be explicitly represented in the system due to its fundamental properties. The proposed framework in which the self is implemented via a set of self-axioms (constraints) about system's own dynamics conforms to this concept.

References

Aleksander, I., & Dunmall, B. (2003). Necessary first-person axioms of neuroconsciousness. *Lecture Notes in Computer Science,* 2686: 630-637.

Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines*, 4, 391-402.

Kant, I. (1781/1929) *Critique of pure reason.* Translated by N. K. Smith. New York: St. Martin's Press.

Samsonovich, A. V., & De Jong, K. A. (2003). Meta-cognitive architecture for team agents. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (pp. 1029-1034). Boston, MA: Cognitive Science Society.

Individualization as influencing semantic alignment in mathematical word problem solving

Emmanuel Sander (emmanuel.sander@univ-paris8.fr)
University Paris 8, Department of Psychology, 2 Rue de la Liberte
Saint-Denis, 93526 France

Nadege Mathieu (nadege.mathieu@wanadoo.fr)
University Paris 8, Department of Psychology, 2 Rue de la Liberte
Saint-Denis, 93526 France

Several previous works from Bassok and colleagues (e.g. Bassok, Wu & Olseth, 1995) put in evidence that, when solving a mathematical word problem, content is used to interpret structure: surface features are used as semantic cues in order to induce an interpretative structure that the participant will rely on in order to solve the problem. For instance, a problem involving doctors choosing other doctors is likely to provoke the inducement of a symetrical structure wereas a problem involving secretaries choosing computers is likely to provoke the inducement of an asymetrical structure (Bassok, Wu & Olseth, 1995). These structures interfere with the mathematical ones and influence problem difficulty, solving procedures and analogical transfer. Bassok (2001) considered this phenomenon as a special case of the cognitive mechanism of structural alignment (Markman & Gentner, 1993) and refered to it as semantic alignment. Two dimensions, namely symmetry-asymmetry and continuity-discreteness, were identified by Bassok and colleagues as influencing semantic alignment for a large range of problems and identification of other dimensions is important for a better understanding of the phenomenon and it's range of application. We conducted 2 experiments in order to show that Individualization (I)-Non Individualization (NI) is also a relevant dimension.

In the first experiment, 80 undergraduate students were equally split among I and NI conditions and solved combinatorial problems in two contexts. For instance, one NI version involved four children choosing one after the other one strawberry among twelve strawberries wheras the I version was identical except that strawberries were replaced by explicitly individualized cakes (a cheese cake, an apple pie, a chocolate cake…). We found significant effect of the condition (Table 1). In the NI condition, students proposed significantly more partitive solutions (e.g. 12/4) which did not require individualizing objects such as sharing 12 objects among 4 people, than in the I condition. The reverse was true for the multiplicative solutions (e.g. 12x4) which required individualization, such as each child having 12 choices. With the same experimental design, we conducted a second experiment including two problems, the 'car problem' and the 'grocery problem' in which, contrary to combinatorial problems, individualization was not a relevant dimension in the mathematical structure.

Table 1: Rates of procedures used by participants (exp. 1)

Procedure	Indiv	Non Indiv
Correct	8%	10%
Partitive	25%	51% *
Multiplicative	28%	4% *
Other	39%	35%

We found again significant effects: Table 2 sums up the results of the 'car problem': John buys a car 10,000 Euros and sells it 12,000 Euros. He buys it back 14,000 Euros and sells it again 16,000 Euros (NI) and John buys a red car 10,000 Euros and sells it 12,000 Euros. He buys a black car 14,000 Euros and sells it again 16,000 Euros (I). We found also significant effect of individualization for the second problem (NI condition involved two kinds of figs wheras I condition involved figs and dates).

Individual protocol analyses confirmed that participants induced different structures depending on the condition. For instance, most of the participants who found 2000 in the 'car problem' considered that there was a loss of 2000 due to the buying of the same object sold for 2000 less. Those results encourage to carry out more extensive studies concerning the influence of individualization in problem solving.

Table 2: Rates of results obtained by participants (exp. 2)

Results	Indiv	Non Indiv
4000	53%	34% *
0	29%	12% *
2000	12%	46% *
Other	6%	8%

References

Bassok, M., Wu, L.L., & Olseth, K.L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory and Cognition, 23*, 354-367.

Bassok, M. (2001). Semantic alignments in mathematical word problems. In Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.) *The analogical mind: Perspectives from cognitive science.* Cambridge, MA: MIT Press.

Markman, A.M., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431-467.

Evidence from an fMRI Experiment for the Minimal Encoding and Subsequent Substantiation of Predictive Inferences[1]

Franz Schmalhofer, Markus Raabe, Uwe Friese, Karin Pietruska and Roland Rutschmann
FirstName.LastName@uos.de and Roland.Rutschmann@Psychologie.Uni-Regensburg.de
Institute of Cognitive Science, University of Osnabrueck
49069 Osnabrueck Germany

Abstract

In an event-related fMRI-experiment reading during a supposed inference generation period was compared to explicit reading. In a subsequent verification task, the verification of inference and explicit statements were compared. The results show systematic but minimal inference processing during encoding (BA 9) and supplementary activities of text related processes, additional inferencing, and situational elaborations when verifying inference as compared to explicit statements.

Introduction

Which brain areas are involved in predictions during reading and which areas in the subsequent utilization of predictive inferences in a verification task? While predictive inferences are supposedly only represented as part of the situation model, the sentences of a text are additionally encoded as text information (Schmalhofer et al., 2002). Previous fMRI experiments have shown involvement of the prefrontal cortex in establishing text coherence (Ferstl & von Cramon, 2001) and in generating inferential bridges (Mason & Just, 2004).

Experiment

In an event-related fMRI-experiment we investigated 1) the reading during a supposed inference generation period (see Table 1, words 13-18) in comparison to explicit reading and 2) the subsequent verification of the respective statement. (e.g. "wine spilled"). Four versions of texts were constructed so that the same statement constituted an explicit, a paraphrase, an inference or an incorrect statement. The collected data were analyzed by SPM2.

Table 1: Sample text material and test statement

Title: Air Travel
Words 1-12 (all conditions): While the flight attendant served the passenger a full glass of wine
Words 13-18 (explicit): turbulence caused the wine to spill.
Words 13-18 (paraphrase): turbulence caused the wine to splash.
Words 13-18 (inference): turbulence occurred which was very severe.
Words 13-18 (control) : the plane was at cruising altitude.
Test statement: wine spilled

13 students from the University of Osnabrueck participated. 108 reading passages and subsequent tasks were presented: The 4 experimental conditions with 18 trials each as well as 18 filler and 18 non-word trials. A trial lasted 27 seconds. Data were collected by a 1.5 Tesla Siemens Sonata scanner.

Results and Discussion

During reading of words 13-18, the comparison between the inference versus the explicit condition showed the medial frontal gyrus (L + R BA 9, 63 Voxels) to be active. For the statement verification, the contrast between the inference and explicit conditions showed three clusters predominantly in the left prefrontal cortex to be differentially active, as can be seen from Figure 1. For more details see Table 2.

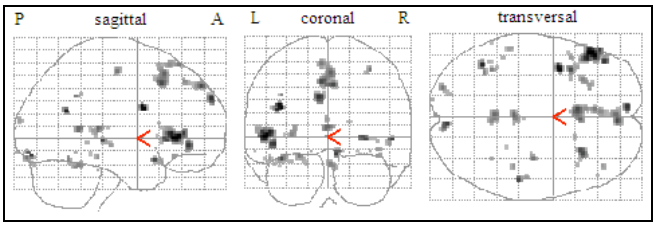


Figure 1: Statement Verification: Inference versus explicit

This experiment confirms previous behavioral results showing minimal predictive inferencing during encoding. It involves the medial frontal gyrus (bilateral). During verification, areas which can be attributed to semantics, inferencing and situational elaborations were observed.

Table 2: Statement Verification: Inference versus explicit

Location of Activated Areas	Z	Voxels
1. Cingulate Gyrus / Superior Frontal Gyrus Medial Frontal Gyrus (L + R BA 8; L BA 6, 32)	3.8	125
2. Medial Frontal Gyrus (L + R BA 9; L BA 6, 8)	3.8	67
3. Inferior Frontal Gyrus (L BA 10, 44, 45, 46, 47)	4.2	168

References

Ferstl, E. C., & von Cramon D. Y. (2001). The role of coherency and cohesion in text comprehension: An event-related fMRI study. *Cognitive Brain Research*, *11*, 325-340.

Mason, R. A., & Just, M. A. (2004). How the brain processes causal inferences in text. *Psychological Science, 15* (1), 1-7.

Schmalhofer, F., McDaniel, M. A., & Keefe, D. (2002). A unified model for predictive and bridging inferences. *Discourse Processes*, *33*(2), 105-132.

[1] Supported by Alexander von Humboldt Foundation, grant III – TCVERL-Deu/1075454. We thank Evelyn Ferstl and Charles A. Perfetti.

Consistent Argument-Predicate Binding Is Important for Predicate-Predicate Linking

Adam Sheya (aasheya@indiana.edu), Rima Hanania (rhanania@indiana.edu)
Department of Psychology and Cognitive Science, Indiana University

Hilmi Demir (hdemir@indiana.edu)
Department of Philosophy and Cognitive Science, Indiana University

The fundamental process of connecting instances to each other is essential to many types of learning: from generalization over instances, to category learning, to learning from analogies. The present work seeks an understanding of these processes by studying how adults learn about relations. Learning about relations requires learning about two kinds of entities: arguments and predicates.

Gentner (2003) proposes that arguments, and particularly arguments that take the form of concrete objects, are psychologically prior to predicates. Further, she has shown that object-object similarities play a key role in the relational mappings that both children and adults make. This suggests that when learners are presented with a set of instances in the form of arguments and predicates, the similarity among arguments may be more important than among predicates in connecting learning instances to each other.

However both Gentner (2003) and Billman and Knutson (1996) have also suggested that systematicity of predicates is important. More specifically, Billman and Knutson propose that what is important when learning is how many cues are systematically predictive of the categories. All cues – arguments and predicates – can contribute to systematicity with the critical issue being the degree to which cues are mutually predictive. Thus, it may be the systematicity relations across a set of instances and not specifically argument and predicate similarity that guides learning.

The present experiment uses a learning task in which object categories are defined by the relational roles of the objects and not by their properties. These relational categories have high systematicity: knowing that object X is in relation P to object Y determines both what other relation X enters into and the relational roles of all other objects. In order to learn this, learners must link one relation to another. In these experiments, we manipulate argument similarity and the systematicity of argument-predicate links; keeping predicate systematicity high and constant.

Design

The experiment consisted of a training and test phase. On each trial there were three objects: two actors (A_1, A_2) and one receiver (R). The actors each performed two actions relative to the receiver (e.g. A_1 might "jump over" R and also circle R). On each trial the actions that define A_1 and A_2 did not change. Participants were assigned to one of three training conditions: (1) low argument similarity (different objects each trial), (2) high argument similarity (same objects each trial) and (3) high argument similarity but low argument-predicate systematicity (same objects but different roles on each trial).

In the test trials, new object triads were used that were not superficially similar to the training objects. On each trial the experimenter demonstrated one of the actions for A_1 or A_2. Since the predicates (actions) are systematically related, if the argument-predicate structure has been learned then participants should infer the correct object and predicate pairs from this single cue. In order to measure learning, participants were asked to perform the demonstration object's second action and the two actions of the other actor.

Results and Discussion

A test trial was scored as correct if the actions were paired correctly and the correct receiver was used for every action. Participants failed to learn the argument-predicate structure in the low argument similarity condition (Mean percent of trials correct=16%) and in the high argument similarity and low argument-predicate systematicity condition (M=22%), but they did learn the argument-predicate structure in the high argument similarity condition with high argument-predicate systematicity (M=65%). Our results indicate that systematicity matters in learning. However systematicity of predicates alone is insufficient because this was present in all conditions. The systematicity that was crucial for learning in this case was the systematicity between arguments and predicates. This type may be critical to the learning process because it facilitates the linking of distinct temporal events. The arguments may thus serve as the indexes in working memory that bind one instance to another and thus enable learning across them. The next question is whether objects or arguments in general are privileged in this role or whether any common index to all learning instances would do.

References

Billman, D. & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 458-475.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Golden-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195-235). Cambridge, MA: MIT Press.

Is Color Photography Flatter:
The Difference of Depth Perception between Chromatic and Achromatic Photos

Suejin Shin (sjshin@yonsei.ac.kr)
Center for Cognitive Science, Yonsei University
134 Shinchondong, Seodaemungu, Seoul, KOREA

Introduction

Expression of depth is one of the most important factors in constructing the creative vision of a photographer. Photographers usually know how to create, with certain artistic sense or taste, depth impression in the process of flattening objects in a three-dimensional space. To emphasis the feeling of depth, photographers tend to choose black and white processing. This is because there has been a myth: color photography is flatter than black and white. From the era of early color photography, pleasure of flatness has been pursued rather than feeling of depth(Newhall, 1982).

The aperture and focal length of lenses are the major factors in changing depth impression(London et al, 2002). Depth impression increases when the aperture size is enlarged, resulting in shallower depth of field. When the focal length of lens becomes longer making the discrepancy of relative size among subjects smaller, depth impression decreases (Shin, 2002). This study is performed to investigate the effects of the presence of color as a photographic technique on depth perception.

Method

Fifty (N=50) Yonsei University undergraduates and graduate students were assigned two experiments. The stimuli were taken in color using two techniques: (a) aperture; f/2.8, f/5.6, f/11, f/22, and (b) focal length of lens; 28mm, 50mm, 70mm, 105mm, which were then duplicated and transformed to gray scale images. The conditions were consistently maintained to reveal the effects of the specific techniques. Every photograph included two same-sized mannequins positioned at different distances from camera. Participants were asked to compare two photographs at a glance, and were forced to identify the photograph in which the two mannequins appear to be closer to each other.

Results and Discussion

The frequency was analyzed by the method of paired comparison(Thurstone1927a; 1927b). In the results(Figure 1, 2), the smaller aperture size and the longer focal length decreased depth perception in both chromatic and achromatic images. Color severely decreased depth impression at variations of focal length. In the mean time, the flattening effect of color was relatively weak along aperture variations. This implies that spatial frequency is a strong factor in giving a feeling of depth even in color photographs.

Black and white process is still common in the field of photography as fine art. From the results of this work, the relatively intense feeling of depth can be one of the reasons for that.

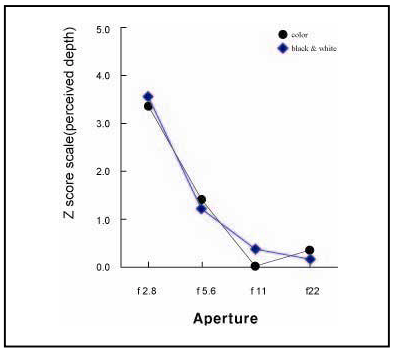


Figure 1: The effect of aperture on depth perception.

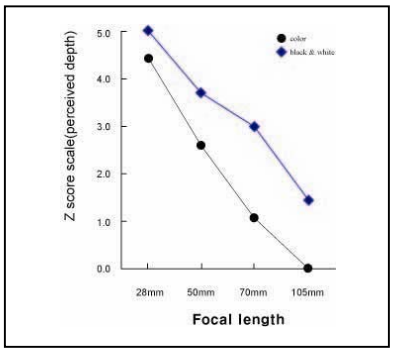


Figure 2: The effect of focal length on depth perception.

Acknowledgments

This work was supported by Korea Research Foundation Grant (KRF-2002-074-HS1003).

References

London, B. et al.. (2002). *Photography, 7th ed.* NJ: Pearson Education Inc..

Newhall B. (1982). *The History of Photography*. New York: Museum of Modern Art.

Shin, S. (2002). The Effect of Aperture and Focal-length on Depth Perception. *AURA, Vol. 9*, 122-129.

Thurstone, L. L. (1927a). A Law of comparative judgment. *Psychological review, 34*, 273-286.

Thurstone, L. L. (1927b). Psychological analysis. *American journal of psychology, 38*, 368-389.

Learning through verbalization (1): Understanding the concept of probability

Hajime Shirouzu, Naomi Miyake, ({nmiyake; shirouzu}@sccs.chukyo-u.ac.jp)
& Hitoshi Izumori (izumori@ra2.so-net.ne.jp)
School of Computer and Cognitive Sciences, Chukyo University
101 Tokodate, Kaizu-Cho, Toyota, 470-0393 JAPAN

Schematizing experiences plays a critical role in learning. Verbalizing experiences at a proper abstraction level has been identified as important for effective schematization (Shirouzu *et al.*, 2002), or learning (Chi *et al.,* 1989), but its details need to be studied further. For instance, in a statistics class, a dramatic demonstration can help students grasp basic concepts like the law of large numbers. Though the students remember them well, what they could verbalize differs depending on the class activities and has different effects on learning. We report here that the students asked to verbalize a demo one hour after could express the important aspects of the event 18% more than their counterparts who did the same twelve weeks later.

Learning the concept of probability

When asked what it means that "The probability of getting ONE pip when you roll a die is one-sixth," it is not rare that even a college student answers that you get ONE once per six rolls of a die. To change this misconception a curriculum was devised. In Activity 1 each student rolled a die 50 to 100 times, counted each pip, and the class tallied the results to yield a histogram of over 3000 trials. This was followed by Activity 2 using a deformed die, with four sides of 1.5 lengths of the other two. Each student rolled the die 200 times, checked the probability of appearances of the pips of ONE and SIX (on shorter sides). Then the class collected all the data to histogram them. The comparison of these two patterns aims to clarify the relationship between the probability and the likelihood of event occurrences, based on the law of large numbers.

Comparison of two classes

Using the curriculum, two undergraduate classes in cognitive science dept. were taught the concept of probability. Two classes were organized differently to compare the timing effect of abstracting the experiences. While Class 1 emphasized teacher-guided abstraction, Class 2, taught by the same teacher, focused more on the students' own verbalization. In Class 1, the teacher explained the law of large numbers, had the students engage in Activity 1. One week later, he showed to the class the histogram of all the data, explained the law, and engaged the class in Activity 2. The results were tallied three times, for 20, 200, and 1800 trials. The students were only explicitly requested to verbalize the meaning of their experiences twelve weeks later, at the term examination. In Class 2, the class did Activities 1 and 2 consecutively in one day (in two classes), without teacher's explanation of the law. Explicit verbalization was requested at the end of the class, in the form of revisiting the starting question. The students had a chance to discuss among themselves.

Results

The verbal reports of the two classes were categorized in terms of their degrees of abstraction. The reports of category "High" refer to the meaning of the law; "If you roll the die infinitely, the ratio of getting the pip ONE approaches 1/6." "Moderate" reports mention the effect of large numbers; "You get the pip ONE roughly 1/6 times if you roll the die many, many times." "Concrete" reports may refer to their class size as an example of a large number, but not its effects. "Others" often include their previous knowledge about the probability, "The pip ONE occurs 1/6 times because it is one of the equally possible six events." Table 1 summarizes the results.

Table 1: Abstraction levels of verbal reports

	Ratio of answerers	Answer abstraction levels			Other
		High	Moderate	Concrete	
1	85.3%	5.3%	16.0%	45.3%	18.7%
2	92.1%	0%	39.4%	10.5%	42.1%

In Class 2, more students moderately abstracted their experiences than concretely. Lacking the chance to do the same, some 45% of the students in Class 1 reverted to the concrete level answers when tested. In order to bridge concrete experience with abstraction, the "moderate" expressions may play an important role.

Requiring students only to verbalize from memory may have had them focus on resultant pattern of 1/6, bringing them back to their previous "common sense" from textbooks. There seems to be certain duration of time to properly ponder on the exact cause and effect of the "surprising" phenomena, to be able to scrutinize their newness carefully enough to be able to generalize them.

Acknowledgments

This research is supported by JPS to the 1st author and by MECSST and CREST/JST to the 2nd author.

References

Chi, M., et al. (1989). Self-explanations. *Cognitive Science, 13,* 145-182.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002), Cognitively active externalization for situated reflection. *Cognitive Science, 26,* 469-501.

Basic Questioning Strategies for Making Sense of a Surprise: The Roles of Training, Experience, and Expertise

Winston R. Sieck (sieck@decisionmaking.com)
Deborah A. Peluso (debbie@decisionmaking.com)
Jennifer Smith (jsmith@decisionmaking.com)
Danyele Harris-Thompson (dharris@decisionmaking.com)
Klein Associates Inc.
1750 Commerce Center Blvd. North, Fairborn, OH 45324 USA

Information operations (IO) specialists are like US political strategists in foreign lands, and they are concerned with affecting others' decision processes. In order to be effective, IO practitioners must be able to efficiently develop an understanding, frame or theory (i.e. "make sense") about how decisions are made in a particular locale. As in scientific reasoning, when IO specialists observe surprising events, they have an opportunity to dramatically improve their frames (cf. Dunbar, 1995). But to capitalize on such opportunities, they must acquire the skill to ask good questions; questions that admit to a basic lack of understanding or that specifically challenge their frames. In the current study, we ask, "What roles do training and field experience have in acquiring skills for questioning one's frames?"

Method

Participants (n=60) were either laypeople (L) with no military background, novices (N) who were trained in IO, or individuals who had training and IO field experience (F). Of the latter group, 4 were identified as IO "experts" (E) via peer nomination. Participants were presented with a 1-page scenario describing a real situation that had occurred in Kosovo, and that was obtained earlier from an IO expert by CTA elicitation. The synopsis was that buses with armed escorts were used to transport Serb college students to school from their family's enclaves. The regional commander made plans to reduce the escort due to costs. An IO campaign was conducted to convince the *students* that the buses would still be safe. However, once the escort was reduced, the vast majority of students quit riding the bus. This was quite a surprise to US personnel on the scene. The reason as eventually discovered was that, unlike in the US, the Serb *mothers* made the ride/no ride decision for the students. This reason was not disclosed to participants. Instead, they were asked to explain their understanding of the situation in a think-aloud procedure, as well as what they would want to know to inform their understanding.

Results

The protocols were coded for key kinds of inquiries participants made, in particular, inquiries that would lead directly to developing an accurate understanding of the scenario. The two key inquiry types are: "Why are the students not riding?" and "Is someone else influencing the student's decision?" The proportions of participants who asked each of these key decisions by experience level are presented in Table 1. As shown, participants with field experience were 3 to 4 times more likely to ask one of these critical questions than were those with no field experience (trained or not), $\chi^2(1) = 5.31$, $p = .02$ for the "why" question, and $\chi^2(1) = 5.31$, $p = .02$ for the "who" question. The results were not due to the experienced participants simply "knowing" the answer. Only 3 participants hypothesized the correct answer (coded liberally as "family decides" is the reason). Also, accuracy did not depend on field experience, $\chi^2(1) = 0.04$.

Table 1: Proportion who ask each question

	Experience Level			
Key Inquiries	L	N	F	E
"Why not ride?"	.10	.04	.23	.50
"Who decides?"	.05	.09	.23	.50

Discussion

Experienced IO practitioners were much more likely to question important aspects of their frames than laypeople and trained novices. At one level, the kinds of questions they asked were quite "basic," lacking the obvious technical sophistication that might often be assumed to be associated with experience and expertise. Nevertheless, these simple questions were exactly the kind needed to develop a useful frame on which to base decisions and actions, and are quite similar to questioning strategies of experienced scientists. Indeed, the current study represents an early step toward extension of work in scientific reasoning, situation assessment, judgment and other areas to a broader collective higher-order cognitive topic of "sensemaking" (Klein et al. 2004).

Acknowledgments

This research was supported by the Army Research Institute for the Behavioral and Social Sciences (Contract 1435-01-01-CT-3116).

References

Dunbar, K. (1995). How scientists really reason: scientific reasoning in real-world laboratories. In R. J. Sternberg, & J. E. Davidson (Eds.), *The nature of insight*. Cambridge, MA: MIT Press.

Klein, G., Phillips, J., Rall, E. A., & Peluso, D. A. (2004). A Data/Frame model of sensemaking. To appear in R. Hoffman (Ed.), *Expertise out of context*.

The Effects of Working Memory Load on Transitive Inference

Cynthia M. Sifonis (sifonis@oakland.edu)
Department of Psychology, Oakland University
Rochester, MI 48309 USA

William B. Levy (wbl@virginia.edu)
Department of Neurological Surgery, University of Virginia Hlth Sys
Box 800420-Neurosurgery., Charlottesville, VA 22909-0420 USA

Introduction

Cognitive scientists are interested in the brain mechanisms supporting transitive inference (TI) reasoning. In a TI task, participants learn correct responses to premise pairs (A>B, B>C, C>D, D>E) from which they can infer a relational hierarchy (A>B>C>D>E). The hierarchy allows them to respond appropriately to novel pairings of non-adjacent members of the hierarchy (B>D). It has been demonstrated that both the hippocampus and the prefrontal cortex, with its ability to support working memory (WM) (Barrouillet, 1996) are necessary for successful TI in humans. However, the relationship between WM and TI has never been directly tested.

In the current experiment we examine the effect of a working memory load on TI. We load working memory in two ways – by preventing rehearsal with a math task and by using unfamiliar stimuli, making it more difficult to encode the information in WM. We hypothesized that loading working memory would impair participants ability to learn the premise pairs, their ability to engage in TI, and the degree to which they are aware of the TI hierarchy.

Methods

Stimuli and Procedure

During training blocks, Oakland University students learned the correct response to the premise pairs. The stimuli in the **Unfamiliar** conditions consisted of five hiragana characters. Those in the **Familiar** conditions consisted of five familiar shapes (See Figure 1). Following each premise pair, participants in the **Math** conditions were presented with a subtraction task. Feedback as to premise pair and subtraction task accuracy was provided for each of the 40 training trials in each training block.

Following completion of each training block, participants were tested on their discrimination of the four premise pairs, and the transitive inference pair (BD). Feedback was not provided during the testing blocks. If participants reached 80% accuracy on the premise pairs during testing, training ended. Otherwise, they continued on to the next training block. This continued until criterion performance was reached or four testing/training blocks were completed. After testing, all participants were given a questionnaire assessing their awareness of the hierarchical relationship between pairs.

Results

There were significant main effects of math task, $F(1,141) = 6.12$, $p < .05$, and familiarity, $F(1,141) = 8.14$, $p < .01$, on avg. premise pair performance during testing. Participants given the math task performed less accurately (72% correct) on the premise pairs than those not given a math task (77%). Those given unfamiliar stimuli performed less accurately (76%) than those given familiar stimuli (83%). There was a marginal main effect of familiarity, $p < .10$, on average TI performance during testing. Those given unfamiliar stimuli were less accurate on the BD pairs (68%) than those given familiar stimuli (75%). There was a marginal effect of familiarity on awareness of the TI hierarchy, $X^2 = 3.09$, $p = .10$. Participants given unfamiliar stimuli were more likely than those given familiar stimuli to engage in TI while reporting no conscious awareness of the hierarchy.

Discussion

Our predictions were confirmed. Both preventing rehearsal of premise pairs during the ITI and making it more difficult to encode the information in WM adversely affected participants' ability to engage in transitive inference. With a working memory load, participants were less accurate on both the premise and transitive inference pairs during testing. Additionally, loading WM affected participants' awareness of the TI hierarchy even when they performed accurately on the novel TI pair. This suggests that in humans the prefrontal cortex and the hippocampus interact to support deductive reasoning tasks such as transitive inference.

Figure 1: Familiar and unfamiliar stimuli.

References

Barrouillet, P. (1996). Transitive inferences from set-inclusion relations and working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*, 1408-1422

Now You See It, Now You Don't: Can People Mentally Impose Spatial Category Boundaries?

Vanessa R. Simmering (vanessa-simmering@uiowa.edu)
John P. Spencer (john-spencer@uiowa.edu)
Department of Psychology, University of Iowa
E11 Seashore Hall, Iowa City, IA 52242 USA

Introduction

Several accounts of spatial memory biases propose that people "mentally impose" spatial category boundaries (e.g., Huttenlocher, Hedges, & Duncan, 1991). However, in most tasks that have reported categorical biases, adults have used boundaries aligned with either visible lines or axes of symmetry. This raises a fundamental question: can people mentally impose a category boundary in the absence of perceptual structure supporting such a division? In our previous research, we have demonstrated that category boundaries can be created and destroyed in a spatial memory task by changing the perceptual cues available in the task space (Simmering & Spencer, 2004). Thus, in the present study, we added or deleted perceptual structure to see if people could maintain a categorical division in the absence of relevant perceptual information.

Method & Results

One behavioral signature of using a category boundary in spatial recall is drift away from the boundary over delay (e.g., Spencer & Hund, 2002). In the current experiments, we used direction of drift as an indication of whether participants were using the category boundary. In Experiment 1, the presence of perceptual support for the boundary alternated across blocks. Participants' responses showed drift away from the category boundary only when the perceptual support was provided. Figure 1 shows the switch in drift direction based on the presence of perceptual structure (negative values indicate drift away from the boundary). In both conditions (solid and dashed lines), performance depended on the available perceptual structure. This suggests that people need perceptual support to impose a category boundary.

In Experiment 2, the presence of perceptual support varied randomly across trials. Although imposing the category boundary should have been simpler in the experiment, participants were still unable to impose the reference without perceptual support (see dotted line in Figure 1). This provides further evidence that people need perceptual support to impose a category boundary.

Further analysis of the effects in Experiment 2 suggested that the order of trials may influence participants' ability to mentally impose the spatial category boundary. That is, whether support for the boundary was present on the just-previous trial seemed to influence performance, but the number of trials available for this analysis was too small to determine the reliability of this effect. Experiment 3 was designed to test this more directly by providing the most supportive conditions for mentally imposing the category boundary. In this experiment, we designed mini-blocks of trials in which a trial with no perceptual support followed sets of 1 or 3 trials with perceptual support. In the mini-blocks with 3 trials, the memory for the category should be strongest, and therefore easiest to impose on the following trial. Data collection for this experiment is in progress.

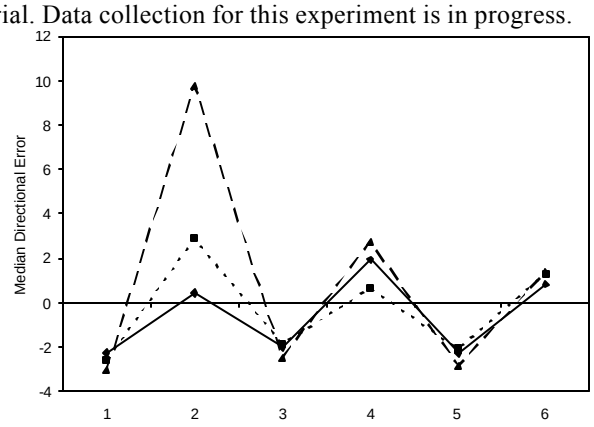


Figure 1: Directional error across blocks for Experiments 1 (solid and dashed lines) and 2 (dotted line).

Conclusion

This series of experiments suggests that adults are unable to mentally impose a category boundary without perceptual support. Even when a category boundary has been used on previous trials, when perceptual support is removed, adults' performance indicates a failure to impose the boundary.

Acknowledgments

NIMH RO1 MH62480 awarded to John P. Spencer
NSF BCS 00-91757 awarded to John P. Spencer

References

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352-376.

Simmering, V. R. & Spencer, J. P. (2004). Creating and destroying frames of reference for spatial working memory. *Manuscript in preparation*.

Spencer, J.P. & Hund, A.M. (2002). Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General, 131*, 16-37.

Perception of temporal continuity in discontinuous moving images.

Tim J. Smith (tim.smith@ed.ac.uk)
Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh,
2 Buccleuch Pl., Edinburgh, EH8 9LW, UK

Introduction

The perception of short visual durations is dependent on the internal and external allocation of attention. When attention is concentrated on a single visual event the perceived duration of that event will be longer than if attention is divided or distracted ('Watched Pot Illusion': Block, George & Reed, 1980). Similarly, when we perform voluntary saccadic eye movements to an object with a discernible temporal signature we perceive the first duration following the saccade as being longer than the subsequent duration ('Stopped Clock illusion': Yarrow et al, 2001).

These effects show how ecological time perception can lead to perceptual discontinuities. The opposite effect can be seen in motion picture perception: the perception of temporal continuity from discontinuous visual events. An action filmed from two different camera positions appears temporally continuous if two frames (83.3ms) of the action are overlapped during the cut between shots (Anderson, 1996). This technique of 'continuity editing' is well established yet the perceptual foundations for it have rarely been empirically investigated.

The aim of this study was to show how 'continuity editing' can be explained as the natural result of time perception under different viewing conditions (fixation, peripheral change, and saccadic eye movements).

Methods

Twenty subjects (10 male, 10 female; 19-33 years) were shown animations depicting a series of letters within photo-realistic scenes and asked to judge whether the presentation duration of a target letter was longer or shorter than all other letters. A Modified Binary Search procedure (MOBS: Yarrow et al, 2001) was used to identify a presentation duration perceived by the subjects as being equal to 1000ms under nine viewing conditions: Saccade target relocation (3) x Background (3) (see Figure 1). The presentation order of the conditions was either blocked or randomized to create predictable and unpredictable viewing conditions.

Results and Discussion

The following effects were identified in this study:

- Predictable viewing conditions lead to perceptual extension of fixation duration (estimate = 898ms; t=-3.096 df=9 p=.013). Unpredictable viewing conditions lead to accurate fixation duration perception (estimate=991ms).
- Unexpected peripheral change leads to perceptual shortening of fixation duration (duration = 1049ms; t=-3.180 df=9 p=.011). This effect disappears when coinciding with an unexpected saccade target relocation.

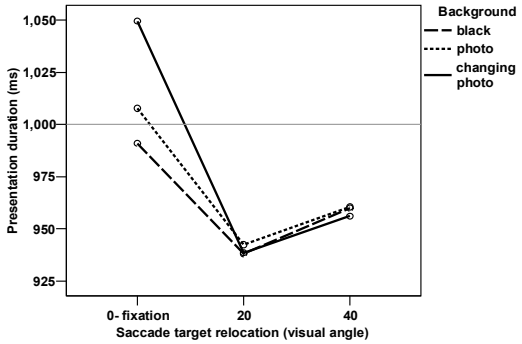


Figure 1: Presentation duration perceived as being equal to 1000ms under the nine randomized viewing conditions.

- Small saccade target relocations (20°) lead to the perceptual extension of post-saccadic durations (blocked: duration decrease=67ms, p=.041; random: duration decrease=53ms, p=.054, one-tail). Large target relocations (40°) only lead to a similar extension when they are unexpected (duration decrease = 57.2ms, p=.023).

These results show how involuntary capture of attention by peripheral change is perceptually under-compensated and voluntary redirection of attention (saccades) over-compensated when perceiving visual durations. These effects are moderated by expectancy.

This allows us to conclude in favour of and explain in more detail the ecological basis of 'continuity editing': perceived temporal continuity is created by overlapping one frame (42.5ms) of a visual event across a cut when the focus of attention remains in the same location but the periphery changes, and omitting one frame when the focus relocates.

Acknowledgments

This project was conducted under financial support from EPSRC and ICCS, and the supervision of John Lee, Helen Pain, and Graeme Ritchie.

References

Anderson, J. (1996) *The Reality of Illusion: An Ecological Approach to Cognitive Film Theory.* Southern. Illinois University Press

Block, R.A., George, E.J., & Reed, M.A. (1980) A Watched Pot sometimes boils: a study of duration experience. *Acta Psychologica 46,* 81-94.

Yarrow, K., Haggard, P., Heal, R., Brown, P., & Rothwell, J.C. (2001) Illusory perceptions of space and time preserve cross-saccadic perceptual continuity. *Nature, 414,* 302-305

Perceiving Narrated Events

Nicole K. Speer (nkspeer@artsci.wustl.edu)
Jeffrey M. Zacks (jzacks@artsci.wustl.edu)
Jeremy R. Reynolds (jrreynol@artsci.wustl.edu)
Department of Psychology, Washington University
Campus Box 1125, One Brookings Drive
St. Louis, MO 63130 U. S. A.

Introduction

The perception of events, such as viewing a baseball game, is typically studied using movies of real-world events (e.g., Zacks, Tversky, & Iyer, 2001). However, people frequently perceive events by reading, hearing, or talking about events. Current models of text comprehension suggest that the process of perceiving events in narrated activity may be driven in part by changes in various dimensions of the narrated situation (e.g., Zwaan, Radvansky, Hilliard, & Curiel, 1998). Four experiments were conducted to determine a) whether people are able to reliably perceive event structure in narratives using a paradigm employed to study event structure in real-world activities, and b) which dimensions of the narrated situation are relevant to the perception of event structure in narratives.

Materials & Method

The stories used in all current studies were excerpts from *One Boy's Day* (Barker & Wright, 1951). Written in the style of a narrative, this book provides a detailed record of the activities of a 7-year old boy (Raymond) during a single day in the 1940's. The four stories used in these studies described Raymond waking up, playing in the schoolyard, working on an English lesson, and attending a Music class.

Task Design

In the first experiment, 32 participants were asked to listen to each narrative twice: Once while identifying large units of activity (coarse segmentation), and once while identifying small units of activity (fine segmentation). In the second experiment 32 participants were asked to read the narratives twice on paper, and place a line between words to identify coarse and fine segments of activity in the same stories. In a third experiment, clause-by-clause reading times were collected from 32 participants. In a final experiment, 32 participants used a 7-point scale to rate the predictability of the activity described in each clause given prior information from the story.

Analysis

For each study, the data were analyzed at the level of clauses. In the first two studies, clauses were considered to be event boundaries if a participant segmented at least once during that clause. Each clause was coded for changes on one of six dimensions: temporal references, changes in the foregrounding of characters, their spatial locations, the objects with which they were interacting, their goals, and the causal relations between their actions. The number of syllables (or the duration of the spoken clauses) and punctuation were also coded for each clause.

For each study, these variables were used to predict the patterns of large and small segmentation, reading time, or predictability ratings for each participant. The coefficients generated from these regressions were used to measure the influence of the independent variables in each study.

Results & Conclusions

Participants identified larger units of activity during coarse than fine segmentation, indicating that they were able to perceive structured activity in the narratives.

When identifying large units of activity, participants' patterns of segmentation were related to changes in the foregrounding of characters, their locations and goals, and the causal relations between their actions. In contrast, when identifying small units of activity, patterns of segmentation were more strongly related to changes in characters' interactions with objects. These results were consistent across presentation modalities, and suggest that fine-grained events are closely tied to physical interactions, whereas coarse-grained events are more tied to goals and plans.

Situational changes were also associated with slower reading times (see also Zwaan, et al., 1998) and lower ratings of predictability, suggesting a role for transient changes in predictability in the perception of event structure.

Acknowledgments

We thank Sherry Beaudreau, Brett Hyde, Ben Oberkfell, and Rolf Zwaan for their assistance in the construction and coding of the materials, as well as Sarah Berson, Rahal Kahanda, Jean Vettel, and Richard Zernickow for help with data collection. This research was supported by the National Science Foundation and the Office of Naval Research.

References

Barker, R. G., & Wright, H. F. (1951*). One boy's day: A specimen record of behavior*. New York: Harper & Brothers.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*, 29-58

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E., & Curiel, J. M. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading*, *2*, 199-220.

Laypersons Searching for medical information on the Web: The Role of Metacognition

Marc Stadtler (stadtlm@uni-muenster.de)
Rainer Bromme (bromme@uni-muenster.de)
Department of Psychology, University of Muenster
Fliednerstr. 21, 48149 Münster, Germany

Introduction

Today, the WWW is a very prominent resource of health-related information, both, for medical experts and laypersons (e.g. Fox, 2003). The latter often retrieve these information to make an informed decision. However, one cannot expect laypersons to deal with these information effectively offhand. We rather assume that to succeed, laypersons need to actively guide their search process on a metacognitive level, since metacognitive strategies are known to play an important role in the comprehension of complex documents (e.g. Hill & Hannafin, 1997). However, it is yet unclear whether laypersons spontaneously guide their web search on a metacognitive level and - in case they do so - whether the use of metacognitive strategies is related to search success.

Method

To answer this question we carried out a study in which 20 university students with little medical knowledge participated. Their task was to search the WWW for information on cholesterol in order to help a fictitious friend make a knowledge based decision: "Is a medical treatment of my high level of cholesterol necessary?". Participants were provided with 11 pre-selected websites containing controversial information on the topic. Search time was limited to 35 minutes. Knowledge acquisition, decision conflict and detailedness of written credibility assessments functioned as measures of search success. Cognitive processes were ascertained using a think-aloud procedure. Verbal protocols were analyzed using a category system which comprises the categories *Planning*, *Monitoring*, *Evaluation* and *Elaboration*. Inter-rater reliability was 82% across all categories.

Results

Results show that participants differ considerably in their metacognitive activity (see Table 1). Participants' metacognitive activity is rather consistent across the four categories (Cronbach's $\alpha = .78$). Interestingly, metacognitive activity is positively related to knowledge acquisition. Correlation coefficients range from $r = .45$, $p < .05$ (Monitoring), to $r = .57$, $p < .01$ (Evaluation). No significant correlation could be obtained for the relationship of Planning and knowledge acquisition ($r = -.18$, *ns.*).

Results concerning subjectively experienced decision conflict reveal a negative but nonsignificant correlation with metacognitive activity ($r = -.23$, *ns.*). The assumption that better knowledge of the topic cholesterol is related to subjectively experienced decision conflict could be confirmed only partially. While factual knowledge did not correlate significantly ($r = -.13$, *ns.*), comprehension of the subject matter was significantly correlated with scores on the Decision Conflict Scale ($r = -.49$, $p < .05$).

Table 1: Mean number of metacognitive statements and standard deviations for each category.

Category	*M*	*SD*
Planning	10.90	6.61
Monitoring	13.95	6.86
Evaluation	19.00	12.60
Elaboration	11.65	9.48

Finally, analysis of participants' written credibility assessments show that the more participants evaluate information during the search process, the better they are able to report on the credibility of information after their search ($r = .46$, $p < .05$).

To summarize, in the present study the importance of metacognitive strategies for a successful web search could be demonstrated. The results point to the need for metacognitive interventions which support laypersons in dealing with complex technical information on the WWW. Therefore, we have developed the computer based tool *met.a.ware.* The tool enables laypersons to systematically store the information they have found on the web. For this, laypersons have to assign the information gathered to different tabs, which are labeled with aspects of the topic cholesterol. Furthermore, laypersons are prompted to engage in metacognitive activities each time they add information to the system. In ongoing experiments, different types of metacognitive prompts (i.e. evaluating information and monitoring ongoing comprehension) are tested against each other. Thereby, we seek to seperately examine the contributions of different metacognitive activities to a successful web search. First results from our current experiments point to the supportive character of *met.a.ware*.

References

Fox, S., & Fallows, D. (2003). *Internet health resources*. (Vol. 2003): Pew Internet and American Life Project. Retrieved on 2004-05-07 from: http://www.pewinternet.org

Hill, J. R., & Hannafin, M. J. (1997). Cognitive strategies and learning from the World Wide Web. *Educational Technology Research and Development, 45*(4), 37-64.

Optimal Auditory Categorization on a Single Dimension

Sarah C. Sullivan (sullivans@boystown.org)
Center for Hearing Research, Boys Town National Research Hospital
425 North 30th Street, Omaha, NE, 68131, USA

Andrew J. Lotto (lottoa@boystown.org)
Center for Hearing Research, Boys Town National Research Hospital
425 North 30th Street, Omaha, NE, 68131, USA

Randy L. Diehl (diehl@psy.utexas.edu)
Department of Psychology, University of Texas
1 University Station A8000, Austin, TX, 78712, USA

Introduction

A number of auditory tasks, including speech perception, require listeners to categorize stimuli on the basis of one or more features of the input. In many cases, especially speech, there is no one-to-one mapping between values along continuous features and discrete categories (e.g., phonemes). How then do perceptual systems categorize stimuli under uncertainty? One possible solution is that perceptual systems identify and use statistical information inherent in the acoustic environment. We propose that perceivers incorporate distributional knowledge about the acoustic environment with the information provided by the signal in order to make optimal (i.e., maximized accuracy) categorical decisions. Statistical approaches such as this are widely used in vision research but are rarely applied to auditory or speech perception. Our goal in this study was to develop a framework that will provide testable hypotheses about the nature of statistical (distributional) learning in auditory perception in general and specifically in speech perception.

Methods

In this experiment, participants were presented non-speech sounds sampled from two overlapping distributions. The sounds consisted of 25 narrow-band noise bursts varying in center frequency from 1000 to 1360 Hz. Three different conditions were created by varying parameters of the training distributions. The distributions varied in the ratio of stimuli in each category (i.e., prior probabilities of each category) as well as the amount of overlap between the two distributions. Figure 1 displays the distributions for one of these conditions. The listeners were asked to identify the sounds as belonging to one of two categories ("A" or "B") and feedback was provided after each trial. Due to the overlap between the distributions, there was no deterministic relationship between center frequency and category label for many stimuli.

Results

In decision tasks such as these, optimal performance requires listeners to create a criterion boundary on the dimension (i.e., a particular frequency). Stimuli on either side of this boundary should receive different category labels. Within as few as one block of training trials, most listeners displayed a stable category boundary. Boundaries were estimated from categorization functions averaged across several training blocks. These boundaries varied as a function of the distribution characteristics and were statistically equivalent to the point of distribution intersection. Slopes from categorization functions were steeper than the slopes of the training distributions; suggesting that listeners were more likely using a criterion bound as opposed to simply probability matching the distributions. In general, listeners were responding in a near optimal manner with minimal experience with the training distributions. The ability to use distributional information to map from continuous dimensions to category labels is also essential for speech sound categorization.

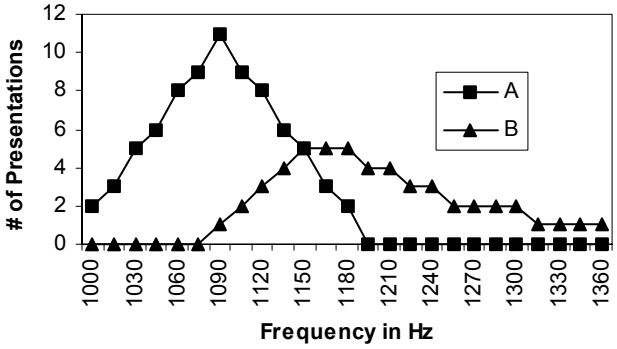


Figure 1: Stimulus Distributions

Acknowledgements

This research was funded by the National Institute on Deafness and Other Communication Disorders and the National Science Foundation.

Implicitly Learned Sequences Structure the Perception of Human Activity

Khena M. Swallow (kmswallo@artsci.wustl.edu)
Jeffrey M. Zacks (jzacks@artsci.wustl.edu)
Department of Psychology, Washington University in St. Louis
Campus Box 1125, One Brookings Drive, St. Louis, MO 63112 USA

Introduction

People are able to learn and use temporal sequences to guide their perception and behavior. This ability has been demonstrated in visual search tasks (Olson & Chun, 2001) and is present in infants as young as eight months (e.g., Saffran, Aslin, & Newport, 1996). Does seeing the same sequence of activity repeated in different contexts cause that sequence to be treated as a coherent perceptual unit? If temporal sequence learning influences the perception of others' activity, observers should be able to use sequences of activity to prepare for an event predicted by those sequences. Furthermore, these sequences should have a direct effect on the structure people perceive in others' activity: implicitly learned sequences of activity should be perceived as units of activity. The goal of this study was to determine the consequences of temporal sequence learning for the way observers understand the actions of others.

Experiment 1

The first experiment addressed the hypothesis that participants can learn sequences of human activity and use those sequences to aid in a target detection task. A series of pictures (presented for 750 ms with no ISI) of a man with his arm in six different positions and forming thirteen different hand gestures were presented to participants. Eight participants were asked to monitor the gestures and press the correct button whenever they saw either of two target gestures. Within the series of pictures a sequence of seven arm positions was repeated 320 times. Each repetition of the sequence was separated by two to twelve pseudo-randomly selected arm positions. For the first three-quarters of the task, a target gesture immediately followed the sequence. In the last quarter of the task the sequence did not predict when the target would appear. Response times steadily decreased while the target followed the sequence but then increased once the target was no longer predicted by the sequence. None of the participants discovered the sequence nor were they able to demonstrate knowledge of the sequence in a cued-generation recognition test.

Experiment 2

The first experiment demonstrated that observers learn sequences of human activity and can use this knowledge to prepare for important, task-related activity. In a second experiment we sought to determine whether these learned sequences of human activity are treated as perceptual units and, if so, whether this perception depends upon the predictive value of the sequence. Twenty-four participants performed the same target detection task described in the first experiment. However, the predictiveness of the sequence was manipulated across two groups (predictive and nonpredictive groups) and the task was shortened. The performance of these two groups was significantly different: The group for whom the sequence was predictive showed better task performance than the group for whom the sequence was not predictive. Participants also performed a segmentation task in which they were asked to identify boundaries between units by pressing a button when one natural unit of activity ended and another began (Zacks & Tversky, 2001). This task used the stimuli and sequence from the target detection task (though no target gestures were presented). As they performed the segmentation task both groups of participants chose boundaries relative to the sequence of activity they learned in the first task.

Conclusions

As observers watch others perform everyday activities they build up structured representations of their behavior (Zacks & Tversky, 2001). These structured representations influence how they break behavior into smaller units of activity. The results of these experiments suggest that the structure people impose on their experience is in part due to implicitly learned sequences of human activity. These sequences are learned and used to predict the occurrence of important events. Moreover, these sequences influence the way we subsequently perceive the actions of others.

Acknowledgments

Portions of this research were supported by NSF grant 0236651 to Jeffrey M. Zacks. The authors would like to thank Henry L. Roediger III for his help with the recognition task and Nicole K. Speer for her helpful comments.

References

Olson, I. R., & Chun, M. M. (2001). Temporal contextual cuing of visual attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1299-1313.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*(1), 3-21.

Within-Language Attention Control and Second Language Proficiency

Marlene Taube-Schiff (marlene_taubeschiff@yahoo.ca)
Norman Segalowitz (norman.segalowitz@concordia.ca)
Department of Psychology, and the Centre for the Study of Learning and Performance, Concordia University
7141 Sherbrooke Street West, Montréal, QC H4B 1R6 Canada

Introduction

This study investigated the role linguistic attention control might play in second language (L2) proficiency. Cognitive linguists have proposed that language, beyond referring to events, objects and their properties, also *directs attention* towards relationships between elements in a message (e.g., Slobin, 1996). This is especially true of function words, grammatical morphemes, etc. For example, in *The book is under the table*, the meaning of the preposition *under* is not represented by sensori-motor/perceptual experiences in the same way as it is for *book* and *table*. Instead, *under* directs attention to the relationship between *book* and *table*. Such grammatical elements pose challenges for L2 learners due to these attention-directing functions (Slobin, 1996).

This research used the *alternating runs paradigm* (Rogers & Monsell, 1995) to study attentional control when processing grammatical elements. This paradigm requires responses to two tasks that repeat and alternate predictably (e.g.,...AABBAA...), creating a sequence of repeat and shift trials. Typically, reaction times (RTs) are slower on shift than on repeat trials, resulting in *shift costs* that reflect the burden that shifting places on the attention system.

Decontextualized simple stimuli are often used to investigate attentional task shifting processes. Recently, Taube-Schiff and Segalowitz (2003) found significant shift costs during performance of first language (L1) grammatical judgment tasks involving contextualized sentence-like stimuli. The current two experiments aimed to clarify the specificity of linguistic attention in the grammatical domain by asking the following questions: (1) Does degree of grammatical similarity between tasks affect shift costs? (2) Does attention control in L2 differ for shifts between grammatical elements versus non-grammatical elements? (3) Are linguistic attention shift costs similar in L1 and L2?

Method

Bilingual undergraduate participants (Expt. 1: N=24; *M*=24 years and Expt. 2: N=32; *M*=22 years; L1=English; L2=French) performed an *alternating-runs task* involving 2-alternative forced choice conditions, with trials predictably alternating between repeat and shift trials. Stimuli were displayed on a computer screen and consisted of target words embedded in sentence-like fragments, appropriately counterbalanced for their occurrence in specific sentence contexts. In Experiment 1, participants were tested in two conditions in L1, each involving the following two tasks: In a Grammatically-Different (GDIFF) condition, verb targets were judged for temporal meaning (past versus present tense) and prepositions for location meaning (above versus below). In a Grammatically-Similar (GSIM) condition, prepositions were judged for either one type of location meaning (above versus below) or another type (near versus far). In Experiment 2, participants were tested in L1 and L2 in the GSIM condition (same as in Expt. 1) and in a Non-Grammatical (NOUN) condition in which noun targets were judged with respect to non-grammatical category membership (air versus water craft, and 2- versus 4-wheel vehicles).

Results

Repeated measures ANOVAs were conducted comparing shift and repeat trials to obtain shift costs. In Expt. 1, shift costs were significantly greater for the location task in the GDIFF versus the GSIM condition. In Expt. 2, a significant interaction effect revealed shift costs were significantly greater in L2 in the GSIM condition than in the NOUN condition. Finally, shift costs were significantly greater in L2 than L1, in the GSIM and not in the NOUN condition.

Discussion

The main findings from these studies were: (1) Increased grammatical similarity between tasks decreased shift costs, suggesting a lower attentional burden. (2) There was a greater impact on attention control in L2 when shifting attention between grammatical versus non-grammatical elements, and (3) Linguistic attention shifts costs were greater in L2 than L1, but only significantly so in the grammatical judgment tasks. Results speak to psychological distinctions within the grammatical system and provide additional support for the idea that grammatical elements are more difficult to master in L2 (Slobin, 1996).

Acknowledgments

This research was funded by a grant to NS from the Natural Sciences and Engineering Research Council of Canada.

References

Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207-231.

Slobin, D. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96). Cambridge, U.K.: Cambridge University Press.

Taube-Schiff, M., & Segalowitz, N. (2003). Reconfiguration and inertial processes in attention switching during reading. Cognitive Science Society Conference. Boston.

Intra-clause Constraints in Think-Aloud Protocols

Stacey A. Todaro (shaberk1@niu.edu)
Joseph P. Magliano (jmagliano@niu.edu)
Keith K. Millis (kmillis@niu.edu)
Department of Psychology, Northern Illinois University
DeKalb, IL 60115

Danielle S. McNamara (d.mcnamara@mail.psyc.memphis.edu)
Department of Psychology, University of Memphis
Memphis, TN 38152-3230

Christopher C. Kurby (ckurby@niu.edu)
Department of Psychology, Northern Illinois University
DeKalb, IL 60115

Clauses in verbal protocols produced during reading reflect relationships among entities (e.g., arguments) and events (e.g., verb predicates) in the reader's unfolding situation model (e.g., Trabasso & Magliano, 1996). The content of the clauses come from three knowledge sources: the current sentence, the prior text, and the reader's world knowledge. In this study, the relative impact of text elements, as pertaining to dimensions of situation model construction during reading was examined (Zwaan & Radvansky, 1998). Specifically, we measured the extent to which producing an argument (or predicate) influences the likelihood of producing a predicate (or argument) within each knowledge source.

Method

The study included 64 participants enrolled in a critical thinking class at Northern Illinois University. Participants read and self-explained two of four science texts. The four texts were adopted from high-school textbooks on life sciences. Self-explanations were collected after each sentence was presented.

Protocol Analysis

Reader's utterances were parsed into clauses containing main verbs. The verb predicates and arguments within each clause were identified as belonging to one of three sources: the current sentence, the prior text, or world knowledge.

Results and Discussion

We computed the extent to which one constituent type (i.e., verb predicate vs. argument) determines the use of the other within a knowledge source with the following two equations:

Equation 1: Argument Determines Predicate (ADP) =
p (generate P | generate A) - p (generate P | not generate A)

Equation 2: Predicate Determines Argument (PDA) =
p (generate A | generate P) – p (generate A | not generate P)

Table 1 presents the mean constraint scores as a function of the source of a verb (i.e., current sentence, prior text, or world knowledge). A 2 (Constituent Constraint: argument or predicate) X 3 (Source: current sentence, prior text, or world knowledge) repeated measures ANOVA was conducted on the constraint scores. This analysis yielded a main effect for constraint score, such that ADP scores (M = .46) were significantly different from PDA scores (M = .51), $F(1, 126) = 32.03$, $MSE = .008$, $p < .01$. This main effect was qualified by a significant Constituent Constraint X Source interaction ($F(2, 126) = 142.94$, $MSE = .005$, $p < .01$). Post hoc analyses revealed that constraint scores differed across the knowledge sources. With respect to current sentence and prior text, verb predicates constrained the arguments more than arguments constrained the verb predicates. With respect to world knowledge, the opposite pattern was found.

These data suggest that intra-clause constraints may be source dependent. Specifically, when readers describe information from the current sentence, or are accessing information from the prior discourse, they tend to describe the events and entities associated with those events. On the contrary, when entities from world knowledge are activated, readers must construct the events which than link them to the current discourse information.

Table 1: Constraint scores as a function of source and constituent.

Source	ADP Score	PDA Score
Current sentence	0.47	0.52
Prior text	0.37	0.57
World knowledge	0.53	0.43

References

Trabasso, T. & Magliano, J.P. (1996). Conscious understanding during comprehension. *Discourse Processes, 21,* 255-287.

Zwaan, R.A., & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123(2),* 162-185.

Eye Scanpaths Influence Memory for Spoken Verbs

Alexia C. Toskos (atoskos@indiana.edu)
Rima Hanania (rhanania@indiana.edu)
Stephen Hockema (shockema@indiana.edu)
Program in Cognitive Science, Indiana University
1033 E. Third St., Bloomington, IN 47405 USA

Introduction

The human body is designed to interact with its environment. Information from all sensory modalities is integrated to allow for navigation and decision making. Developing theories of embodied cognition suggest that thought may be grounded in low-level motor activity and sensory modalities.

The present study tests whether eye motion affects memory for spoken verbs. Vision is a dynamic process in which saccades are used to gather information from multiple points in space. However, these eye fixations are not random. Eye scanpaths are often consistent with thought during the absence of any visual stimulus (Spivey & Geng, 2001, & Laeng and Teodorescu, 2002). The present question is whether feedback from low-level motor activity also plays a role in high-level cognitive processes. If natural eye movements are congruent with thought, perhaps it is possible to influence thought by controlling eye motion.

Richardson, Spivey, Barsalou, & McRae (2003) found that certain verbs carry either a vertical or horizontal spatial orientation, and that spatial orientation is activated upon stimulus presentation. Thus, it is hypothesized that horizontal or vertical eye scanpaths will either enhance memory for words whose spatial orientation is congruent with that of the motion and/or inhibit memory for those words that are incongruent with the motion.

Method

Participants were 45 sighted undergraduates at Indiana University. Apparent motion software was used to create a black circle flipping back and forth vertically or horizontally. As participants tracked the stimulus, a pre-recorded list of verbs played. The list contained 20 of the verbs that Richardson et al. (2003) found to carry spatial meaning. Ten verbs were played while the participant observed apparent motion in one direction, and the next ten words were played while the participant observed apparent motion in the other direction. After a two minute pause, participants were given a list of 40 verbs and asked to circle those which appeared in the previously heard list.

Congruent instances consisted of "vertical" verbs that were presented while participants' eyes moved vertically or "horizontal" verbs while participants' eyes moved horizontally. Incongruent instances consisted of "vertical" verbs presented during horizontal motion or "horizontal" verbs during vertical motion. According to the hypothesis, performance on the recognition task should be better for congruent verbs than for incongruent verbs.

Results and Discussion

The results suggest that eye movements do prime memory for verbs, with vertical eye movements enhancing recognition of verbs with vertical spatial orientations over those with horizontal orientations. The effect was weaker in the horizontal condition. A 2(eye motion direction) x 2(order) x 2(congruency) analysis of variance for a within subjects design yielded a reliable interaction between eye motion direction and order, $F(1,43)=12.470$, $p<.001$. Horizontal eye movements only primed memory for horizontal verbs when this direction of motion was performed first. The analysis also yielded a reliable congruence x direction of motion interaction. The vertical condition yielded a greater difference between congruent and incongruent instances than did the horizontal condition, and the effect was reliable under both orders of presentation.

An ongoing study replicates these findings using a between subjects design (to eliminate order effects) and longer scanpaths. What one thinks is known to determine how one moves. The present findings show that how one moves (at least how the eyes move) determines what is remembered. Apparently, verb meanings are represented in a form close to the sensorimotor surface.

Acknowledgments

Special thanks to Linda Smith for her insights and contributions to this study.

References

Laeng, Bruno & Teodorescu, Dinu-Stefan. (2002) Eye Scanpaths During Visual Imagery Reenact Those of Perception of the Same Visual Stimulus. *Cognitive Science,* 26, 207-231.

Richardson, Daniel C; Spivey, Michael J; Barsalou, Lawrence W; McRae, Ken. (2003) Spatial representations activated during real-time comprehension of verbs. *Cognitive Science,* Vol 27(5), 767-780.

Spivey, M. J. & Geng, J. J. (2001) Oculomotor mechanisms triggered by imagery and memory: Spontaneous eye movements to objects that aren't there. *Psychological Research*, 65, 235-241.

Use of Spatial Transformations in Graph Comprehension

Susan Bell Trickett (stricket@gmu.edu)
Department of Psychology, George Mason University
4400 University Dr., Fairfax, VA 22030 USA

J. Gregory Trafton (trafton@itd.nrl.navy.mil)
Naval Research Laboratory, NRL Code 5513
Washington, DC 20375 USA

Introduction

Current theories of graph comprehension are largely silent about the processes by which inferences are made from graphs (Freedman & Shah, 2002; Pinker, 1990), although it is apparent that people are able to make such inferences. In Trickett & Trafton (2004), we proposed that people use spatial reasoning, in the form of spatial transformations (Trafton et al., in press) to answer inferential questions. This paper is an extension of our earlier study, in which we standardized the graphs presented, so that the distance from the x and y axes was identical for all conditions, we removed typing time from the RT measure. Finally, we expanded the experiment with an additional "middle extension" condition.

Method

8 graduate students and faculty at GMU participated. Participants were shown 40 unlabelled line graphs presented in random order, 10 in each of 4 conditions. They were asked for the value of the y axis at a point on the x axis. The 4 conditions were: read-off (arrow beneath line), near (arrow slightly beyond line), middle (arrow a greater distance) and far (far beyond end of line) (see Figure 1).

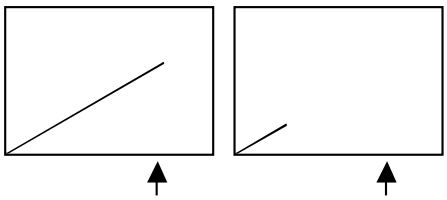


Fig. 1: Schematic of readoff (left) and far (right) conditions.

The readoff condition required no spatial transformations. However, in the near, middle and far conditions, we hypothesized that participants would mentally extend the line (i.e., use spatial transformation) to locate its intersection with the perpendicular from the red arrow. Spatial transformation theory predicts that longer extensions take longer; thus, we predicted that participants would be fastest in the read-off (no extension) condition, increasingly slower in the near and middle conditions, and slowest in the far (longest extension) condition. We also predicted that accuracy would decrease with increased use of spatial transformations, as people must move further from "anchor points" on the graph to obtain needed information—i.e, most accurate in the read-off condition, decreasingly accurate in the near and middle conditions, and least accurate in the far condition.

Results and Discussion

We measured accuracy as the absolute value of the correct response minus the participant's response. Response times (RT) represent the time taken to reach an answer.

Consistent with our hypothesis, participants were most accurate on the read-off task, decreasingly accurate on the near and middle tasks, and least accurate on the far task, repeated measures ANOVA $F(3, 15) = 12.43$, $p < .01$, linear trend $F(1, 5) = 13.93$, $p < .05$.

RT data also supported our hypothesis. Participants were fastest on the read-off task, increasingly slower in the near and middle tasks, and slowest on the far task, $F(3, 15) = 7.44$, $p < .05$, linear trend $F(1, 5) = 10.99$, $p < .05$. The linear trend is consistent with the idea that a longer extension takes more time to execute than a shorter one. If this is true, it should take a measurable amount of time more for each extension. In order to calculate how long each extra extension took, we did a linear regression, using the distance the participants had to extend the line (recall that the distance along the x and y axes was constant). This analysis was significant, $r = .43$, $p < .01$. The analysis yielded the following formula: Response Time = 4.8 + .63, where 4.8 seconds is the baseline time to read information from the graph and .63 is the amount of extra time required to extend the line each centimeter distance required. This result supports our hypothesis that participants used spatial transformations, by indicating a systematic relationship between response time and the distance mentally traveled. As participants had to draw longer mental extensions to the graph, their response times systematically increased. Thus, we propose that a comprehensive theory of graph comprehension should accommodate spatial reasoning.

References

Freedman, E. G., & Shah, P. (2002). *Toward a Model of Knowledge-Based Graph Comprehension.* Paper presented at the Diagrammatic Representation and Inference, Second International Conference, Diagrams 2002, Callaway Gardens, GA, USA.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73-126). Hillsdale, NJ: Lawrence Erlbaum.

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (in press). Overlaying images: Spatial transformations of complex visualizations. *Foundations of Science.*

Trickett, S. B., & Trafton, J. G. (2004). *Spatial transformations in graph comprehension.* Paper presented at the Diagrams 2004, Cambridge, UK.

Cross-Category Effects in Spatial Working Memory

Wendy W. Troob (wendy-troob@uiowa.edu)
Vanessa Simmering (vanessa-simmering@uiowa.edu)
John Spencer (john-spencer@uiowa.edu)
Department of Psychology, University of Iowa
E11 Seashore Hall, Iowa City, IA 52242 USA

Introduction

Most models of memory and spatial categorization predict that people select relevant categorical information at the time of stimulus encoding (e.g., Huttenlocher, Hedges, & Duncan, 1991). Following encoding, unselected category information has no influence on subsequent memory and categorization responses. In contrast, the Dynamic Field Theory (DFT), a neural network model of spatial working memory, suggests that unselected information can still exert an influence following encoding (Schutte, Spencer, & Schöner, 2004). In particular, the network's activation continues to be affected by "unselected" categorical information during memory delays.

To investigate this issue, memory targets were placed in separate spatial categories, but close to a category boundary (e.g. to the left and right of the midline axis of the task space). Participant's experience with the targets was varied by changing the relative frequency of trials to each target. The critical question concerned whether or not the longer-term memory of items in the unselected (e.g., right) category would affect memory for items in the adjacent (e.g., left) category during memory delays. If the predictions of the DFT are correct, such cross-category interactions would be expected.

Method

Participants were seated at a table with a homogeneous surface in a dimly lit room. Two dots aligned with the table's vertical axis were presented 15cm to right of midline. Previous work has demonstrated that these dots form a salient reference axis in spatial recall tasks (Simmering & Spencer, 2004). A target appeared for 2s and participants were asked to recall the location after delays of 0, 10, or 20s. We examined performance in four conditions: no bias (targets -5° to the left of the axis and 5° to the right of the axis), bias right (targets at -5° and 5° with twice as many trials to 5°), plus 10 (targets at -5°, 5°, and 10°), plus 80 (targets at -5°, 5°, and 80°). Importantly, participant's experience responding to the left (-5°) target was the same in all conditions.

Results

According to the DFT, performance to the left target should differ across conditions based on the frequency and spatial distribution of targets in the unselected, right category. This is precisely what we found. Repeated and different exposures to targets in the right category exerted significant effects on responses to items in the left category, even though the number and type of trials to items in the left category was identical across all conditions.

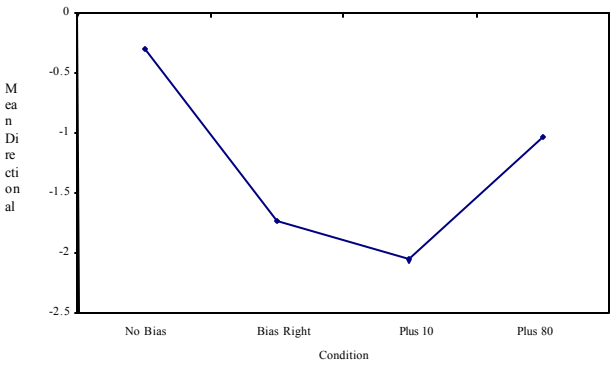


Figure 1: Directional error at the left target across conditions. As predicted, Bias Right and Plus 10 differs significantly from No Bias.

Discussion

Our results are consistent with the proposal that information from both selected and unselected categories can exert an influence on spatial memory performance. Current studies are examining these cross category effects more closely. For instance, the DFT predicts that memory biases to the left target should vary systematically with the distance between the left and right targets.

Acknowledgments

NIMH RO1 MH62480 awarded to John P. Spencer
NSF BCS 00-91757 awarded to John P. Spencer

References

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352-376.

Schutte, A. R., Spencer, J. P., & Schöner, G. (2004). Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. *Child Development, 74*(5), 1393-1417.

Simmering, V. S., & Spencer, J. P. (2004). Reference-related inhibition produces enhanced position discrimination near axes of symmetry. *Manuscript submitted for publication*.

Computational Accuracy and Conceptual Understanding of Statistics: Effects of Thinking Before Plugging and Chugging

David L. Trumpower (trumpower@marshall.edu)
Department of Psychology, Marshall University
One John Marshall Drive, Huntington, WV 25755 USA

Introduction

Mathematics in general and statistics in particular are notoriously difficult topics for many students. Part of the difficulty may be that students view unfamiliar equations and procedures as being extremely complex and, as a result, focus on learning the computational procedures involved in solving problems at the expense of developing a conceptual understanding of the principles which underlie those procedures. One remedy for this situation is to present conceptual, rather than computational, equations to students. Atkinson, Catrambone, and Merrill (2003) found that learners who were trained to perform t-tests using conceptual equations were better able to transfer their knowledge to solve Analysis of Variance (ANOVA) problems than were learners trained to solve t-tests using computational equations.

Even students who posses both computational skill and conceptual knowledge, however, often fail to make use of their conceptual knowledge (Trumpower, 2003). Anecdotal evidence is provided by students who are able to perform a particular statistical computation accurately, but who then fail to draw accurate conclusions based on the computation. It is possible that this situation is caused by students' tendencies to solve statistics problems by immediately looking for appropriate equations and then plugging numbers into those equations (i.e., "plugging and chugging") before thinking about what the equations and numbers actually represent. If so, then forcing learners to answer some simple conceptual questions before performing computations may create links between conceptual and procedural knowledge, and thus allow them to draw more accurate conclusions. This hypothesis was tested in the present study.

Method

Twenty-eight undergraduate psychology students at Marshall University who had not previously taken any statistics courses served as participants. All participants were first asked to study a booklet that described a procedure for performing an independent-groups t-test and provided a solved example. Participants were then asked to perform 2 independent-groups t-tests and to state their conclusions based on the results of the tests. Half of the participants were randomly assigned to a conceptual condition in which they were asked to answer a series of conceptual questions before performing the t-tests. The other half were assigned to a procedural condition in which they performed both t-tests before answering the conceptual questions.

Results

Participants' computations and conclusions for the two t-tests were scored for accuracy. Also the percentage of conceptual questions answered correctly was determined for each participant. A one-way ANOVA revealed no significant difference between the percentage of conceptual questions answered correctly by participants in the conceptual and procedural conditions, $F<1$. A 2 Condition (conceptual, procedural) x 2 Assessment Type (computations, conclusions) mixed factorial ANOVA with repeated measures on the second factor revealed a significant main effect of Assessment Type, $F(1, 26) = 58.40$, $p < .001$, that was qualified by a significant Condition by Problem Type interaction, $F(1, 26) = 8.03$, $p < .01$. Participants in the procedural condition made more accurate computations than participants in the conceptual condition, whereas participants in the conceptual condition drew more accurate conclusions than participants in the procedural condition.

Discussion

Although participants in both conditions were equally capable of answering simple conceptual questions, those who answered them before performing computations were better able to draw accurate conclusions from the computations than those who performed the computations first. However, answering the conceptual questions first appears to have interfered with performing the computations. This latter finding may have been due to increased cognitive load of participants in the conceptual condition, as they may have been thinking about implications of the conceptual questions while performing computations.

References

Atkinson, R. K., Catrambone, R., & Merrill, M. M. (2003). Aiding transfer in statistics: Examining the use of conceptually oriented equations and elaborations during subgoal learning. *Journal of Educational Psychology, 95,* 762-773.

Trumpower, D. L. (2003). *Development of problem solving performance and structural knowledge in physics problem solvers*. Doctoral dissertation. Department of Psychology, University of New Mexico, Albuquerque.

Learning To Be Fluently Disfluent

Mija M. Van Der Wege (mvanderw@carleton.edu)
Department of Psychology, Carleton College
One North College Street, Northfield, MN 55057 USA

E. Christena Ragatz (christa.ragatz@highstream.net)
School of Education, Hamline University
1804 West 138th Street, Burnsville, MN 55337 USA

Introduction

Disfluencies are a normal part of everyday conversation, yet until the past few years, they were viewed only as speech errors and little attention had been paid to their potential functions. Recent research has determined that some of these disfluencies help to coordinate conversational processes. For example, speakers use both *um* and *uh* to indicate a delay in speech production, with *um* indicating a longer delay than *uh* (Smith & H. H. Clark, 1983; H. H. Clark & Fox Tree, 2002). These signals not only are systematically conveyed by the speaker but also appear to be used by the listener when recognizing words (Fox Tree, 2001).

Children are observed to use *um* and *uh* as young as two-years-old, about the same time that they are starting to learn many other content and function words (Van Der Wege, 1996). However, the proper adult-like use of these disfluencies requires cognitive abilities that most other words do not. A child must conceive of a listener as a separate being with a separate understanding from his own.

Preschool-aged children are often described as egocentric, unable to take another's perspective when reasoning or using language (Flavell, 2001; Glucksberg, Krauss, & Weisberg, 1966). Nevertheless, two-year-old children frequently make spontaneous repairs to their speech for the benefit of their listener, indicating that they have an awareness both of the communicative purpose of language and of what their listeners may or may not understand (E. V. Clark & Anderson, 1979).

The current study addresses whether this level of metalinguistic awareness is sufficient for the child to use disfluencies in systematic ways or if the ability to take another's perspective is critical.

Method

Sixty children between the ages of 3 years, 11 months and 6 years from two different preschools were interviewed. All children completed two tasks with a familiar adult. First, they recounted the details of a story, heard for the first time the morning of the interview. Second, they described and discussed a favorite toy.

All conversations were transcribed and digitized. Using a sound editing computer program, the lengths of the pauses surrounding all *um*s and *uh*s were measured and recorded.

Results and Discussion

When talking about their toys, older preschool children (age 5-6 years) used *um* and *uh* systematically, in the same way that adults do (i.e., *um* preceded longer pauses than *uh*). Younger children (age 3-4 years) did not appear to distinguish between their use of the two disfluencies. The distinction was also not seen when the children were recounting the details of the story that they had heard. This was a significant three-way ANOVA interaction ($F[4,2210] = 4.01$, $p < 0.01$).

Apparently, children begin to use *um* and *uh* early in their linguistic lives, recognizing the need to mark speech difficulties. However, they do not have the ability to use these disfluencies in an adult-like manner until later, when they have sufficient cognitive resources to take another's perspective.

Acknowledgements

I'd like to thank Michelle Yount and Meghan Parkinson of Mount Holyoke College and Rachael Klein of Carleton College for their help in coding this data.

References

Clark, E. V. & Anderson, E. S. (1979). Spontaneous repairs: Awareness in the process of acquiring language. *Papers and Reports on Child Language Development, 16*, 1-12.

Clark, H. H. & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition, 84,* 73-111.

Flavell, J. (2001). *Cognitive Development*. Upper Saddle River, NJ: Prentice Hall.

Fox Tree, J. E. (2001). Listener' uses of *um* and *uh* in speech comprehension. *Memory and Cognition, 29*, 320-326.

Glucksberg, S., Karuss, R. M., & Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology, 3*, 333-342.

Smith, V. & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32*, 25-38.

Van Der Wege, M M. (1996). *The development of repair strategies*. Unpublished manuscript, Department of Psychology, Stanford University, California.

Toward a Graded Model of English Phonology

Brent Vander Wyk (bcv@andrew.cmu.edu)
Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

James L. McClelland (jlm@cnbc.cmu.edu)
Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA

Introduction

Traditional linguistics seeks to specify the universal and absolute properties of phonology and produce a set of inviolable rules. These rules serve to make binary distinctions between allowable forms and disallowed ones. In a thoughtful analysis by Harris (1994) such a set is described. However, this approach has difficulty accounting for graded differences in frequency among phonological forms that do not violate the rules, except to acknowledge certain preferences. For example, post-vocalic stops are more frequent in the context of a short vowel than a long vowel, though they occur with both.

The alternative approach taken in this investigation is to use a set of graded constraints to determine the frequency of a phonological form. This single mechanism captures both the binary and graded patterns assuming that the degree of concordance with the constraints is what determines the frequency. The less concordant a form is with the constraints the lower its frequency, and an unattested form is one that is extremely discordant. The approach taken here is similar to a graded version of Optimality theory (Boersma, 2000; Prince & Smolensky, 1993). In the following model, concordance with constraints is motivated by the observation that more complicated forms tend to be less frequently used than simpler forms, perhaps because as more phonetic material is added to a syllable there is an overall compression that makes articulation and perception difficult.

Graded Model

Phonotactic constraints differ in word internal and word-final contexts, and weaken both across morphological boundaries and from onset to rhyme. Rather than deal with all the sources of complexity at the outset of the investigation we limit our analysis to the rhymes of monosyllabic monomorphemic words. Furthermore, only those rhymes that contain a stop and follow principles of sonority sequencing (Harris, 1994) were considered. Vowels were categorized as either long or short with further distinctions ignored. So, rhymes under consideration have the following form: a long or short vowel followed by an optional liquid, optional nasal, or optional coronal fricative followed by a requisite stop followed by an optional coronal fricative or optional coronal stop. This yields a set of 64 possible rhymes, consisting of a vowel plus up to two additional phonemes, of which 38 are realized in English.

Each rhyme was characterized as a set of complexity-adding features, such as the presence of a fricative after the stop. The frequency of occurrence, measured as the average number of words that use the rhyme per vowel, was then predicted by a linear function which starts with a positive baseline and assesses a weighted penalty for each of the features. The weight was adjusted so that the model would predict the average number of words that use each rhyme type per vowel. Only those forms that occur in the language were allowed to contribute to weight adjustment.

The features used in the model and their final weighted penalty values were: presence of a long vowel (-5.46), stop voicing (-4.79), features to indicate whether the stop is labial (-4.75) or back (-3.22), the presence of a pre-stop homorganic nasal (-10.06), a pre-stop liquid (-13.96), a pre-stop coronal fricative (-12.93), a post stop coronal fricative (-12.80), and a post-stop coronal stop (-15.93). The positive baseline was 21.21. In terms of predicting which terms are attested in English, the model predicts a positive frequency of occurrence for 32 of the 38 that do occur, and a zero frequency of occurrence for 21 the 26 forms that do not. Among the forms that actually occur in English the model predicts 89% of the variance in the frequency per vowel.

Discussion

Previous linguistic work characterized phonology using a set of rules remaining largely unconcerned with graded patterns in the language. The goal for this investigation was to formulate a simple graded constraint model that could account for the binary distinction of which forms are attested in English and which are not, as well as account for the variations in frequencies among occurring forms. Based on the results from this simple model, continued development seems justified.

References

Boersma, P. (2000). *Learning a grammar in Functional Phonology*: Oxford University Press.

Harris, J. (1994). *English Sound Structure*. Oxford: Blackwell.

Prince, A., & Smolensky, P. (1993). *Optimality theory: Constraint interaction in generative grammar.* Technical Report TR2. New Brunswick, NJ: University Center for Cognitive Science, Rutgers.

Natural language computer tutoring vs. human tutoring vs. text studying□□

VanLehn, Kurt
vanlehn@cs.pitt.edu□□

Studies of human tutoring suggest that the participants' use of natural language might be crucial to the effectiveness of human tutoring. In order to study the impact of natural language on learning, we compared 2 kinds of human tutoring (spoken and computer-mediated) with 2 kinds of natural-language-based computer tutoring (Why2-Atlas and Why2-AutoTutor) and 2 kinds text studying. Students solved qualitative physics problems by writing paragraph-long explanations and (in some conditions) discussing them with a tutor. Results from 5 experiments suggest that natural language tutoring is more effective than studying a text without a tutor unless (a) the students are motivated to self-explain the text thoroughly, (b) they have the prior knowledge to successfully self-explain the text, and (c) the content of the text matches the content of the assessments. If all three conditions are met, as they were in some of our experiments, then studying a text elicits the same learning gains as tutoring, even human tutoring. These results are consistent with current theories of cognitive skill acquisition, and with the benefits of tutoring in practical settings where students often lack appropriate engagement, prior knowledge and texts.

Sufficiency: A surprisingly stretchy concept.

Niki Verschueren (Niki.Verschueren@psy.kuleuven.ac.be)
Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)
Géry d'Ydewalle (Géry.Dydewalle@psy.kuleuven.ac.be)
Laboratory of Experimental Psychology
University of Leuven
Tiensestraat 102, 3000 Leuven, Belgium

Introduction

The concepts of necessity and sufficiency play a central role in explaining the reasoning performance. It is often argued that how people interpret the necessity and sufficiency expressed by a conditional relation has a causal impact on the number and types of inferences that are drawn (see e.g. Thompson, 2000). It is however not yet clear whether the formal definitions of necessity and sufficiency reflect the way reasoners use and interpret these concepts.

In logic, one proposition is a necessary condition of another when the second cannot be true while the first is false, and one proposition is a sufficient condition for another when the first cannot be true while the second is false. Research on conditional reasoning revealed that logical conceptions and definitions are not necessarily psychologically relevant or valid. The current experiment will verify whether participants adhere to the logical definitions of the concepts of necessity and sufficiency.

Experiment

A total of 28 first-year psychology students were asked to indicate whether each of four cause-effect combinations are possible or impossible. Figure 1 gives an example of the task for sufficiency. According to the logic definition, we should observe the pattern listed in Table 1 (the definition of sufficiency does not relate to the third combination). For necessity, participants should accept the first and the last combination and reject the third. When a reasoner considers a cause-effect combination possible, the answer is scored as 1; when it is considered impossible it is scored as 0

Figure 1: Example of the possibility-task.

The cause is sufficient for the effect		
Combinations	Possible	Impossible
1. Cause occurs – Effect occurs	x	
2. Cause occurs – No effect		x
3. No Cause – Effect occurs		
4. No Cause – No effect	x	

Table 1 displays the results. According to the formal conceptualisation of necessity the 'no cause-effect' combination is illegal, whereas the combination 'cause–no effect' is irrelevant. As expected, the irrelevant combination was more often considered possible than the illegal combination, *Wilcoxon T* = 15, *Z* = 2.35, *N non-ties* = 14, $p < .05$. For sufficiency, the difference between the irrelevant 'no cause – effect' and illegal 'cause–no effect' combination was not significant. Surprisingly, the illegal combination was considered possible by 60.7% of the participants.

Table 1: Percentage of trials in which each combination was considered possible.

	Cause Effect	Cause No Effect	No Cause Effect	No Cause No Effect
Sufficient	100	**60.7**	46.4	85.7
Necessary	96.4	57.1	14.3	92.9

When we look at the patterns of relevant combinations for a sufficient cause, there were 8 participants (29%) who considered the 'cause–effect', 'cause– no effect' and the 'no cause–no effect' combinations respectively possible, impossible and possible, whereas there were 16 participants (57%) who found all three combinations possible. For necessity, there were 22 participants (79%) that considered the 'cause-effect', 'no cause-effect' and 'no cause- no effect' respectively possible, impossible and possible, whereas only 3 participants (11%) considered all three combinations possible. The 'no cause–effect' combination is thus understood as a combination that contradicts necessity, the combination 'cause–no effect' does not contract sufficiency.

Conclusion

Whereas the subjective conceptualisation of necessity parallels the formal definition, the subjective concept of sufficiency is less stringent than the formal concept. A cause can be considered sufficient to grant the effect, even when the effect does not always follow. However, when causal rules are used to make predictions, it can be adaptive to label a cause that increases the probability of the effect as subjectively sufficient. The observed divergence between the subjective and formal definition raises doubt on the claim that reasoners assess the formal level of sufficiency to derive conditional conclusions.

References

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition, 76, 209-268.*

Toward an Integrated Understanding of the Generation of Place-Fields in the Different Sub-fields of the Hippocampal Region.

Renan W.F. Vitral.[1,2] (renan@cns.bu.edu)
[1]Department of Cognitive and Neural Systems, Boston University.
677 Beacon Street, Boston, MA, 02215. USA

[2]NIPAN – CNPq, Department of Physiology,
Federal University of Juiz de Fora, Brazil.

Introduction

The generation of the "Place Fields" in the hippocampal region is a fundamental phenomenon for the processing of the declarative memory. The electrophysiological attributes are associated to the concomitant generation of theta-gamma activity, while, at the behavioral level, they are expressed as environmental exploration and attention phenomena, when studied in animals. However there were not described robust topographical pathways that sustain a direct association of the afferent pathways from hippocampus to specific cortical regions that would be recipients of the functional processes for the consolidation of memory and the navigational abilities.

Another topic refers to the generation of the "Place Fields" in the several sub-fields of the hippocampal region: the integrated role of these attributes still stays obscure.

Methods

In previous studies we developed a computational neural net based on mathematical attributes of the Gated Dipoles and the opponent processing for the modeling of the generation of place fields. We showed on this model that the temporary entrance of frequencies at the net, more than the topographical organization, would sustain a memory system for the construction of place fields. We used also attributes of self-organizing maps to construct the system.

In this work we expanded the developed net, with the integration of the several hippocampal sub-fields, based on anatomical and electrophysiological data, trying to get a more reliable model of the hippocampal function when working on the cited frequencies.

Results

In our results the concomitant generation of the Place Fields in different sub-fields and in an independent way would work in a combinatory system. So, while the afferent connections to the hippocampus show a progressive topographical condensation with a transduction for aleatory functional patterns in each sub-field, it happens a temporary filtration and accentuation of the developed information based on the temporal properties of the little variations on each frequency arriving the hippocampus. These attributes, associated with the internal abilities of frequency generation, could create a combinatory amplification of the capacities of storage of information that temporally would construct a diverse and accentuated basis for the efferent processes of the hippocampal system.

Conclusions

Finally, our suggestion is that the hippocampal complex functions as a transduction system from topographical to frequency-dependent abilities with a combinatory generation of memory traces based on the temporary weight of the established connections among cells and subfields to subserve the several roles attributed to it, as the participation on the memory consolidation, navigational abilities, emotional expressions, and so, adaptively timed learning.

Acknowledgments

Funding: IBRO Post-doctoral Fellowship Award. CAS/CNS/Boston University. Federal University of Juiz de Fora.

References

Burgess, N., & O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus, 6,* 749-762.

Grossberg, S. (1989). Classical-Conditioning - the Role of Interdisciplinary Theory. *Behavioral and Brain Sciences, 12,* 144-145.

Grossberg, S., & Merrill, J. W. L. (1992). A Neural-Network Model of Adaptively Timed Reinforcement Learning and Hippocampal Dynamics. *Cognitive Brain Research, 1,* 3-38.

Grossberg, S., & Merrill, J. W. L. (1996). The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience, 8,* 257-277.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks, 6,* 895-905.

Vitral, R. W. F. (2004). CA3 place cells in a fixed environment: place fields, hippocampal oscillations, behavioral correlates, and opponent processing. *Proceedings of the Eighth International Conference on Cognitive and Neural Systems* (pp. 12). Boston, MA.

Early Word Learning: How Infants Learn Words that Sound Similar

Julia Wales (jwales@purdue.edu)
Department of Psychological Sciences, 703 Third St.
West Lafayette, IN 47907 USA

George Hollich (ghollich@purdue.edu)
Department of Psychological Sciences, 703 Third St.
West Lafayette, IN 47907 USA

Introduction

The development of both phonological perception and semantic acquisition has been well studied, yet the connection between these remains a mystery. Research in phonological perception has demonstrated that infants are born with the ability to make fine phonetic discriminations (Cristophe, Jacques, & Sebastian-Galles, 2001). However, word-learning research has suggested infants do not make use of this knowledge in learning similar sounding words (Werker & Tees, 2002). Why? The current study suggests that infants have detailed phonetic representations for newly acquired words but suppress this information under certain circumstances.

Method

Infants from the West Lafayette/Lafayette area were tested at four different ages: 14, 18, 22, and 26 months.

The present study used the splitscreen preferential looking paradigm. This paradigm presents two objects, one on each side of a screen while audio stimuli requests one of the objects. The experiment consisted of two sequences. The first sequence attempted to teach infants a novel word (e.g. "chab") while the second sequence attempted to teach a second novel word that was phonologically similar to the previously taught word (e.g. "chas"). The auditory stimuli for this sequence were presented in a different voice (differing in gender) from the second sequence. The order of voices and specific words were counterbalanced across subjects.

Each sequence consisted of four types of trials. Infants first had a training trial where they were presented with a single object on the screen and the novel label for that object. This was always followed by a salience trial where the object the infant heard labeled was presented on one side of the screen and another object was presented on the other side. Auditory stimulus was played that was not intended to direct attention to either object (e.g "What do you see?"). The infants then saw the two test trials (label and similar), the order of which were counterbalanced to control for order effects. In the label condition, the object that the infants saw in training was requested. In the similar condition, the similar word was requested. The logic of the procedure was that if infants learned the word in the test phase, they should look longer at it during the label trials than in the salience trials or the similar trials.

Results and Discussion

Although performance did increase with age, even at 14 months, infants looked significantly longer at the labeled object when it was requested during the first sequence. Furthermore, they did not look longer at the label object in the salience or similar trials. In those trials, infants looked longer at the unlabeled object. This switch in looking suggests that infants noticed and rejected phonetic differences when the voice was the same, demonstrating that they can make fine phonetic discriminations in the context of a word-learning task. However, infants did not notice the same phonetic differences when the talker was changed in the second sequence of trials.

There are two logical explanations for the results from the second sequence of trials. Perhaps infants pragmatically noticed the switch between voices and assumed that the phonetic differences in this case were not meaningful. This strategy may cause mislabeling when the task is to learn similar sounding words, but it may ultimately lead to more successful labeling in the real world where phonetic signals are more variable and normalization is key. Alternatively, it is possible that the task of attaching meaning to the second word caused the difficulty. Specifically, it is possible that infants lost track of which word went with which object (something even adults will do, on occasion). In this case, the switch in voice was irrelevant. Even if the voice had been the same, infants would have had the same difficulties in the second block of trials.

Ongoing studies are examining whether it was the change in talker or the learning of a second word that caused the difficulties in the second block of trials. However, the current results suggest that infants do posses fine phonetic distinctions in a word learning task, even at 14 months, and that they will ignore these distinctions when memory or pragmatic conditions dictate.

References

Cristophe, A., Jacques, M., & Sebastian-Galles, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy, 2(3)*, 385-394.

Werker, J. F. & Tees, R. C. (2002). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 25*, 121-133.

Learning OT Grammars of Syllable Structure

Adam T. Wayment (awayment@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles St. Baltimore, MD 21218 USA

1. Introduction

Optimality Theory (OT) (Prince & Smolensky 1993) has been widely adopted in phonology and has also been successfully applied to syntax, semantics, and pragmatics. One reason OT has been so rapidly accepted is that its initial presentation was closely tied to a connectionist realization (Blutner, et al. forthcoming). Goldwater and Johnson (2003) suggest that another reason for OT's recent dominance is that there are algorithms for learning constraint rankings. However, these elements of OT's success (learning algorithms and connectionist realizations) have not yet been unified in a connectionist network that learns constraint rankings.

A grammar in Optimality Theory is defined by a set of ranked violable constraints. The function GEN takes a base form as input and generates an infinite set of candidates. EVAL then selects the optimal realization from among these candidates, obeying the criterion that the ordering of constraints is strictly dominant. Archangeli (1997) reviews how re-ranking a few violable constraints (ONSET, PEAK, NOCODA, *COMPLEX, FAITHC, AND FAITHV) accounts for a large number of the syllable structures attested in the languages of the world. Within this domain of syllable structure, we explore the dilemma of learning a constraint ranking in a connectionist network. The goal is to design a network that can learn the well-formedness of test syllables, based on positive training data generated from a particular ranking of the violable constraints.

2. Fixed-Point Membership

Now, let **G** be a harmonic grammar. The task is to determine if an input form w is in $L(\mathbf{G})$. Define language membership as follows $w \in L(\mathbf{G})$ iff EVAL(GEN(w)) = w. In words, membership is equated with being a fixed-point of OT generation. This concept of membership provides a powerful framework, in which recognition can be performed via a fixed point test on an input form.

Translating this notion to a Harmonic Network—where input is equivalent to clamping the initial activation state—if, after the network is allowed to harmonize, the input pattern is the same as the output pattern, then the input form is in the language of the harmonic grammar described by the weights of the network. If however, the input and output pattern differ, then the input form is not in the language prescribed by the grammar because the more harmonic activation state corresponds to some other output form, so the form fails the fixed-point test. Thus, learning well-formedness is tantamount to learning the identity map. Poverty of the stimulus issues bear heavily on this problem because we require that the identity map be learned from only positive data.

3. Data, Representation, and Learning

Sample data consists of valid phonetic combinations of consonants and vowels, e.g. for a CV language, [ba], [mi], and [po] may be present in the training. Network evaluation is determined by way of the well-formedness scores of a random collection of test syllables of various structure (CV, CVC, CCV, etc.). Only those syllables that are in the language of the net's grammar will be fixed points, and all forms that are not in the language will not.

We represent syllables in a fully connected-symmetric network made by copying a set of fillers (one unit per consonant or vowel) for every role position (peak, onset, etc.). We allow for complex onsets and codas, by placing multiple filler sets in a position. Thus, 'tint' =[tInt], is represented by activating the *t* unit in the Onset filler set, the *I* unit in the Nucleus filler set, the *n* unit on in the first coda filler set, and the *t* unit in second coda filler set.

We compare two different algorithms for learning the weights of the network: backpropagation through time and Boltzmann learning. Preliminary experiments show that the first is unsuccessful at learning the elements in the complement of the training set are ungrammatical, whereas because of its negative phase of training, Boltzmann learning can learn the well-formedness of syllables.

4. References

Archangeli, D (1997) *Optimality Theory: An Overview*.

Blutner, R., Hendriks P., and de Hoop, H. (2003) *Optimal Forms and Meanings*. Book draft.

Goldwater, S. & Johnson, M. 2003. Learning OT Constraint Rankings Using a Maximum Entropy Model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. 111-20.

Prince, A. and Smolensky, P. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers Univ.

Emergence of features in visual stimuli

Alice Welham (a.k.welham@ex.ac.uk) and A.J. Wills (a.j.wills@ex.ac.uk)
School of Psychology, University of Exeter, Perry Rd., Exeter, EX4 4QG. UK.

It has been suggested that new perceptual features can be "created" when they are necessary for a particular task. For instance, by "unitization" (Goldstone, 2000), components which were previously processed separately become represented as a wholistic unit. Certain associative theories (McLaren, Kaye and Mackintosh, 1989) explain unitization as the establishment of connections between reliably co-occurring elements of a stimulus. By this account, after unitization, sampling a subset of featural elements causes retrieval of the whole feature. Given that the model assumes that only a proportion of elements are sampled on any presentation, unitization could lead to an increase in subjective salience of a feature.

This account does not require that the feature is necessary for a task (e.g., diagnostic of a category) for unitization to occur, merely that its elements co-occur. Experiments 1 and 2 indicate firstly that features emerge through simple pre-exposure as well as when they are diagnostic, and secondly, that the process of emergence may increase the collective salience of the feature's components.

Method

Our stimuli consisted of 75% trial-unique random noise, and 25% "feature", which could occur in any of the four corners of a stimulus. There were four "non-obvious" features (NOF condition) and four "control features" (control condition), which were horizontal lines, vertical lines, and two types of square. Figure 1 shows an example of each, with the feature in the top left.

Forty-eight undergraduate students from Exeter University participated in each of Experiments 1 and 2, for course credits or 4 GBP. In both experiments, half of the participants were in the NOF condition and half in the control condition. Every participant completed a training phase followed by a test phase. The training phase consisted of repeated exposure to two of the four features (of the participant's feature type condition). Stimuli were displayed one after another on a computer monitor, and each stimulus contained one feature, in variable location. In Experiment 1, the training phase was a binary choice category learning task in which each feature was diagnostic of a category, and in Experiment 2, participants had to judge the aesthetic appeal of each stimulus on a 9-point scale.

The test phase (identical for both experiments) involved all four features (two trained and two untrained) from that participant's condition. In the first task, pairs of stimuli containing a common feature (the remainder of each stimulus was independently randomly created) were presented for 2 seconds each, after which a similarity judgment was made on a scale of 1 (not at all similar) to 9 (very similar). This was followed by a triad task, in which

Figure 1: Non-obvious feature stimulus (left) and control stimulus (right)

participants were presented with three stimuli (X, Y and Z) simultaneously, and had to decide which two were the most similar. X and Y shared 25% in the form of one of the "features", and X and Z shared 75% but in the form of trial-unique, randomly created noise. Of principle interest are differences in test phase performance with features that have been trained as opposed to untrained.

Results and discussion

In the NOF condition of both experiments, the number of times that the X and Y pair in the triads task was chosen as more similar than the X and Z pair was significantly greater for trained than untrained features. Contrastingly, training had no effect on control features' salience. The sequential similarity judgment task showed similar results. In the NOF condition of both experiments, similarity judgments were higher for pairs of stimuli containing trained than untrained features. This was not seen for the control features, whose salience significantly *decreased* with training in Experiment 2 (and did not change in Experiment 1). For both test phase tasks, effects of training were not significantly different for the two experiments.

The results indicate that novel features, which are presumably not represented prior to the experiment, became more salient through training. This is not dependent on their explicit usefulness. Theories of the allocation of attention to existing attributes (e.g., Kruschke 1996) would have trouble accounting for the increase in salience due to simple pre-exposure of a feature, and the McLaren et al. model can predict that the unitization process itself may be responsible.

References

Goldstone, R.L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 86-112.

Kruschke, J.K. (1996). Dimensional Relevance Shifts in Category Learning. *Connection Science, 8, 225-247.*

McLaren, I.P.L., Kaye, H. and Mackintosh, N.J. (1989). An Associative Theory of the representation of Stimuli. In *Parallel Distributed Processing* (ed R.G.M. Morris). Clarendon Press, Oxford.

All parts are not created equal: SIAM-LSA

Peter Wiemer-Hastings
peterwh@cti.depaul.edu
DePaul University
School of Computer Science,
Telecommunications, and Information Systems
243 S. Wabash
Chicago IL 60604

The ability to assess the similarity of objects in the world is fundamentally important to our survival. Many theories have been proposed for modeling human similarity judgments. Most of these theories involve comparing the sets of features of the compared items to determine the overlap between them. Many of them completely ignore the structure of the objects and the relationships between the parts. Goldstone (1994) showed that such systems fail to account for human similarity ratings of structured data. His SIAM system used a (non-learning) connectionist architecture to create correspondences between objects and their features in different scenes. Excitatory connections reinforced coherent mappings between objects (e.g. ObjectA to ObjectC and ObjectB to ObjectD). Inhibitory connections fought against redundant or contradictory mappings. Likewise, connections between the features of objects either supported or inhibited each other and the corresponding object–object connections. SIAM's connectionist architecture allowed it to take into account the structure of the scenes and the objects as well as the similarity of the features.

Goldstone examined similarity ratings of visual scenes. His approach represented a scene as a spatially related set of objects (for example, pairs of schematic butterflies). Each object has a set of parts each of which has some value. For example, one of Goldstone's butterflies could be represented as: `(object1 (head square) (tail zig-zag) (body-shading white) (wing-shading checkered))`.

In previous research, we have explored the use of Latent Semantic Analysis (LSA) for judging the semantic similarity of a given sentence to a set of alternative target sentences. Although LSA has been shown to match the reliability of raters with intermediate domain knowledge, the correlation between LSA and human ratings is still somewhat disappointing, generally below 0.5 in a number of studies (Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999). In recent research, we have pursued the general hypothesis that including structural knowledge would improve the correspondence between human and LSA ratings. We found that by performing syntactic analysis of the source and target sentences and separately comparing their subjects, objects, and verbs with LSA, we could reduce the error by over 10% (Wiemer-Hastings & Zipitria, 2001).

In the current research, we explored the use of SIAM to combine the analysis of the structural aspects of the sentences with the semantic similarity ratings provided by LSA. To map this approach to sentences, we broke the inputs into subject, verb, object, and indirect object parts. Thus, a simple representation of the sentence "The dog bit a man" as an object would be: `(object1 (verb "bit") (subject "The dog") (object "a man"))`. The advantage of SIAM-LSA over the previous model (Structured LSA, or SLSA) is that its connectionist architecture allows the different components to "compete" for correspondence, instead of relying on a direct mapping of subject, verb, and object segments. Our basic hypothesis was that SIAM-LSA would provide a closer match to human ratings than SLSA. A secondary hypothesis was that providing a salience value to give differential weight to the different structural components of the sentences would better match human ratings.

In our experiment, we compared human ratings with the basic SIAM-LSA system and the system augmented with salience values. Our results did not support the basic hypothesis. In fact, SIAM-LSA performed worse than LSA or SLSA. When we included empirically derived weights which accentuated verb and object matches but completely devalued subject matches, the ratings correlated with human ratings $r = 0.59$, another 10% reduction in the error over SLSA. In accordance with (Resnik, 1993), this suggests that humans essentially ignore the role of syntactic subjects when matching sentence meanings.

References

Goldstone, R. (1994). Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology*, *20*(1), 3–28.

Resnik, P. (1993). *Selection and Information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). How Latent is Latent Semantic Analysis?. In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence*, pp. 932–937 San Francisco. Morgan Kaufmann.

Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for Syntax, Vectors for Semantics. In *Proceedings of the* 23^{rd} *Annual Conference of the Cognitive Science Society* Mahwah, NJ. Erlbaum.

Testing Simple Rules for Human Foraging in Patchy Environments

Andreas Wilke (wilke@mpib-berlin.mpg.de)
International Max Planck Research School LIFE, Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

John M. C. Hutchinson (hutch@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Peter M. Todd (ptodd@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Background

Food in a natural environment is often distributed in patches, spots of higher resource abundance than in the surrounding area. For an animal or human searching on the sea shore, each patch might be a rock pool. Models of animal foraging have considered the situation where such patches vary in their initial quality (return rates), where this may be hard to judge because food items are hidden, and where foraging progressively depletes the resource. As animals learn about and simultaneously deplete a patch they should eventually decide to move because greater success is expected elsewhere.

The optimal strategy in such a situation is given by the *Marginal Value Theorem* (MVT): leave a patch when the instantaneous rate of return falls below the long-term return rate in the whole environment when following the optimal policy (Charnov, 1976). However, the MVT does not offer a mechanistic solution if mean return rate in the environment is not known and if foraging is a succession of discrete events in which items are encountered stochastically (McNamara, 1982). Behavioural ecologists have both derived optimal departure rules in these circumstances and investigated the performance of sub-optimal rules of thumb (such as giving up after a constant time) which may be computationally simpler (Iwasa, Higashi & Yamamura, 1981; Green, 1984; Bell, 1991). Which rules perform well depends on whether patches are evenly dispersed in quality or some are very good and the others very poor. In the former environment finding an item should decrease the tendency to stay, whereas in the latter the opposite is true. This theory indeed explains why related species of insect utilising differently dispersed resources use different rules.

Hypotheses

We propose that humans also should be adapted to decide when to give up on one food patch and move to another, and that they may apply similar simple heuristics as animals have been shown to use. But because humans are intelligent generalists, feeding on some foods which are evenly dispersed across patches and on some which are aggregated in a few better patches, we further predict that humans are sensitive to this aspect of our environment and are able to adapt our heuristics accordingly. Additionally we propose that the patch-leaving heuristics that we use in foraging tasks are also used to decide when to give up on other tasks. We have designed two computerised experiments to test these hypotheses.

Methods

External search: the fishing task

Participants are given a virtual landscape in which they have to monitor ponds (i.e. patches), forage for fish and decide on how long to stay at each pond. All ponds appear equal, but the number of fish in each varies. Each participant experiences either a dispersed, aggregated or Poisson distribution of fishes per patch, and we will also vary the mean travel time between ponds. The probability of finding a fish is proportional to the number left in the pond. Participants see only the number of fish caught at the current pond (and must judge times and rates without reference to a clock). They receive payment at the end depending on the total number of fish caught at all ponds in a fixed time.

Internal search: the word puzzle task

Participants are presented with a modified anagram task in which they search for words from memory. Meaningful words must be generated out of meaningless sequences of letters. Analogously to the first task, participants experience one of three types of patch quality distribution, must decide when to switch to the next sequence, and are paid by their overall success. We attempt to match the environmental parameters in these two tasks as closely as possible.

References

Bell, W. J. (1991*). Searching Behaviour: the behavioural ecology finding resources*. Kluwer Academic Press.

Charnov, E. L. (1976). Optimal foraging: the marginal value theorem. *Theoretical Population Biology*, *9*, 129-136.

Green, R. F. (1984). Stopping rules for optimal foragers. *American Naturalist*, *123*, 30-43.

Iwasa, Y., Higashi, M. & Yamamura, N. (1981). Prey distributions as a factor determining the choice of optimal foraging strategy. *American Naturalist*, *117*, 710-723.

McNamara, J. M. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology*, *21*, 269-288.

Acquisition of polymorphous concepts

A. J. Wills (a.j.wills@ex.ac.uk), Lyn Ellett and S. E. G. Lea
School of Psychology, University of Exeter, Perry Rd. Exeter. EX4 4QG. ENGLAND.

In a polymorphous concept, features are characteristic rather than defining. In Figure 1a, a triangle, an upwards arrow and a pound sign are characteristic of category A. Stimuli are members of category A if they contain more features characteristic of A than features characteristic of B. Dennis, Hampton and Lea (1973) found that polymorphous concepts took considerably longer to acquire to an errorless criterion than either conjunctive or disjunctive rules; conjunctive rules being precisely the sort of structure rejected as "unnaturalistic" by much of contemporary categorization research.

Humans are not the only species to find the acquisition of polymorphous concepts very difficult. In one study with pigeons (von Fersen & Lea, 1990), separate training on each of the stimulus feature-pairs was eventually required in order to train the concept. If it could be demonstrated, with appropriate control groups, that this sort of pre-training was more effective than an equal length of training on the full problem, this would present a challenge to some theories of learning in both pigeons and in people.

Method

The left-hand panel of Figure 1b shows a stimulus containing all five features characteristic of category A. From the outside in, the five feature-pairs are a) flankers (fine/coarse), b) trapezium, c) stars/blobs, d) colored square (yellow/blue), and e) lines (orientation).

Sixty undergraduate students from Exeter University participated for course credit or 4 GBP. Standard category acquisition procedures were followed throughout - stimuli were presented one at a time, a category decision requested ("category A or B?") and feedback given immediately after each decision.

There were three between-subject conditions. In the SINGLE condition, feature pairs were trained one at a time. For example, a participant might first be trained on the problem "stars -> category A / blobs > category B", and would then move on to the next feature-pair. The order in which the five feature-pairs were trained was randomized across participants. Once all five feature-pairs had been trained individually, participants were moved, in the second phase, to the full polymorphous set of 32 (2^5) stimuli for four blocks of trials. Subjects in the POLY condition received the same total number of training trials as the subjects in the SINGLE condition, but all trials were with the full polymorphous stimuli.

Subjects in the SINGLE (REV) condition received single-feature training in the same manner as the SINGLE group. The difference was that the category associations of three out of the five feature pairs were (unbeknownst to the subject) reversed prior to the polymorphous training phase. Thus, if they had initially been trained that "stars -> category A / blobs > category B", then in the polymorphous phase, "blobs" were characteristic of category A and "stars" were characteristic of category B.

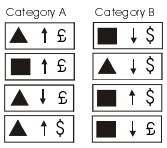


Figure 1a


Figure 1b

Results and discussion

Participants in SINGLE condition were considerably more accurate on the polymorphous problem than participants who had done that problem throughout, but they were also slower (longer RTs).

If these results were entirely due to general motivation or strategic factors then one might expect the reversal in the SINGLE (REV) condition to have relatively little effect. In contrast, if the SINGLE group is more accurate and slower because specific categorical knowledge acquired in phase one is transferred to phase two, then this reversal between the phases should dramatically affect performance. In fact, participants in the SINGLE(REV) condition were significantly worse at the polymorphous problem than participants in either of the other two conditions, but their reaction times were comparable to those in the SINGLE condition. Our working hypothesis is that the specific categorical knowledge acquired in the single feature-pair phase does indeed facilitate polymorphous categorization, but that there may also be important strategic/motivational effects.

Exemplar models (e.g. Nosofsky, 1986) explain aquisition of categorical knowledge by stating that we store labelled instances of categories. In "broad-brush" terms, it seems difficult to explain, from an exemplar-based account, why trading an exact copy of the stimulus you need to make a decision about for stimuli that contain only small parts of it would be beneficial to categorization accuracy.

References

Dennis, I., Hampton, J. A., & Lea, S.E.G. (1973). New problem in concept formation. *Nature, 243*, 101-102.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorisation relationship. *Journal Of Experimental Psychology: General, 115*(1), 39-57.

von Fersen, L., & Lea, S. E. G. (1990). Category discrimination by pigeons using five polymorphous features. *Journal of the Experimental Analysis of Behavior, 54*(2), 69-84.

This research was supported by the BBSRC (U.K.).

Are Relations Directly Detected at Initial Encoding?

Aaron S. Yarlas (yarlasa@gvsu.edu)
Department of Psychology, Grand Valley State University
1 Campus Drive, Allendale, MI 49401 USA

Vladimir M. Sloutsky (VSloutsky@hec.ohio-state.edu)
Center for Cognitive Science, The Ohio State University
1961 Tuttle Park Place, Columbus, OH 43210 USA

This study examines how representations for relations are formed during initial stimulus encoding. One possibility is parallel encoding of elements and relations, such that detection of relations does not require binding, but rather involves matching a new stimulus to a relational template or schema that is retrieved from LTM. A second possibility is a serial account: that there is no direct detection of relations, but rather binding occurs only after elements have detected, at which point their configuration is encoded.

These two possibilities make differing predictions regarding the encoding of elements and relations. First, the first possibility predicts that elements and relations should be represented comparably, whereas the second possibility predicts that relations should be represented less often than elements. Second, the two possibilities differ in their expectation of illusory binding (i.e., binding of target elements to a distracter relation, or distracter elements to a target relation). The first possibility predicts that illusory binding should be symmetrical: given that both elements and relations are identified, there should be both binding of distracter elements to a target relation and binding of a distracter relation to target elements. On the other hand, the second possibility predicts that illusory binding should be asymmetrical: given that elements, but not relations, are identified prior to binding, there should be binding of distracter elements to a target relation, but not binding of a distracter relation to target elements, since this latter case requires an identified relational schema.

To distinguish between the two possibilities, we incorporated an immediate recognition procedure. The general procedure involved subjects receiving on each trial two study items in succession (presented on a computer screen), one a target and one a distracter, with order of presentation randomly counterbalanced across trials. Subjects then received two recognition items simultaneously on the screen, with one of these items being an old item (i.e., identical to the target study item), and the other a foil. Subjects' task was to choose which of these items had been presented during the study phase (i.e., the target). Subjects' choices and latencies were recorded.

The stimuli used were three horizontally-aligned shapes, with elements being shapes of objects and relations being the patterns among 3 shapes within each arrangement. Three relations were used (ABA, AAB, and ABB), with A and B representing different shapes (e.g., an ABA relation might be circle-square-circle). A second within-participants factor was the type of foil paired with an Old target (in the forced-choice recognition task). There were 5 types of foils: E_{New}/R_{Target} foils (same relation as target, but new elements), E_{Target}/R_{New} foils (same elements as target, but a new relation), $E_{Target}/R_{Distractor}$ foils (same elements as target, but relation from the distracter item), $E_{Distractor}/R_{Target}$ foils (same relation as target, but elements from the distracter item), and E_{New}/R_{New} foils (new elements and relations).

The two possibilities predict different patterns of accuracy across foil types. If relations are detected directly, then there should be no difference in accuracy between E_{New}/R_{Target} foils and E_{Target}/R_{New} foils, since participants should be equivalently sensitive to violations of both elements and relations. However, a different pattern was found: participants made fewer choices of E_{New}/R_{Target} foils than E_{Target}/R_{New} foils, indicating that they were more likely to have encoded elements than relations.

Comparisons among foils also afford examination of illusory binding for elements and relations. If participants directly detect elements during encoding, they should be more likely to choose a foil containing elements from the distracter study item ($E_{Distractor}/R_{Target}$ foils) than elements not presented in the study phase (E_{New}/R_{Target} foils) of that trial. At the same time, if participants directly detect relations during encoding, they should be more likely to choose a foil containing relations from the distracter study item ($E_{Target}/R_{Distractor}$ foils) than relations not presented in the study phase (E_{Target}/R_{New} foils). Consistent with both possibilities, there was evidence of illusory binding for distracter elements onto target relations; however, consistent only with the second possibility, there was not symmetrical illusory binding for relations. In other words, participants were not more likely to choose the foil containing distracter relations than the one containing new relations, indicating that distracter relations were not represented above and beyond relations never presented during the trial.

Thus, data clearly support the second possibility, that unlike elements, relations are not detected directly during encoding, but more likely representations for relations emerge from configural binding of elements. Of course, it could very well be the case that this is only true for unlearned relations such as those used here. Future studies will involve training of relations to examine whether well-learned relations are detected directly at encoding.

Acknowledgments

This research was supported by NSF grant REC-0208103.

What is similar in phonological-similarity effect?

Michael C. W. YIP

School of Arts & Social Sciences, The Open University of Hong Kong

myip@ouhk.edu.hk

Introduction

A robust finding in working memory research is that to recall a set of phonologically similar words is much more difficult than to recall a set of phonologically dissimilar words, which is the well-known phonological-similarity effect (Conrad & Hull, 1964). This finding points out that the capacity of information retention in our working memory store more or less depends on the phonological nature of the to-be-memorized information. The more similar (phonologically) of the to-be-memorized item, the more difficult to retain in the working memory store. However, most of the Chinese people have the subjective experience that to immediately recall a set of colloquial slogans in television advertisement is much more easier than to immediately recall a set of common sentences due to the similarity of prosody. There is also evidence showing that rhyming of verbal information usually enhances our memorization ability (Fallon, Groves, & Tehan, 1999). Therefore, how to explain these contradicting observations is very important in order to get a fuller understanding to the operation of the working memory model (Baddeley, 1992).

In the memory study done by Saito (1998), he reported that intonation of a sentence might make a contribution to participants' recall performance (see also Pennington & Ellis, 2000). Following to this point and together with our aforementioned subjective experience, we can see that prosodic information may be useful to our recall performance to the verbal information to an extent, simply like to recall a colloquial slogan in advertisement for a brief period of time. Reviewing the relevant literature so far, there are a lot of empirical works conducted on this issue in the domain of language research: comprehension and production (Sevald & Dell, 1994; Slowiaczek, McQueen, Soltano, & Lynch, 2000; Soto-Faraco, Sebastián-Gallés & Cutler, 2001). However, little consideration has been given to how these different phonological characteristics of a word affect the recall performance in working memory so far despite of their interdependency.

Hence, the major objective in the present study is to examine how the phonological characteristics of a word influence the recall performance in working memory, which is a theoretically interesting but still unexplored question.

Experiment

A typical word span task with Chinese words as the materials was used to examine the phonological characteristics of a word on the recall performance. The main variable in the present experiment is the different degree of phonological similarity, whether those Chinese words presented in the testing lists shared any phonological characteristics (onset, rime and tone) among themselves or not (see Yip, 2004 for details).

Procedure

Participants were asked to read aloud lists of displayed Chinese words on the computer screen one by one. And then, they were asked to recall the Chinese words from the list out loud as many as possible, and the experimenter counted the correctness of their verbal responses at the end of each list. Altogether, each participant received forty lists with a total of 400 Chinese words in the experiment within two sessions with a break. Each session included 100 phonologically similar items and 100 phonologically dissimilar items. The order of presentation for the lists was randomly assigned in the two sessions. The whole experiment lasted for forty minutes.

Results and Discussion

Two main findings in the present study were concluded.

First, the present results indicate that one major source of phonological-similarity decrement comes from the overlapping of the segmental information of the to-be-memorized materials. This phonological overlapping among the to-be-memorized words poses difficulties for participants to perceive and to rehearse because of the acoustic confusion among the words, which is consistent with the previous research findings.

Second, the prosodic information of the to-be-memorized materials seems to be retained longer in the working memory. This overlapping of tonal information among words even produces a phonological-similarity facilitatory effect. Finally, based on the present results, the traditional concept of the term "similar" in the phonological-similarity effect should be re-conceptualized. Because similarity in prosodic information, unlike the similarity in segmental information, will not create any interference effect in working memory, but a facilitatory effect will occur in working memory instead.

References

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556-559.

Conrad, R. & Hull, A. J. (1964). Information, acoustic confusion, and memory span. *British Journal of Psychology, 55*, 429-432.

Fallon, A. B., Groves, K. & Tehan, G. (1999). Phonological similarity and trace degradation in the serial recall task: When CAT helps RAT, but not MAN. *International Journal of Psychology, 34,* 301-307.

Pennington, M. C. & Ellis, N. C. (2000). Cantonese speakers' memory for English sentences with prosodic cues. *Modern Language Journal, 84*, 372-389.

Saito, S. (1998). Effects of articulatory suppression on immediate serial recall of temporarily grouped and intonated lists. *Psychologia, 41*, 95-101.

Sevald, C. A. & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition, 53,* 91-127.

Slowiaczek, L. M., McQueen, J. M., Soltano, E. G. & Lynch, M. (2000). Phonological representations in prelexical speech processing: Evidence from form-based priming. *Journal of Memory and Language, 43,* 530-560.

Soto-Faraco, S., Sebastián-Gallés, N. & Cutler, A. (2001). Segmental and suprasegmental cues for lexical access in Spanish. *Journal of Memory and Language*, *45***,** 412-432.

Yip, M. C. W. (2004). How similar is phonological-similarity effect? Manuscript.

Computing Semantic Representations: A Comparative Analysis

Xiaowei Zhao (xzhao2@richmond.edu)
Ping Li (pli@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23173 USA

How can we formally capture the complex semantic relationships of the human lexicon? This question has been the focus of much recent computational studies. The ability to represent semantics faithfully in formal mechanisms not only is important for understanding the nature of the lexical system of natural languages, but also has significant implications for understanding the mental representation of meaning and its processing and acquisition.

Two best-known models in this regard are the Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and the Hyperspace Analog to Language (HAL; Burgess & Lund, 1997). Both of them are based on large-scale computational analyses of human speech corpora. The LSA model represents the corpora as a high-dimensional co-occurrence matrix of words in texts, and reduces its dimensions using singular value decomposition. The HAL model builds a semantic word co-occurrence matrix, which is weighted according to co-occurrence frequency. In contrast to these two models that automatically extract meanings by computational algorithms, a third model, the WordNet (Miller, 1990), is a computational thesaurus that provides semantic classification of the English lexicon in terms of hyponyms, synonyms, and antonyms, as well as searchable word entries with semantic definitions. Harm (2002) developed a system to extract the semantic features of the WordNet definitions so that lexical entries can be represented as feature-based vectors. In this study, we examine the virtues and drawbacks of the three models with respect to their ability to represent semantics accurately.

Because of our interest in modeling a developmental lexicon, we selected as our test vocabulary 600 words from the MacArthur Communicative Development Inventories (CDI; Dale & Fenson, 1996). The vocabulary can be divided into four major grammatical categories (nouns, verbs, adjectives, and closed-class words). The nouns can be further divided into 12 subcategories according to their meanings (e.g., clothes, toys, food, etc). The LSA, HAL and WordNet matrices used in our analyses were made available either by the authors or by their electronic distributions.

To examine the accuracy of word classification and representations of the three models, we used a simple k-nearest neighbor (kNN) classifier (Duda, Hart & Stork, 2000). The average classification rates of 4 grammatical categories and the 12 noun subcategories were treated respectively with a 5NN classifier. Figure 1 presents the results. It shows that the WordNet vectors give the best classification rates overall, followed by HAL and then LSA for the 4 grammatical categories, and by LSA and then HAL for the 12 noun subcategories. The best performance of the WordNet model indicates that the lexicographic and psycholinguistic analyses of words can yield accurate lexical-semantic representations, although it comes with a price: a significant amount of work is required to hand-code the features of words by human researchers. The better performance of HAL for the major grammatical categories indicates that HAL captures important information about grammatical relationships of words because of its representation and weighting of word sequences (word-to-word co-occurrence matrices). Finally, the better performance of LSA for the noun subcategories indicates that LSA is able to capture more subtle semantic differences and relationships among words, because a word's representation in this model involves a large number of other words in text (word-to-text co-occurrence matrices).

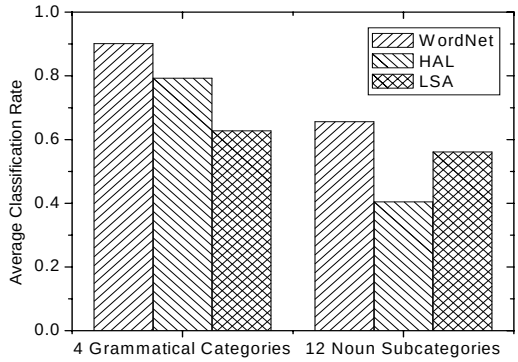


Figure 1: Average classification rates by a 5NN classifier

Acknowledgments

This research was supported by a grant from the National Science Foundation (BCS-0131829).

References

Burgess, C. & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 1-34.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28,* 125-127.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed.). John Wiley and Sons.

Harm, M. (2002). Building large scale distributed semantic feature sets with WordNet. *Technical Report PDP-CNS-02-1*, Carnegie Mellon University.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Miller, G.A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235-312.

Reasoning about Ecological Systems

Corinne Zimmerman (czimmer@ilstu.edu)
Renée M. Tobin (rmtobin@ilstu.edu)
Andrea Cossey (alcosse@ilstu.edu)
Department of Psychology, Illinois State University
Campus Box 4620, Normal, IL 61790 USA

Introduction

Ecological systems are quite complex and dynamic, and are often poorly understood (Groves & Pugh, 2002). The multiple cause-and-effect relationships and second-order effects in such systems are difficult to learn and teach (Hogan, 2000). One example is the use of "bio-control," where the introduction of a species to prey or feed on an unwanted species is used as an alternative to chemical herbicides or pesticides. Because of the complexity of ecological systems and the potential side effects and long-term consequences of such actions it is often difficult to predict precisely the how the system will change over time. Therefore, it is important to understand how individuals think about the ecological systems and the environmental problems that they are being asked to make decisions about.

The purpose of this study was to examine students' reasoning about an ecological management proposal. Such reasoning can be influenced by many factors including conceptual understanding of ecological systems, perceived and actual scientific knowledge, the way information is presented and the influence of other individuals who support or oppose the proposal. In the current study, we focus on two of these factors. First, we were interested in exploring whether the process of self-evaluation (i.e., making students aware of their perceived and actual scientific knowledge) would affect reasoning and decision making. Attitudes, thoughts and beliefs are often automatic or non-conscious, but the conscious evaluation of one's own beliefs or attitudes may be one method for changing thoughts or behaviors (e.g., Bargh & Chartrand, 1999) and so requiring individuals to reflect on their current understanding of science could affect beliefs, reactions, or decisions. Second, we wanted to determine if an ecological management proposal described as a species *introduction* would invoke different mental models than one described as a species *reintroduction*. We hypothesized that the word "reintroduction" may support the inference that the species was "meant" to be part of the ecosystem.

Methods

Eighty undergraduates read and evaluated a brief news article (294 words) that described a proposed initiative to introduce wolves to the Rocky Mountain region. The article was adapted from an online newsletter ("Poll shows strong support for wolves," 2001). Two versions were created with the initiative described as either an introduction or a reintroduction. Participants were asked a number of questions, including whether or not they would support the initiative if required to vote today, their certainty and confidence in their decision, and whether they felt qualified to vote on such an issue. Students also completed a questionnaire to assess perceived and actual background knowledge either before or after reading and evaluating the proposal. Five items assessed perceived scientific knowledge. Actual background knowledge was assessed with a 20-item multiple-choice test covering basic ecological knowledge. Participants were randomly assigned to one of four conditions created by crossing *topic* (introduction vs. reintroduction) with *order of self-evaluation* (before or after evaluating the proposal).

Results

A relationship between the *topic* and voting decision was evident. Participants in the reintroduction condition were more likely to vote in support of the initiative (87.5%) than those in the introduction condition (62.5%) (χ^2 (1) = 6.67, $p < .01$), supporting the idea that this subtle, one-word manipulation may invoke different mental models. *Order of self-evaluation*, however, did not influence voting decisions. We predicted that people who took the test before making a decision would be less certain, confident, and feel less qualified than people who took the test after they made their decision. There was a main effect of order (F (1,76) = 7.21, $p = .009$) on this composite variable, but no main effect of topic or interaction between order and topic (Fs ≈ 1). This effect was not due to differences in either perceived or actual knowledge (Fs ≈ 1). Making individuals aware of their own knowledge did not affect the decision itself, but it did affect certainty and confidence with which they made their decisions.

References

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54*, 462-479.

Groves, F. H., & Pugh, A. F. (2002) Cognitive illusions as hindrances to learning complex environmental issues. *Journal of Science Education and Technology*, *11*, 381-390.

Hogan, K. (2000). Assessing students' systems reasoning in ecology. *Journal of Biological Education, 35*, 22-28.

Poll shows strong support for wolves (2001). Retrieved May 2, 2003, from http://www.sinapu.org/Pages/wolfPoll.html.

Author Index